Politecnico di Bari

Repository Istituzionale dei Prodotti della Ricerca del Politecnico di Bari

Model based security and monitoring systems for resilient industrial control systems: A systematic approach to Cyber Physical Systems Security

This is a PhD Thesis

Original Citation:
Model based security and monitoring systems for resilient industrial control systems: A systematic approach to Cyber Physical Systems Security / Bonagura, V.. - ELETTRONICO. - (2026).

Availability:
This version is available at http://hdl.handle.net/11589/309580 since: 2026-11-09

Published version
DOI:

Publisher: Politecnico di Bari

Terms of use:

(Article begins on next page)

20 November 2026

Italian National Ph.D. Program in Autonomous Systems

ACADEMIC DISCIPLINE: SYSTEMS AND CONTROL ENGINEERING (IINF-04/A)

Final Dissertation

Model based security and monitoring systems for resilient industrial control systems: A systematic approach to Cyber Physical Systems Security

---

by

Bonagura Valeria

Valeria Bonagura

_____

Administrative Headquarters:

Politecnico di Bari – Department of Electrical and Information Engineering

Hosting University:

Università degli Studi di Roma Tre – Department of Civil, Computer Science and Aeronautical Technologies Engineering

Referees:

Prof. Karen Eguiazarian

Prof. Tiago Cruz,

Supervisors:

Prof. Stefano Panzieri

_____

Prof. Federica Pascucci

_____

Coordinator of Ph.D Program

Prof. Mariagrazia Dotoli

_____

Course no. 38, 01/11/2022-31/10/2025

LIBERATORIA PER L'ARCHIVIAZIONE DELLA TESI DI DOTTORATO

Al Magnifico Rettore
del Politecnico di Bari

Il/la sottoscrittoa Valeria Bonagura nata a Roma il 12\12\1998
residente a Roma in via Lorenzo Bonincontri 63 e-mail valeria.bonagura@uniroma3.it
iscritto al 3° anno di Corso di Dottorato di Ricerca in sistemi autonomi (DAUSY) ciclo XXXVIII
ed essendo stato ammesso a sostenere l'esame finale con la prevista discussione della tesi dal titolo:

DICHIARA

1) di essere consapevole che, ai sensi del D.P.R. n. 445 del 28.12.2000, le dichiarazioni mendaci, la falsità negli atti e l'uso di atti falsi sono puniti ai sensi del codice penale e delle Leggi speciali in materia, e che nel caso ricorressero dette ipotesi, decade fin dall'inizio e senza necessità di nessuna formalità dai benefici conseguenti al provvedimento emanato sulla base di tali dichiarazioni;
2) di essere iscritto al Corso di Dottorato di ricerca in sistemi autonomi (DAUSY) ciclo XXXVIII, corso attivato ai sensi del "*Regolamento dei Corsi di Dottorato di ricerca del Politecnico di Bari"*, emanato con D.R. n.286 del 01.07.2013;
3) di essere pienamente a conoscenza delle disposizioni contenute nel predetto Regolamento in merito alla procedura di deposito, pubblicazione e autoarchiviazione della tesi di dottorato nell'Archivio Istituzionale ad accesso aperto alla letteratura scientifica;
4) di essere consapevole che attraverso l'autoarchiviazione delle tesi nell'Archivio Istituzionale ad accesso aperto alla letteratura scientifica del Politecnico di Bari (IRIS-POLIBA), l'Ateneo archivierà e renderà consultabile in rete (nel rispetto della Policy di Ateneo di cui al D.R. 642 del 13.11.2015) il testo completo della tesi di dottorato, fatta salva la possibilità di sottoscrizione di apposite licenze per le relative condizioni di utilizzo (di cui al sito http://www.creativecommons.it/Licenze), e fatte salve, altresì, le eventuali esigenze di "embargo", legate a strette considerazioni sulla tutelabilità e sfruttamento industriale/commerciale dei contenuti della tesi, da rappresentarsi mediante compilazione e sottoscrizione del modulo in calce (Richiesta di embargo);
5) che la tesi da depositare in IRIS-POLIBA, in formato digitale (PDF/A) sarà del tutto identica a quelle **consegnate**/inviate/da inviarsi ai componenti della commissione per l'esame finale e a qualsiasi altra copia depositata presso gli Uffici del Politecnico di Bari in forma cartacea o digitale, ovvero a quella da discutere in sede di esame finale, a quella da depositare, a cura dell'Ateneo, presso le Biblioteche Nazionali Centrali di Roma e Firenze e presso tutti gli Uffici competenti per legge al momento del deposito stesso, e che di conseguenza va esclusa qualsiasi responsabilità del Politecnico di Bari per quanto riguarda eventuali errori, imprecisioni o omissioni nei contenuti della tesi;
6) che il contenuto e l'organizzazione della tesi è opera originale realizzata dal sottoscritto e non compromette in alcun modo i diritti di terzi, ivi compresi quelli relativi alla sicurezza dei dati personali; che pertanto il Politecnico di Bari ed i suoi funzionari sono in ogni caso esenti da responsabilità di qualsivoglia natura: civile, amministrativa e penale e saranno dal sottoscritto tenuti indenni da qualsiasi richiesta o rivendicazione da parte di terzi;
7) che il contenuto della tesi non infrange in alcun modo il diritto d'Autore né gli obblighi connessi alla salvaguardia di diritti morali od economici di altri autori o di altri aventi diritto, sia per testi, immagini, foto, tabelle, o altre parti di cui la tesi è composta.

Rome, 15\01\2026 Firma Valeria Bonagura

La sottoscritto, con l'autoarchiviazione della propria tesi di dottorato nell'Archivio Istituzionale ad accesso aperto del Politecnico di Bari (POLIBA-IRIS), pur mantenendo su di essa tutti i diritti d'autore, morali ed economici, ai sensi della normativa vigente (Legge 633/1941 e ss.mm.ii.),

CONCEDE

- al Politecnico di Bari il permesso di trasferire l'opera su qualsiasi supporto e di convertirla in qualsiasi formato al fine di una corretta conservazione nel tempo. Il Politecnico di Bari garantisce che non verrà effettuata alcuna modifica al contenuto e alla struttura dell'opera.
- al Politecnico di Bari la possibilità di riprodurre l'opera in più di una copia per fini di sicurezza, back-up e conservazione.

Rome, 15\01\2026 Firma Valeria Bonagura

Valeria Bonagura

Model based security and monitoring systems for resilient industrial control systems

A systematic approach to Cyber Physical Systems Security

Thesis submitted for the degree of Philosophiae Doctor

Italian National Ph.D. Program in Autonomous Systems
Università degli studi di Roma Tre

Tutors
Prof. Engr. *Federica Pascucci*
Prof. Engr. *Stefano Panzieri*



Roma Tre

2025



Dissertation submitted for the degree of *Philosophiae Doctor*
Italian National Ph.D. Program in Autonomous Systems

Cycle:
38th

Administrative Headquarters:
Politecnico di Bari

Hosting University:
Università degli studi Roma Tre

Title:
Model based security and monitoring systems for resilient industrial control systems: A systematic approach to Cyber Physical Systems Security

Ph.D Candidate:
Valeria Bonagura, Università degli Studi Roma Tre (Rome, Italy)

Tutors:
Prof. Engr. Federica Pascucci, Università degli Studi Roma Tre (Rome, Italy)
Prof. Engr. Stefano Panzieri, Università degli Studi Roma Tre (Rome, Italy)

Coordinator:
Prof. Engr. Mariagrazia Dotoli, Politecnico di Bari (Bari, Italy)

External Reviewers:
Prof. Engr. Karen Eguiazarian, Tampere University (Tampere, Finland)
Prof. Engr. Tiago Cruz, University of Coimbra (Coimbra, Portugal)

Last version:
November 6, 2025

All rights reserved. No part of this publication may be reproduced or transmitted, in any form or by any means, without permission.

Abstract

Cyber–Physical Systems are the technological backbone of modern industrial, energy, and transportation infrastructures. They integrate computation, communication, and physical processes to enable automation, real-time control, and large-scale coordination. Yet, this tight interconnection also introduces fragility: faults or cyber attacks in the digital layer can quickly propagate through physical feedback loops, compromising safety and reliability. Ensuring the security and resilience of such systems therefore requires a unified approach, where estimation, control, communication, and interdependence are treated as mutually reinforcing aspects of the same problem.

This thesis develops a comprehensive framework for model-based security and monitoring in resilient control systems. The central idea is that resilience begins with trustworthy perception — the ability of a system to maintain an accurate representation of its physical state even in the presence of deception or uncertainty. To this end, hybrid estimation architectures are designed to detect and isolate faults or intrusions in real time. By combining stochastic filtering with deterministic reasoning, these observers achieve a balance between sensitivity to attacks and robustness against modeling errors, allowing for consistent reconstruction of compromised signals.

Building on this foundation, the research investigates how the reliability of information changes over time in adversarial environments. It introduces a formal metric to quantify how long a system operates with incorrect or outdated information, capturing the dynamics of misinformation caused by attacks such as data falsification or denial of service. This concept is analyzed within a game-theoretic framework that models the strategic interaction between defenders and adversaries, providing analytical insight into optimal attack and protection strategies.

The work then extends these ideas to distributed and cooperative settings. In networked systems, agents must coordinate their actions through potentially unreliable communication channels. To address this challenge, the thesis proposes a secure consensus mechanism that integrates trust assessment directly into the coordination process. Each agent evaluates the reliability of its neighbors and adjusts its interactions accordingly, enabling agreement and stability even in the presence of malicious participants or uncertain data. The same principles are also adapted to ensure privacy-preserving coordination, allowing agents to cooperate without disclosing sensitive internal information.

Finally, the thesis scales this perspective up to the systemic level, focusing on the interdependencies among critical infrastructures such as energy, telecommunications, transport, and finance. Using an agent-based simulation framework, it models how local cyber incidents can trigger cascading effects across sectors and how uncertainty — both statistical and epistemic — influences resilience. This modeling approach provides a tool for national-scale risk assessment and supports the design of coordinated protection strategies for interconnected infrastructures.

Overall, the thesis presents an integrated view of security and resilience in cyber–physical systems, linking low-level estimation and detection to high-level systemic analysis. By connecting mathematical modeling, information dynamics, and cooperative control, it contributes to the development of infrastructures that are not only intelligent and interconnected but also robust and trustworthy in the face of evolving cyber threats.

Contents

Preface **xiv**

Introduction

1 Introduction **2**
1.1 Context and Motivation . . . 2
1.2 Preliminaries and Notation . . . 3
1.2.1 Notation and Conventions . . . 3
1.2.2 Graph–Theoretic Representation . . . 4
1.2.3 Multi–Agent Dynamical Model . . . 5
1.2.4 Consensus in Multi–Agent Systems . . . 6
1.2.5 State Observers . . . 8
1.2.6 Game Theory . . . 10
1.2.7 Information Freshness . . . 11
1.2.8 Evidence Theory . . . 13
1.2.9 Holistic and Reductionistic Approaches . . . 14
1.3 State of the Art . . . 15
1.3.1 Control-Theoretic Foundations of CPS Security . . . 15
1.3.2 Dynamical Modeling of Adversarial Behavior . . . 16
1.3.3 Observer-Based Detection as the Local Defense Layer . . . 18
1.3.4 Information Freshness as a Systemic Security Variable . . . 20
1.3.5 Adversarial Resilience in Networked Consensus . . . 22
1.3.6 Modelling Cascading Effects in Interdependent Critical Infrastructures . . . 24
1.3.7 Unified Perspective . . . 26
1.4 Thesis Overview and Contributions . . . 27
1.5 Thesis Organization . . . 29
References — Introduction . . . 32

Observer Based Attack Detection and Isolation

2 Introduction to Observer-Based Attack Detection and Isolation **45**
2.1 Literature review on Model Based Attack Detection and Isolation . 47
2.1.1 Observer-Based Detection Strategies . . . 53
2.2 Thesis Contributions . . . 63

3 Exploiting sparsity for efficient Kalman filtering **64**
3.1 System Model . . . 65
3.1.1 Continuous- and Discrete-Time Models . . . 65
3.1.2 Centralized Kalman Filtering . . . 65
3.1.3 Steady state behavior . . . 67
3.1.4 Steady-state under the invertibility of R . . . 67
3.2 Decaying Properties . . . 67
3.2.1 Generalities . . . 67
3.2.2 SED formalism and its decaying properties . . . 68
3.2.3 SED properties . . . 70
3.2.4 Weak-SED behavior of $\bar{P}_r$ . . . 72
3.3 Simulation Results . . . 74
3.3.1 Simulation scenario . . . 74

3.3.2 The search for stationary points . . . 76
3.3.3 Tracking properties . . . 77
3.4 Implications on Kalman System Design . . . 78
3.4.1 Distance-based pruning for spatially localized filters . . . 78
3.4.2 Key Performance Indicators . . . 79
3.4.3 Numerical results . . . 80
3.5 Discussion . . . 82

4 Multirate Interlaced Kalman Filter and Attack Detection 83
4.1 Problem Statement . . . 84
4.1.1 System Model . . . 84
4.1.2 Attacker Model . . . 85
4.2 Observer Design . . . 86
4.2.1 Multirate Interlaced Kalman Filter . . . 86
4.2.2 Stability Analysis under Asynchronous Operation . . . 89
4.2.3 Consistency and Information Content . . . 90
4.2.4 Consistency of Multirate Interlaced Kalman Filter . . . 91
4.3 Attack Detection Method . . . 93
4.4 Simulations . . . 94
4.5 Discussion . . . 96

5 Sliding Mode Observer for Distributed Attack Reconstruction 99
5.1 System Model . . . 100
5.1.1 Networked System Model . . . 100
5.1.2 Sensor Attack Model . . . 101
5.1.3 Problem formulation . . . 103
5.2 Modular Distributed Observer Architecture . . . 103
5.2.1 Functional Overview . . . 104
5.2.2 Architecture Diagram and Information Flow . . . 104
5.2.3 Observer Decomposition . . . 104
5.3 Level 1 – Sliding-Mode Output Regulator . . . 105
5.3.1 Output Regulator Design . . . 105
5.3.2 Error Dynamics . . . 106
5.4 Level 2 – State–Attack Observer . . . 110
5.4.1 Dynamic Constraint on the Sliding Layer . . . 110
5.4.2 Attack–State Coupling Dynamics . . . 111
5.4.3 Detectability and Observer Synthesis for the Extended Dynamics . . . 112
5.5 Main Result . . . 116
5.6 Numerical Results . . . 116
5.6.1 Network Configuration and Setup . . . 116
5.6.2 Observer Implementation . . . 117
5.6.3 Attack Scenarios . . . 118
5.6.4 Results and Discussion . . . 118
5.7 Discussion . . . 118
References — Observer Based Anomaly Detection and Isolation . . . 121

Information Freshness in Adversarial Multi-Agent Systems

6 Introduction to Information Freshness in Adversarial Multi-Agent Systems 137
6.1 Literature Review on Information Freshness under Adversarial Conditions . . . 139
6.1.1 Metrics Classification . . . 140
6.1.2 Models of Agent Interaction . . . 142
6.1.3 Solution Strategies . . . 144
6.2 Thesis Contributions . . . 147

6.3 Modeling Formalism Used in This Thesis . . . 148
6.3.1 Problem Setting and Architecture . . . 148
6.3.2 Drift Rate: Physical Origin and Modeling Role . . . 149
6.3.3 Markov Modeling of Correctness and Attacks . . . 150

7 False Data Injection 154
7.1 False Data Injection: System-Level Impact . . . 155
7.2 Static Game-Theoretic Modeling with Complete Information . . . 156
7.2.1 Game-Theoretic formulation . . . 156
7.2.2 Derivation of the Nash Equilibrium . . . 157
7.2.3 Numerical Results . . . 159
7.3 Bayesian Game-Theoretic Modeling with Incomplete Information . 160
7.3.1 Bayesian Game- Theoretic Formulation . . . 162
7.3.2 Derivation of the Bayesian Nash Equilibrium . . . 163
7.3.3 Numerical Results and Discussion . . . 165
7.4 AoII in Multi-Sensor Systems . . . 168
7.4.1 Modeling Approach and Motivation . . . 170
7.4.2 Return Times and System-Level AoII . . . 171
7.4.3 Game-Theoretic Analysis . . . 171
7.4.4 Main Results . . . 173
7.4.5 Numerical diagnostics for the multipliers . . . 176
7.4.6 Numerical Results . . . 178
7.5 Discussion . . . 182

8 Denial-of-Service Attacks 184
8.0.1 System Modeling under DoS . . . 184
8.1 Game-Theoretic Formulation and Nash Equilibrium . . . 186
8.1.1 Numerical Results . . . 188
8.1.2 Discussion . . . 190

9 Data Substitution Attacks 192
9.1 System Modeling under Data Substitution . . . 192
9.2 Game-Theoretic Formulation . . . 194
9.3 Analytical Derivation of the Nash Equilibrium . . . 195
9.4 Discussion . . . 197

10 Comparative Analysis 198
10.1 Adversarial Strategies in Perspective . . . 198
10.2 Equilibrium Properties and Attack Feasibility . . . 199
10.3 Simulation-Based Comparative Assessment . . . 200
10.4 Discussion . . . 201
References — Information Freshness in Adversarial Multi-Agent Systems . 205

Secure and Resilient Consensus in Networked Multi-Agent Systems

11 Introduction to Secure and Resilient Consensus in Networked Multi-Agent Systems 210
11.1 Literature review on Resilient Consensus . . . 212
11.2 Thesis Contribution . . . 214

12 Privacy-Preserving Consensus 215
12.1 Problem Setting . . . 215
12.2 Dynamic Reputation Assignment . . . 216
12.3 Weight Correction . . . 218
12.4 Simulations . . . 221
12.5 Discussion . . . 222

13 Privacy-Preserving Consensus **224**
13.0.1 Privacy-Preserving Consensus . . . 225
13.1 Synchronization in networks of nonlinear continuous-time dynamical systems . . . 226
13.2 Proposed approach . . . 229
13.3 Security and Privacy . . . 231
13.3.1 Security Analysis . . . 231
13.3.2 Privacy Analysis . . . 232
13.4 Simulations . . . 234
13.5 Discussion . . . 239
References — Secure and Resilient Consensus in Networked Multi-Agent Systems . . . 241

Observering and Simulation of Critical Infrastructure Interdependencies

14 Introduction to Critical Infrastructure Interdependencies Modelling **246**
14.1 Literature Review on Modeling Interdependencies and Cascading Effects in Critical Infrastructures . . . 247
14.1.1 Conceptual Foundations of Interdependency Modeling . . 248
14.1.2 Regulatory and Governance Frameworks . . . 249
14.1.3 Modeling Cascading Effects in Interdependent Systems . . 250
14.1.4 Advanced Research Directions . . . 253
14.2 System Model and CISIApro 2.0 . . . 254
14.3 Thesis Contribution . . . 256

15 National Cybersecurity Perimeter: Modelling Systemic Risk with CISIApro 2.0 **257**
15.1 Regulatory and Systemic Framework of the Italian National Cybersecurity Perimeter . . . 258
15.2 System Model . . . 259
15.3 Cyber scenarios and results . . . 260
15.4 Discussion . . . 263

16 Managing Uncertainty in Interdependency Models: A Possibilistic Extension of CISIApro 2.0 **266**
16.1 System Model . . . 267
16.2 Mathematical background . . . 269
16.3 Results . . . 270
16.4 Discussion . . . 274

17 Fuzzy Modelling of Cyber–Physical Interdependencies: Toward Hybrid Possibilistic Reasoning in CISIApro 2.0 **276**
17.1 Modelling Approach . . . 277
17.2 Mathematical Background for Fuzzy Sets . . . 277
17.3 Case Study and Results . . . 278
17.4 Discussion . . . 283
References — Observering and Simulation of Critical Infrastructure Interdependencies . . . 284

Conclusions

Appendices **291**

A Proof of Theorems in Chap. 3 **292**
A.1 Proof of Theorem 3.2.1 . . . 292
A.2 Proof of Theorem 3.2.2 . . . 292
A.3 Proof of Theorem 3.2.3 . . . 293
A.4 Proof of Theorem 3.2.4 . . . 294

A.5 Proof of Theorem 3.2.5 294
A.6 Proof of Theorem 3.2.6 294
A.7 Proof of Theorem 3.2.7 296
A.8 Theorem 3.2.8 296
A.9 Proof of Theorem 3.3.1 297
A.10 Proof of Theorem 3.4.1 297

B Proof of Theorems in Chap. 4 299
B.1 Proof of Lemma 4.2.1 299
B.2 Proof of Lemma 4.2.2 299
B.3 Proof of Theorem 4.2.1 301
B.4 Proof of Theorem 4.2.2 302

C Proof of theorem in Chap. 4 304
C.1 Proof of the existence and uniqueness of K^* 304

List of Figures

3.1 Decaying structure in Kalman filter matrices 66
3.2 Comparison of weak-SED and strong-SED decay 69
3.3 True vs. strong-SED decay comparison 69
3.4 Flattening of function g_k. 72
3.5 Estimated (x, β, λ) triplets for weak-SED fitting 73
3.6 Estimated weak-SED parameter λ for Z_k 75
3.7 Decay properties of $\bar{P}_r$ and excess threshold τ behavior 76
3.8 Convergence of weak-SED estimates for Z_k 77
3.9 Steady-state behavior of weak-SED parameters of $\bar{P}_r$ 78
3.10 KPI comparison for distance-based truncation 81
3.11 Impact of network size on performance metrics 81

4.1 Example of physical–cyber interconnections 87
4.2 Physical connection graph 94
4.3 Monte Carlo validation of Eq. 4.27 95
4.4 Monte Carlo analysis with reduced compensation on subsystem 1 96
4.5 Comparison of detection statistics and thresholds 96
4.6 Batch detection statistic under FDI on subsystem 1 97

5.1 Modular observer architecture 103
5.2 Directed interconnection topology among the six agents. 117
5.3 Estimated vs. true attack signals. Gray shading indicates active attack windows. 118
5.4 Norm of the attack estimation error $\|\boldsymbol{e}_{\delta}^{(i)}(t)\|$ across all agents. 119
5.5 Norm of the state estimation error $\|e_{x}^{(i)}(t)\|$ for each agent. 119

6.1 AoII dynamics 150
6.2 Nominal correctness dynamics 151
6.3 Adversarial correctness dynamics 152

7.1 Continuous-time Markov chain under False Data Injection. 155
7.2 Adversary's utility $u_{\mathrm{M}}(p, q)$ 160
7.3 Transmitter's utility $u_{\mathrm{N}}(p, q)$ 160
7.4 Equilibrium transmission rate $p^{\star}$ with and without adversary 161
7.5 Equilibrium transmission rate $p^{\star}$ with and without adversary ($K = 0.1$, $d = 1$). 161
7.6 Equilibrium values of $p^{\star}$ and $q^{\star}$ 162
7.7 Comparison between transmission rates at BNE and benchmarks $\alpha = 1$ and $\alpha = 0$. 166
7.8 Injection rates for triangular, uniform, and Bernoulli $f_K(k)$. $C = 4$, $d = 0.001$ 166
7.9 Ratio of Δ_{F} expectations under different $f_K(k)$ 167
7.10 Expected utility of the transmitter 167
7.11 Bayesian Nash Equilibrium under triangular information distribution . . 168
7.12 Bayesian Nash Equilibrium under triangular distribution and varying medians 168
7.13 Equilibrium transmission and injection rates under asymmetric priors . . 169
7.14 Equilibrium performance ratio under different prior distributions 169
7.15 Continuous-time Markov model of sensor correctness dynamics 170
7.16 $\mathcal{L}_{\lambda_T}(\lambda_T; \lambda_M)$ for different (λ_M, C_i). 177
7.17 $\mathcal{L}_{\lambda_M}(\lambda_M)$ under varying totals and attack costs 177
7.18 Equilibrium transmission p_i and injection rates q_i 179

7.19 Equilibrium transmission p_i and injection rates q_i . . . 180
7.20 Equilibrium transmission p_i and injection rates q_i . . . 180
7.21 Equilibrium transmission p_i and injection rates q_i . . . 181
7.22 Equilibrium transmission p_i and injection rates q_i . . . 181
7.23 Equilibrium transmission p_i and injection rates q_i . . . 182

8.1 Continuous-time Markov chain for information correctness under Denial-of-Service attack . . . 185
8.2 Continuous-time Markov chain determining the malicious agent's transition between active and inactive states . . . 185
8.3 Comparison of transmitter activity rate p . . . 189
8.4 Ratio between the attacker rate r and the transmitter rate p . . . 189
8.5 Comparison of AoII as a function of increasing transmission price C . . . 190
8.6 Comparison of transmitter activity rate p . . . 191

9.1 Continuous-time Markov chain for information correctness under Data Substitution attack . . . 193
9.2 Continuous-time Markov chain describing the attacker's operational states 194

10.1 Transmission rate at NE, varying the transmitter's cost of action. . . . 201
10.2 Transmission rate at NE, varying the opponent's cost of action. . . . 201
10.3 Average AoII. . . . 202
10.4 Average AoII. . . . 202
10.5 Ratio between enemy activity rate r and legitimate agent transmission rate p, varying C. . . . 203
10.6 Ratio between enemy activity rate r and legitimate agent transmission rate p, varying K. . . . 203

12.1 Agent v_i computes $r_k^{(i,j)}$ for all $u^{(j)} \in \mathcal{N}_i$ . . . 218
12.2 Simulation setup and behavior of the traditional consensus under attack . 222
12.3 Temporal evolution of the weighted consensus algorithm under attack . . 222

13.1 Balanced and unbalanced ring topologies . . . 233
13.2 Temporal evolution of the hidden consensus variable in the Chua oscillator network . . . 235
13.3 Temporal evolution of the shared state variable in the Chua oscillator network 235
13.4 Temporal evolution of the Chua oscillator dynamics in a ring topology . . 236
13.5 Temporal evolution of the hidden consensus variable in an unbalanced Chua oscillator network . . . 236
13.6 Temporal evolution of the shared state variable $\boldsymbol{z}(t)$ . . . 237
13.7 Temporal evolution of the dynamics $\boldsymbol{x}(t)$ . . . 237
13.8 Unbalanced graph topology with 6 agents. . . . 238
13.9 Temporal evolution of local state variable $\boldsymbol{h}(t)$ . . . 239
13.10 Temporal evolution of the shared state variable $\boldsymbol{z}(t)$ . . . 239
13.11 Temporal evolution of the dynamics $\boldsymbol{x}(t)$ . . . 240

14.1 The Mixed Holistic–Reductionist (MHR) framework . . . 255
14.2 Representation of an infrastructure as an agent in CISIApro 2.0. . . . 255

15.1 Interdependency graph across the eight infrastructures considered in the case study. Directed links denote functional and informational dependencies among sectors. . . . 260
15.2 System-wide synoptic under ransomware on the banking sector . . . 261
15.3 Bank view under ransomware . . . 261
15.4 Airline operator under ransomware . . . 262
15.5 Railway operator under ransomware . . . 262
15.6 Telecommunications operator under ransomware . . . 262

15.7 System-wide synoptic under DDoS on telecommunications 263
15.8 Telecommunications operator under DDoS 263
15.9 Railway operator: significant degradation at stations with single-provider connectivity and ICT coordination bottlenecks. 263
15.10 Government department: online functions and ICT assets degrade due to telecom unavailability. 264
15.11 Electricity distribution: asymmetric impact driven by single-homing versus dual-provider redundancy. 264
15.12 Airline operator: ICT disruption propagates to customer services and airport operations through network dependency. 264

16.1 Uncertainty-aware synoptic dashboard in CISIApro 2.0 268
16.2 Necessity measure for agents in electricity distribution company 2 271
16.3 Possibility measure for agents in electricity distribution company 2 271
16.4 Necessity measure for the telecommunications network 272
16.5 Possibility measure for the telecommunications network 272
16.6 Necessity measure for the airline sector 272
16.7 Possibility measure for the airline sector 273
16.8 Necessity measure for the railway sector 273
16.9 Possibility measure for the railway sector 273
16.10 Necessity measure for the bank sector 274
16.11 Possibility measure for the bank sector 274
16.12 Government department 274

17.1 Triangular fuzzy membership functions 278
17.2 Cyber–operational risk TFN at an electrical substation 279
17.3 Cyber–operational risk TFN at another electric node 279
17.4 Electric distribution grid 2 in use case 1 280
17.5 Telecommunication network 2 in use case 1 280
17.6 Railway network in use case 1 281
17.7 Air transport network in use case 1 281
17.8 Bank network in use case 1 282
17.9 Use case 1 dashboard 282
17.10 Government department node 282
17.11 Government department in use case 2 283

C.1 $f(\alpha)$ and $g(\alpha)$ plot setting $d = 1$ and $C = 2$ 305

List of Tables

1.1 Notation Summary 4
1.2 Summary of thesis structure and main contributions 30

3.1 Eigenvalue ranges for inverse theorems 71

5.1 Dimensions of the local subsystems. 117

7.1 PDF of considered distributions 165

Preface

This thesis is submitted in partial fulfillment of the requirements for the degree of *Philosophiae Doctor* in Autonomous Systems.

List of the Publications Written by the Candidate in the Triennium

International Journal Articles (Published)

[J1] C. Fioravanti, V. Bonagura, G. Oliva, S. Panzieri, C. N. Hadjicostis. *Exploiting the synchronization of nonlinear dynamics to secure distributed consensus.* IEEE Open Journal of Control Systems, 2023.

[J2] V. Bonagura, S. Panzieri, F. Pascucci, L. Badia. *Strategic interaction over age of incorrect information for false data injection in cyber-physical systems.* IEEE Transactions on Control of Network Systems, 2024.

International Journal Articles (Under Review)

[J3] V. Bonagura, S. Panzieri, F. Pascucci, L. Badia. *Controlling Age of Incorrect Information Violation Under Data Drift and Strategic Attacks.* IEEE Transactions on Systems, Man, and Cybernetics, 2025. (under review)

[J4] V. Bonagura, T. Erseghe, C. Foglietta, F. Pascucci. *Exploiting Sparsity for Efficient Kalman Filtering.* IEEE Transactions on Automatic Control, 2025. (under review)

International Journal Articles (Under Preparation)

[J5] C. Foglietta, V. Bonagura, S. Panzieri, F. Pascucci, L. Badia. *Game-Theoretic Analysis of Cyber-Physical Systems Under False Data Injection Attacks.* (under review)

[J6] V. Bonagura, C. Foglietta, F. Pascucci. *Survey on Observer-Based Attack Detection.* 2025. (under preparation)

[J7] V. Bonagura, A. Buratto, L. Badia, F. Pascucci. *Survey on the Use of Information Freshness Metrics for Optimizing Remote Sensing Performances.* 2025. (under preparation)

[J8] V. Bonagura, T. Erseghe, C. Foglietta, F. Pascucci, L. Badia. *Kalman Filter Covariance Decay for Large-Scale Sparse Systems.* 2025. (under preparation)

[J9] V. Bonagura, A. Kasis, M. M. Polycarpou, S. Panzieri, F. Pascucci. *Multi-Agent Collaboration for Accurate Sensor Attack Reconstruction in Cyber-Physical Systems.* 2025. (under preparation)

International Conference Proceedings (Published)

[C1] V. Bonagura, C. Fioravanti, G. Oliva, S. Panzieri. *Resilient consensus based on evidence theory and weight correction.* American Control Conference (ACC), 2023.

[C2] V. Bonagura, C. Foglietta, S. Panzieri, F. Pascucci. *Multirate Interlaced Kalman Filter.* Mediterranean Conference on Control and Automation (MED), 2023.

[C3] V. Bonagura, C. Foglietta, S. Panzieri, F. Pascucci. *Enhancing industrial cyber-physical systems security with smart probing approach.* IEEE International Conference on Cyber Security and Resilience (CSR), 2023.

[C4] V. Bonagura, L. Badia, S. Panzieri, F. Pascucci. *A game of age of incorrect information against an adversary injecting false data.* IEEE International Conference on Cyber Security and Resilience (CSR), 2023.

[C5] V. Bonagura, C. Foglietta, S. Panzieri, F. Pascucci. *Stability Analysis for Multirate Interlaced Kalman Filter.* IEEE Conference on Decision and Control (CDC), 2023.

[C6] C. Foglietta, V. Bonagura, S. Panzieri, M. Rossi, R. Santini, M. Scannapieco, L. Franchina. *The Italian National Security Perimeter for Cyber: Assessing and Modeling Cyber Threats in Interdependent Critical Infrastructures with the CISIApro Simulator.* IFIP International Conference on Critical Infrastructure Protection, 2023.

[C7] V. Bonagura, L. Badia, S. Panzieri, F. Pascucci. *Strategic control against an intruder for timely and accurate updates to a reactive receiver.* European Control Conference (ECC), 2024.

[C8] L. Badia, V. Vadori, E. Grisan, V. Bonagura, F. Pascucci. *Medical self-reporting with adversarial data injection modeled via game theory.* International Conference on Communications, Signal Processing, and their Applications (ICCSPA), 2024.

[C9] V. Bonagura, C. Foglietta, S. Panzieri, F. Pascucci. *Exploiting Interlaced Kalman Filter for Anomaly Detection.* IEEE International Conference on Automation Science and Engineering (CASE), 2024.

[C10] V. Bonagura, L. Badia, S. Panzieri, F. Pascucci. *A Denial of Service Game for Age of Incorrect Information in Cyber-Physical Systems.* IEEE International Conference on Automation Science and Engineering (CASE), 2024.

[C11] V. Bonagura, J. Pisani, G. Cavone, A. Ferrato, C. Foglietta, F. Pascucci. *Machine Learning Techniques for Anomaly Detection in the Hydra Testbed.* International Conference on Critical Information Infrastructures Security (CRITIS), 2024.

[C12] C. Foglietta, V. Bonagura, S. Panzieri, L. Franchina. *Improving Impact Assessment Using Fuzzy Sets in CISIApro 2.0 Model.* International Conference on Critical Information Infrastructures Security (CRITIS), 2024.

[C13] C. Foglietta, V. Bonagura, L. Badia, F. Pascucci, S. Panzieri. *Strategic Interactions in Multi-Sensor Networks Against False Data Injection.* IEEE International Conference on Cyber Security and Resilience (CSR), Crete, Greece, 2025.

[C14] V. Bonagura, L. Badia, C. Foglietta, F. Pascucci, S. Panzieri. *Controlling Age of Incorrect Information Violation Under Data Drift and Strategic Attacks.* 11th International Conference on Control, Decision and Information Technologies (CoDIT), Split, Croatia, 2025.

International Conference Proceedings (Accepted)

[C15] V. Bonagura, G. Cavone, F. Pascucci. *Security by Design with Cost-Constrained Opacity Enforcement for Modbus TCP-Based Industrial Control Systems.* IEEE SMC, Vienna, Austria, 2025. (accepted)

[C16] V. Bonagura, A. Kasis, M. M. Polycarpou, S. Panzieri, F. Pascucci. *Distributed Reconstruction of Sensor Cyber-Attacks in Cyber-Physical Networks.* IEEE Conference on Decision and Control (CDC), Rio, Brazil, 2025. (accepted)

Acknowledgements

I care deeply about thanking all the people who have shaped my journey—both scientifically and personally.

Above all, I want to thank **Prof. Stefano Panzieri** and **Prof. Federica Pascucci** from Roma Tre University for their guidance, trust, and scientific advice throughout these years. Their support and mentorship have been essential to the development of this work and to my growth as a person and as a researcher.

I am also very grateful to **Prof. Lorenzo Badia** and **Prof. Tomaso Erseghe** from the University of Padova for their collaboration and for the many discussions that opened new research directions and perspectives. My thanks also go to **Ing. Camilla Fioravanti** and **Prof. Gabriele Oliva** from Campus Bio-Medico University of Rome for their collaboration and for helping me connect theoretical ideas with practical applications.

A special thank you goes to **Ing. Chiara Foglietta** from San Raffaele Telematic University of Rome. Her clarity, enthusiasm, and constant availability have guided me through every day of this Ph.D. Thank you, Chiara, for your help, patience, and friendship—your presence has made this journey much lighter.

I would also like to express my heartfelt thanks to **Prof. Marios M. Polycarpou** and **Prof. Andreas Kasis** from the University of Cyprus for welcoming me so warmly during my research stay and for their constant generosity, support, and encouragement. Their openness and genuine interest made me feel at home from the very first day, and working with them has been an invaluable learning experience—both scientifically and personally. I am deeply grateful for the time, advice, and kindness they shared throughout my stay in Nicosia.

To all my coauthors, collaborators, and colleagues—thank you for making this a shared experience. Every exchange of ideas, paper, and discussion has contributed to shaping this thesis and to my own development as a researcher.

Beyond the people mentioned above, who directly contributed to my research and publications, I am truly grateful to have met two people who have been fundamental on a personal level. **Elisa Gaetan**, my Ph.D. buddy even from afar—thank you for all the experiences we shared, for the countless conversations, and for being a constant point of comparison in every possible aspect of these three years. Your friendship and humour made this path lighter and much more enjoyable. **Evagoras Makridis**, during my time abroad you were a real point of reference and a wonderful discovery. Your kindness, openness, and generosity made my stay in Nicosia one of the most meaningful and enriching parts of this entire journey.

Introduction

Chapter 1

Introduction

1.1 Context and Motivation

Over the past two decades, the convergence of computation, communication, and control has deeply reshaped industrial and societal systems. Modern infrastructures—such as energy grids, water networks, transport systems, manufacturing plants, and smart cities—are no longer isolated entities. They function as large-scale *Cyber–Physical Systems* (CPS), where digital controllers, sensors, and communication networks interact with physical processes in real time [1], [2]. This integration has enabled high efficiency, automation, and adaptability. However, it has also introduced new fragilities: faults and attacks can now spread across the cyber and physical layers, triggering cascading failures that bypass traditional safety boundaries.

In this interconnected setting, *security* and *resilience* can no longer be treated as separate disciplines. Conventional cybersecurity, focused on data confidentiality and network availability, fails to capture the dynamic link between information and control. Likewise, classical control robustness ignores intelligent adversaries who exploit structure, timing, and feedback to cause instability or misoperation. The merging of these perspectives defines a new domain—*adversarial cyber–physical resilience*—which studies how malicious actions interact with feedback dynamics, information flow, and multi-layer coordination.

The urgency of this shift is evident in the rising number and sophistication of cyber–physical attacks. Incidents such as Stuxnet, the Ukraine power grid blackout, and the Colonial Pipeline ransomware have shown how cyber intrusions can cause real physical and economic harm [3], [4]. These cases reveal a key insight: cyber attacks rarely stay digital. They exploit feedback couplings to corrupt measurements, delay control actions, and destabilize entire systems. As automation and connectivity expand through the Industrial Internet of Things (IIoT), the attack surface grows, challenging existing concepts of safety, reliability, and assurance.

At the systemic level, growing interdependence among Critical Infrastructures (CIs) amplifies these risks. Energy, communication, transportation, and water systems are connected through physical and informational dependencies that carry not only resources but also failures [5], [6]. In such networks, the failure of one node or communication link can cascade across multiple sectors, causing large-scale social and economic impacts. Traditional risk models, which assume independent failures, cannot represent these cross-domain loops or account for intelligent adversaries who deliberately exploit them. Thus, a comprehensive resilience strategy must act across multiple scales—from local control loops to distributed coordination and interdependent infrastructures.

From a methodological viewpoint, this challenge demands a unified analytical framework capturing both cyber–physical dynamics and adversarial behavior. Control theory provides tools for modeling feedback, observability, and stability. Information theory adds quantitative methods for studying data freshness, latency, and uncertainty. Network science offers models of distributed coordination and systemic interdependence. Combining these fields allows security and resilience to be viewed not as static properties, but as emergent features of dynamic systems that sense, decide, and act under uncertainty. This integration also aligns with trends in autonomous and intelligent systems, where learning-enabled controllers and adaptive scheduling blur the boundary between control and computation.

Within this broader context, the main goal of this thesis is to develop a *multi-layer dynamical framework for adversarial resilience* in cyber–physical ecosystems. The work aims to connect the rigor of control theory with the complexity of distributed

infrastructures. At the local level, it develops observer-based detection methods that preserve integrity under deception and denial-of-service attacks. At the network level, it models information freshness as a controllable variable that maintains temporal coherence among agents. At the systemic level, it integrates secure consensus and interdependency modeling to capture how local adversarial events can escalate into large-scale cascading failures. By unifying these mechanisms under a common structure, the framework seeks to transform security from a reactive safeguard into a proactive, adaptive property of system dynamics.

Ultimately, this research addresses the gap between cyber defense and control resilience. Modern infrastructures must remain safe not only under random faults or uncertainty, but also under persistent and intelligent attacks. Achieving this requires systems that can *perceive*, *adapt*, and *learn*—maintaining awareness and stability even when information is incomplete, delayed, or manipulated. This thesis contributes to this goal by formalizing feedback, freshness, and interdependence as the three structural pillars of cyber–physical resilience, and by demonstrating—through analysis and simulation—how their integration enables secure, adaptive, and intelligent defense in adversarial environments.

1.2 Preliminaries and Notation

This section establishes the mathematical language and modeling framework adopted throughout the thesis. Its purpose is twofold: (i) to provide a self–contained reference for all mathematical objects used, avoiding redundant definitions in later chapters, and (ii) to ensure consistent notation across the diverse topics addressed, including estimation, coordination, communication, and systemic analysis. The presentation proceeds from general notation and graph–theoretic preliminaries to multi–agent dynamical models, state observers, measures of information timeliness, consensus mechanisms, and broader formal tools such as evidence theory, holistic modeling perspectives, and game–theoretic reasoning.

1.2.1 Notation and Conventions

Throughout the thesis, we adopt the notation summarized in Table 1.1.

Vectors are denoted by bold lowercase letters (e.g., $\boldsymbol{x}$), matrices by uppercase letters (e.g., A), and sets or spaces by calligraphic symbols (e.g., $\mathcal{X}$). Given a vector $\boldsymbol{p} \in \mathbb{R}^n$, the notation $\texttt{diag}(\boldsymbol{p})$ denotes the $n \times n$ diagonal matrix whose diagonal entries satisfy $\texttt{diag}_{ii}(\boldsymbol{p}) = p_i$. The Euclidean norm is denoted by $\|\cdot\|$. For two matrices $A \in \mathbb{R}^{n\times m}$ and $B \in \mathbb{R}^{p\times q}$, their Kronecker product is denoted by $A \otimes B$.

We denote by $\mathcal{W}$ the set of square $n \times n$ real matrices with zero row sums and nonpositive off-diagonal elements, i.e.,

$$\mathcal{W} = \{W \in \mathbb{R}^{n\times n} \mid W_{ij} \leq 0 \text{ for } i \neq j,\ W\mathbf{1}_n = \mathbf{0}_n\}.$$

Similarly, $\mathcal{W}_S$ denotes the subset of $\mathcal{W}$ containing symmetric and irreducible matrices:

$$\mathcal{W}_S = \{W \in \mathcal{W} \mid W = W^\top,\ W \text{ irreducible}\}.$$

These sets are commonly used to represent Laplacian-like matrices in consensus and networked systems.

Given a differentiable function $g : \mathbb{R}^n \times \mathbb{R} \to \mathbb{R}^n$, we use $\nabla_{\boldsymbol{y}} g(\boldsymbol{y}, t) \in \mathbb{R}^n$ to denote the gradient of $g(\cdot)$ with respect to the first argument, evaluated at $\boldsymbol{x} = \boldsymbol{y}$ and at a given time t. The function $g(\boldsymbol{x}, t)$ is said to be *strongly convex* (see [7]) in $\boldsymbol{x}$ if, for all $t \geq 0$, it satisfies

$$(\nabla_{\boldsymbol{x}} g(\boldsymbol{x}, t) - \nabla_{\boldsymbol{y}} g(\boldsymbol{y}, t))^\top (\boldsymbol{x} - \boldsymbol{y}) \geq \mu \|\boldsymbol{x} - \boldsymbol{y}\|^2, \tag{1.1}$$

for some $\mu > 0$. Equivalently, $g(\boldsymbol{x}, t)$ is strongly convex if and only if its Hessian matrix has eigenvalues lower-bounded by a positive constant.

Table 1.1: Notation Summary

Symbol	Description
x	Scalar quantity
$\boldsymbol{x}$	Column vector (lowercase bold)
M	Matrix (uppercase)
$\mathcal{X}$	Generic set or space
$\mathcal{G} = (\mathcal{V}, \mathcal{E})$	Communication or interaction graph
A_{ij} or $[A]_{i,j}$	(i,j)-th entry of matrix A
$\mathcal{N}_i^{\text{in}}, \mathcal{N}_i^{\text{out}}$	In- and out-neighbor sets of agent i
$\text{col}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N)$	Column vector obtained by stacking $\boldsymbol{x}_i$
$\text{diag}(\boldsymbol{v})$	Diagonal matrix with entries from vector $\boldsymbol{v}$
$\text{blkdiag}(M_1, \ldots, M_N)$	Block-diagonal matrix with diagonal blocks M_i
$\mathbb{R}^n$	n-dimensional real vector space
$\mathbb{N}$	Set of nonnegative integers
$M^\top$	Transpose of M
M^{-1}	Inverse of M, if it exists
$M^{-\top}$	Inverse transpose of M
I_n	Identity matrix of size n
$\mathbf{0}_n$, $\mathbf{1}_n$	All-zero and all-one vectors of dimension n
$\|\boldsymbol{x}\|$	Euclidean norm of vector $\boldsymbol{x}$
$\|M\|$	Induced 2-norm of matrix M
$\|M\|_F$	Frobenius norm of matrix M
$\|\cdot\|$	Absolute value of a scalar
$\succeq, \preceq$	Semidefinite ordering
$\text{tr}(M)$	Trace of matrix M
$\rho(M)$	Spectral radius of M
$\mathbb{E}[\cdot]$	Expectation operator
$\mathcal{N}(\mu, \Sigma)$	Gaussian distribution with mean μ and covariance Σ
$A \otimes B$	Kronecker product of matrices A and B
$A \circ B$	Hadamard (element-wise) product of matrices A and B
e^M	Matrix exponential of M

The notation introduced above is used consistently throughout the thesis to maintain uniformity across the various modelling frameworks—ranging from distributed estimation and control, to information freshness, secure consensus, and systemic interdependency analysis.

1.2.2 Graph–Theoretic Representation

The interaction and communication among agents are described by a directed graph

$$\mathcal{G} = (\mathcal{V}, \mathcal{E}),$$

where $\mathcal{V}$ is the set of nodes (agents) and $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ is the set of edges. An edge $(j,i) \in \mathcal{E}$ means that agent i can receive information from agent j. The set of in–neighbors of agent i is defined as

$$\mathcal{N}_i^{\text{in}} = \{\, j \in \mathcal{V} \mid (j,i) \in \mathcal{E} \,\},$$

and the set of out–neighbors as

$$\mathcal{N}_i^{\text{out}} = \{\, j \in \mathcal{V} \mid (i,j) \in \mathcal{E} \,\}.$$

When $(j,i) \in \mathcal{E}$, we also write $a_{ij} > 0$ to indicate that agent j communicates with agent i. Collecting all communication weights yields the adjacency matrix

$$A = [a_{ij}] \in \mathbb{R}^{N\times N}, \qquad a_{ij} > 0 \text{ if } (j,i) \in \mathcal{E},\ a_{ij} = 0 \text{ otherwise.}$$

The in–degree of node i is $d_i = \sum_{j=1}^{N} a_{ij}$, and the degree matrix is $D = \operatorname{diag}(d_1, \dots, d_N)$. The graph Laplacian is $L_G = D - A$, which satisfies $L_G \mathbf{1}_N = \mathbf{0}_N$ for connected graphs. For undirected communication networks, the adjacency matrix is symmetric, $A = A^\top$, and thus $L_G = L_G^\top \succeq 0$.

The graph distance between two nodes i and j is denoted by $d(i,j)$, representing the minimum number of edges connecting them.

1.2.3 Multi–Agent Dynamical Model

Modern engineered systems increasingly operate as collections of interacting dynamical entities, commonly referred to as *multi–agent systems* (MAS). Each agent represents a self-contained subsystem endowed with sensing, computation, and actuation capabilities, interacting with other agents through communication links or physical couplings. This abstraction captures a broad range of cyber–physical applications, from cooperative robotics and distributed sensing to industrial automation and critical infrastructure management [8], [9], [10], [11], [12]. Within such systems, global behaviors—such as synchronization, consensus, or coordinated control—emerge from local feedback and information exchange among agents, illustrating how complex networked functionality arises from simple interaction rules.

General formulation. Let the network consist of N agents, indexed by the set

$$\mathcal{V} = \{1, 2, \dots, N\}.$$

Each agent $i \in \mathcal{V}$ is described by a local state $\boldsymbol{x}^{(i)}(t) \in \mathbb{R}^{n_i}$, a control input $\boldsymbol{u}^{(i)}(t) \in \mathbb{R}^{m_i}$, and an output $\boldsymbol{y}^{(i)}(t) \in \mathbb{R}^{s_i}$. The dynamics of agent i are governed, in general nonlinear form, by

$$\begin{aligned} \dot{\boldsymbol{x}}^{(i)}(t) &= f^{(i)}\Big(\boldsymbol{x}^{(i)}(t), \{\boldsymbol{x}^{(j)}(t) : j \in \mathcal{N}_i^{\text{in}}\}, \boldsymbol{u}^{(i)}(t), \boldsymbol{w}^{(i)}(t)\Big), \\ \boldsymbol{y}^{(i)}(t) &= h^{(i)}\Big(\boldsymbol{x}^{(i)}(t), \boldsymbol{v}^{(i)}(t)\Big), \end{aligned} \tag{1.2}$$

where:

- $f^{(i)}(\cdot)$ and $h^{(i)}(\cdot)$ denote, respectively, the state-transition and measurement functions of agent i, possibly nonlinear and time-varying;
- $\mathcal{N}_i^{\text{in}}$ represents the set of *in-neighbors* of agent i, i.e., agents whose states influence its dynamics;
- $\boldsymbol{w}^{(i)}(t)$ and $\boldsymbol{v}^{(i)}(t)$ capture process and measurement disturbances.

This formalism naturally extends to discrete time, where interactions evolve at discrete instants $k \in \mathbb{N}$:

$$\begin{aligned} \boldsymbol{x}_{k+1}^{(i)} &= f^{(i)}\Big(\boldsymbol{x}_k^{(i)}, \{\boldsymbol{x}_k^{(j)} : j \in \mathcal{N}_i^{\text{in}}\}, \boldsymbol{u}_k^{(i)}, \boldsymbol{w}_k^{(i)}\Big), \\ \boldsymbol{y}_k^{(i)} &= h^{(i)}\Big(\boldsymbol{x}_k^{(i)}, \boldsymbol{v}_k^{(i)}\Big). \end{aligned} \tag{1.3}$$

The continuous– and discrete–time formulations are structurally equivalent, differing only in the temporal domain over which the coupling unfolds. Such representations have been widely employed in the analysis of cooperative control, distributed estimation, and consensus dynamics [9], [13], [14].

Linearized model. To enable analytical tractability and control design, a linear (or linearized) approximation of (1.2) is typically adopted:

$$\begin{aligned}\dot{\boldsymbol{x}}^{(i)}(t) &= F^{(i,i)}\boldsymbol{x}^{(i)}(t) + \sum_{j\in\mathcal{N}_i^{\text{in}}} F^{(i,j)}\boldsymbol{x}^{(j)}(t) + B^{(i)}\boldsymbol{u}^{(i)}(t) + \boldsymbol{w}^{(i)}(t),\\ \boldsymbol{y}^{(i)}(t) &= H^{(i)}\boldsymbol{x}^{(i)}(t) + \boldsymbol{v}^{(i)}(t),\end{aligned} \tag{1.4}$$

or, in discrete time,

$$\begin{aligned}\boldsymbol{x}_{k+1}^{(i)} &= F^{(i,i)}\boldsymbol{x}_k^{(i)} + \sum_{j\in\mathcal{N}_i^{\text{in}}} F^{(i,j)}\boldsymbol{x}_k^{(j)} + B^{(i)}\boldsymbol{u}_k^{(i)} + \boldsymbol{w}_k^{(i)},\\ \boldsymbol{y}_k^{(i)} &= H^{(i)}\boldsymbol{x}_k^{(i)} + \boldsymbol{v}_k^{(i)}.\end{aligned} \tag{1.5}$$

Here, $F^{(i,i)}$ encodes the internal dynamics of agent i, while $F^{(i,j)}$ quantifies the dynamical influence of neighbor j on agent i through the communication or coupling topology. The matrices $B^{(i)}$ and $H^{(i)}$ represent input and output mappings. In large–scale systems, sparsity of interconnections implies that most $F^{(i,j)}$ are zero, reflecting the underlying network adjacency [8], [10], [15].

Global network representation. By stacking all local states, inputs, and outputs as

$$\boldsymbol{x}(t) = \operatorname{col}\Big(\boldsymbol{x}^{(1)}(t),\dots,\boldsymbol{x}^{(N)}(t)\Big), \quad \boldsymbol{u}(t) = \operatorname{col}\Big(\boldsymbol{u}^{(1)}(t),\dots,\boldsymbol{u}^{(N)}(t)\Big),$$

$$\boldsymbol{y}(t) = \operatorname{col}\Big(\boldsymbol{y}^{(1)}(t),\dots,\boldsymbol{y}^{(N)}(t)\Big),$$

and defining the block matrices

$$F = [F^{(i,j)}]_{i,j=1}^{N}, \quad B = \operatorname{blkdiag}(B^{(1)},\dots,B^{(N)}), \quad H = \operatorname{blkdiag}(H^{(1)},\dots,H^{(N)}),$$

the overall multi–agent dynamics can be compactly written as

$$\dot{\boldsymbol{x}}(t) = F\boldsymbol{x}(t) + B\boldsymbol{u}(t) + \boldsymbol{w}(t), \qquad \boldsymbol{y}(t) = H\boldsymbol{x}(t) + \boldsymbol{v}(t), \tag{1.6}$$

or, in discrete form,

$$\boldsymbol{x}_{k+1} = F\boldsymbol{x}_k + B\boldsymbol{u}_k + \boldsymbol{w}_k, \qquad \boldsymbol{y}_k = H\boldsymbol{x}_k + \boldsymbol{v}_k. \tag{1.7}$$

This compact representation provides the mathematical foundation for distributed estimation, control, and consensus algorithms [12], [16], [17]. In the context of cyber–physical resilience, it enables the explicit modeling of information exchange, delays, and adversarial perturbations across interconnected agents, thus serving as the analytical substrate for the observer–based and freshness–aware coordination frameworks developed in the subsequent chapters.

1.2.4 Consensus in Multi–Agent Systems

A fundamental mechanism enabling coordination and collective intelligence in distributed systems is the *consensus process*. Consensus denotes the capability of a group of agents—each endowed with its own local sensing, computation, and actuation—to reach agreement on specific variables of interest through repeated local interactions. This process underlies a broad range of applications, including distributed estimation, cooperative control, sensor fusion, synchronization, and decision-making in cyber–physical networks [8], [9], [10], [12], [14]. By relying solely on neighborhood communication, consensus mechanisms enable a collection of decentralized entities to act coherently as a single coordinated system, even in the absence of centralized supervision.

Continuous–Time Consensus. Consider a network of N agents represented by a directed graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V} = \{1, 2, \dots, N\}$ is the set of nodes and $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ represents communication links. Each agent $i \in \mathcal{V}$ maintains a local state $\boldsymbol{x}^{(i)}(t) \in \mathbb{R}^{n_i}$, which evolves according to the continuous-time interaction law

$$\dot{\boldsymbol{x}}^{(i)}(t) = \sum_{j \in \mathcal{N}_i^{\text{in}}} a_{ij} \big(\boldsymbol{x}^{(j)}(t) - \boldsymbol{x}^{(i)}(t) \big), \tag{1.8}$$

where $\mathcal{N}_i^{\text{in}}$ denotes the set of in-neighbors of agent i, and $a_{ij} \geq 0$ is the communication weight on edge $(j, i) \in \mathcal{E}$. This interaction drives each agent to adjust its state proportionally to the differences between its value and those of its neighbors, leading to agreement over time.

By stacking all local states into the global vector $\boldsymbol{x}(t) = \text{col}(\boldsymbol{x}^{(1)}(t), \dots, \boldsymbol{x}^{(N)}(t))$, the collective dynamics can be expressed compactly as

$$\dot{\boldsymbol{x}}(t) = -L_G \boldsymbol{x}(t), \tag{1.9}$$

where $L_G = D - A$ is the graph Laplacian, with $A = [a_{ij}]$ the adjacency matrix and $D = \text{diag}\Big(\sum_j a_{ij}\Big)$ the degree matrix. If the graph $\mathcal{G}$ is strongly connected and balanced, the trajectories of (1.9) asymptotically converge to a common value

$$\lim_{t \to \infty} \boldsymbol{x}^{(i)}(t) = \boldsymbol{x}^\star, \quad \forall i \in \mathcal{V},$$

representing global agreement among all agents [8], [9].

Discrete–Time Consensus. In practical cyber–physical implementations, communication and control occur over digital channels, leading to a discrete-time formulation of the consensus process. A standard discrete update rule is given by

$$\boldsymbol{x}_{k+1}^{(i)} = \boldsymbol{x}_k^{(i)} + \varepsilon \sum_{j \in \mathcal{N}_i^{\text{in}}} a_{ij} \big(\boldsymbol{x}_k^{(j)} - \boldsymbol{x}_k^{(i)} \big), \tag{1.10}$$

where $\varepsilon > 0$ is the update gain or step size. In vector form, the network dynamics become

$$\boldsymbol{x}_{k+1} = (I - \varepsilon L_G) \boldsymbol{x}_k. \tag{1.11}$$

For connected graphs, consensus is achieved when $0 < \varepsilon < 1/\Delta_{\max}$, with $\Delta_{\max}$ denoting the maximum node degree [13], [14]. The resulting equilibrium corresponds to a weighted average of the agents' initial states, determined by the structure of $\mathcal{G}$ and the choice of ε.

Secure and Resilient Consensus. In realistic operating environments, multi–agent systems face a variety of non-idealities, including communication delays, packet losses, measurement noise, and intentional disruptions such as false data injection, denial-of-service (DoS), or replay attacks. These phenomena can corrupt inter-agent communication and compromise the convergence of (1.11). *Secure and resilient consensus* seeks to ensure convergence and stability even when a subset of agents or communication channels behaves anomalously or maliciously [10], [12], [15], [18]. Typical approaches embed validation, reweighting, or isolation mechanisms directly into the local update rules—so that each agent can autonomously mitigate the effect of corrupted or inconsistent information. The W-MSR (Weighted Mean-Subsequence Reduced) algorithm [10] and its extensions provide a foundational framework for resilient agreement under bounded adversarial influence, ensuring that normal agents converge despite the presence of misbehaving nodes. Recent advances extend this resilience to dynamic and time-varying topologies, delay-aware consensus [12], and event-triggered communication schemes that jointly preserve stability and bandwidth efficiency [15], [19]. Collectively, these results formalize consensus as a feedback mechanism robust not only to stochastic disturbances but also to strategic, adversarial perturbations.

Privacy–Preserving Consensus. Beyond resilience, modern multi–agent systems increasingly demand protection of private or proprietary information during coordination. *Privacy-preserving consensus* addresses this by enabling agents to compute collective quantities—such as averages or estimates—without exposing their individual states [20], [21], [22], [23], [24]. Differential privacy approaches introduce carefully calibrated noise to obfuscate individual contributions while guaranteeing statistical privacy [21]. Alternatively, cryptographic and homomorphic techniques [22], [25], [26] allow arithmetic operations to be performed directly on encrypted data, achieving exact consensus without revealing local information. Lightweight approaches such as chaotic or state-decomposition encryption [27], [28] offer computationally efficient alternatives for embedded cyber–physical agents. These methods collectively strive to balance two competing objectives: (i) maintaining accuracy and convergence speed, and (ii) preserving the confidentiality of local data, a trade-off that is central to cooperative industrial and IoT applications [29], [30].

Summary. In summary, consensus constitutes the fundamental coordination primitive in multi–agent cyber–physical systems, translating local information exchange into emergent collective behavior. Secure and privacy-aware variants extend this mechanism to adversarial and untrusted environments, embedding resilience and confidentiality directly into the feedback dynamics of distributed control. Together, these developments enable cyber–physical ecosystems to operate coherently, securely, and autonomously across scales—from sensor networks and robotic teams to interdependent critical infrastructures.

1.2.5 State Observers

In many engineering and control applications, the full state of a dynamical system cannot be directly measured due to sensing limitations, noise, or communication constraints. *State observers* are estimation algorithms designed to reconstruct the internal state of a system using available inputs and output measurements, possibly affected by uncertainties or disturbances. They play a crucial role in monitoring, control, and fault detection, enabling real-time supervision and decision-making even in partially observable environments [31], [32], [33]. Depending on the assumptions made about the system and noise characteristics, observers can be classified into stochastic (probabilistic) and deterministic (robust) families. This subsection briefly introduces both perspectives and outlines their distributed extensions for large-scale systems.

From the standpoint of *Observer-Based attack detection*, state observers are particularly valuable because they provide an analytical link between the system's expected behavior—predicted from the mathematical model—and its actual observed outputs [34], [35]. By continuously comparing estimated outputs with real sensor data, an observer allows the computation of a *residual signal*, that is, the discrepancy between measured and predicted values. Under nominal conditions, this residual remains small and statistically consistent with process and measurement noise. However, when a fault, malfunction, or malicious attack perturbs the system, the residual exhibits an abnormal deviation, which can be used as an indicator of anomalous behavior [36], [37]. This mechanism forms the foundation of Observer-Based detection: rather than relying solely on data-driven patterns, it exploits physical consistency enforced by the underlying system dynamics [38], [39].

Moreover, state observers naturally support both *attack detection* and *attack reconstruction*. In addition to signaling the presence of an anomaly, appropriately designed observers can isolate which state or measurement channel is affected, and under certain structural conditions, even estimate the injected attack signal itself [40], [41], [42]. This capability is essential in cyber–physical systems, where distinguishing between benign disturbances and intentional manipulations is critical for resilience [43], [44]. In distributed settings, networks of observers allow local detection and estimation while sharing limited information among neighboring nodes [12], [15], thus ensuring scalability and fault tolerance without central coordination.

In this context, observers provide the theoretical and computational foundation for a systematic Observer-Based security framework—one that unifies monitoring, estimation, and diagnosis within a consistent dynamical model of the system [38], [45].

Kalman Filtering and Distributed Stochastic Estimation

The Kalman filter represents the cornerstone of stochastic estimation for linear systems subject to Gaussian noise [31], [46]. Consider a discrete-time system described by (1.7), where $\boldsymbol{w}_k \sim \mathcal{N}(0, Q_k)$ and $\boldsymbol{v}_k \sim \mathcal{N}(0, R_k)$ are mutually uncorrelated zero-mean Gaussian processes. The Kalman filter recursively computes the optimal (minimum variance) estimate of the system state given all measurements up to time k, through two main phases [16], [17]:

Prediction:

$$\begin{aligned} \hat{\boldsymbol{x}}_{k|k-1} &= F_{k-1}\hat{\boldsymbol{x}}_{k-1|k-1} + B_{k-1}\boldsymbol{u}_{k-1}, \\ P_{k|k-1} &= F_{k-1}P_{k-1|k-1}F_{k-1}^{\top} + Q_{k-1}. \end{aligned} \tag{1.12}$$

Correction:

$$\begin{aligned} K_k &= P_{k|k-1}H_k^{\top}(H_k P_{k|k-1}H_k^{\top} + R_k)^{-1}, \\ \hat{\boldsymbol{x}}_{k|k} &= \hat{\boldsymbol{x}}_{k|k-1} + K_k\big(\boldsymbol{y}_k - H_k\hat{\boldsymbol{x}}_{k|k-1}\big), \\ P_{k|k} &= (I - K_k H_k)P_{k|k-1}. \end{aligned} \tag{1.13}$$

Here, $\hat{\boldsymbol{x}}_{k|k}$ denotes the estimated state, $P_{k|k}$ its error covariance, and K_k the Kalman gain weighting model and measurement confidence. When system parameters are constant, these equations converge to a steady-state solution characterized by the algebraic Riccati equation [33].

For large-scale or networked systems, centralized implementation of the Kalman filter is often impractical due to communication overhead and the lack of a global data fusion point [12]. This has motivated the development of *distributed Kalman filters*, in which each agent maintains a local state estimate and exchanges partial information with its neighbors. Distributed implementations can rely on consensus-based fusion, covariance intersection, or diffusion strategies to achieve approximate global optimality without centralized coordination [12], [15], [17]. Such filters enable scalable and resilient estimation in multi-agent and cyber–physical environments where data sharing is constrained or intermittent [45], [47].

Sliding Mode and Robust Deterministic Observers

While stochastic observers assume probabilistic models of uncertainty, many real-world systems are affected by unstructured disturbances, abrupt faults, or unknown inputs that violate Gaussian assumptions. In such cases, *sliding mode observers* (SMOs) offer a robust alternative [40], [41], [48]. They belong to the class of variable-structure systems and use discontinuous correction terms to enforce insensitivity of the estimation error to certain disturbances.

Consider the continuous-time linear system

$$\begin{aligned} \dot{\boldsymbol{x}}(t) &= A\boldsymbol{x}(t) + B\boldsymbol{u}(t) + E\boldsymbol{d}(t), \\ \boldsymbol{y}(t) &= C\boldsymbol{x}(t), \end{aligned} \tag{1.14}$$

where $\boldsymbol{d}(t)$ represents an unknown disturbance, fault, or attack signal. A generic sliding mode observer can be written as

$$\dot{\hat{\boldsymbol{x}}}(t) = A\hat{\boldsymbol{x}}(t) + B\boldsymbol{u}(t) + L(\boldsymbol{y}(t) - \hat{\boldsymbol{y}}(t)) + G\,\mathrm{sgn}(\boldsymbol{y}(t) - \hat{\boldsymbol{y}}(t))\,, \tag{1.15}$$

where $\hat{\boldsymbol{y}}(t) = C\hat{\boldsymbol{x}}(t)$ is the estimated output, L is a linear feedback gain ensuring nominal convergence, and G scales the discontinuous correction term. The discontinuity $\mathrm{sgn}(\cdot)$

enforces a sliding motion of the estimation error on a predefined manifold, ensuring robustness against matched uncertainties and unknown inputs [15], [42].

Sliding mode observers are deterministic by nature and are particularly suited to safety-critical or adversarial contexts, where it is necessary to reconstruct both the state and potential faults or attacks affecting the system [40], [44]. They can achieve finite-time convergence and strong disturbance rejection, albeit at the cost of possible chattering phenomena, which are mitigated in practice through boundary layer or adaptive schemes [48], [49].

In distributed settings, robust observers can be extended by allowing each agent to estimate its local state based on local measurements and limited information from its neighbors [15], [45]. Such *distributed robust observers* blend principles from sliding mode control and consensus theory, providing networked estimation strategies capable of maintaining accuracy and stability under uncertainty, communication delays, or partial data availability [12], [33].

1.2.6 Game Theory

In many engineered, economic, and cyber–physical systems, multiple autonomous decision-makers interact within a shared environment, each pursuing objectives that may be cooperative, competitive, or even adversarial. The behavior of such systems cannot be fully understood by analyzing agents in isolation, since the outcome of each agent's decision depends on the simultaneous actions of others. *Game theory* provides the mathematical foundation for describing and analyzing these strategic interactions [18], [50], [51]. It offers a structured framework for reasoning about decision-making, stability, and equilibrium in distributed settings where information, incentives, and goals are interdependent.

Basic Concepts. A game is defined by three main elements: (i) a set of players $\mathcal{P} = \{1, 2, \dots, N\}$; (ii) a set of admissible strategies $\mathcal{S}_i$ available to each player i; and (iii) a utility or payoff function

$$u_i : \mathcal{S}_1 \times \cdots \times \mathcal{S}_N \to \mathbb{R},$$

which quantifies the reward (or cost) that player i receives for a given combination of strategies. A strategy profile $\boldsymbol{s} = (s_1, \dots, s_N)$ represents the simultaneous choice of all players. Each player seeks to maximize its own utility, anticipating how others will act—thus, interdependence and strategic reasoning are inherent to the model [18].

Nash Equilibrium. The central concept in non–cooperative game theory is the *Nash equilibrium*, representing a steady state of mutual best response. A strategy profile $\boldsymbol{s}^\star$ is a Nash equilibrium if no player can unilaterally improve its utility:

$$u_i(s_i^\star, s_{-i}^\star) \geq u_i(s_i, s_{-i}^\star), \qquad \forall s_i \in \mathcal{S}_i,\ \forall i \in \mathcal{P}, \tag{1.16}$$

where $s_{-i}^\star$ denotes the equilibrium strategies of all players except i. At equilibrium, all decisions are self-consistent: given the strategies of others, each player's choice is optimal. Depending on the game, equilibria may be unique, multiple, or non-existent in pure strategies, in which case *mixed strategies*—probabilistic distributions over actions—are used. The existence of at least one mixed-strategy Nash equilibrium in finite games is guaranteed by Nash's theorem [50]. In cyber–physical contexts, Nash equilibria describe operating points where competing agents—defenders, attackers, or autonomous controllers—reach a balance of incentives, often interpreted as conditions of adversarial stability [52].

Bayesian Games and Bayesian Nash Equilibrium. In many realistic scenarios, players do not possess perfect information about others' preferences, costs, or capabilities. Such situations are captured by *Bayesian games*, where each player has a private *type* θ_i drawn from a known probability distribution over a type space Θ_i. The type represents private

information—such as risk tolerance, available resources, or system configuration—that influences the player's utility function $u_i(s_1, \dots, s_N; \theta_i, \theta_{-i})$. Players form beliefs about the types of others and choose strategies accordingly [52].

A *Bayesian Nash equilibrium* is a set of strategies $\{s_i^\star(\theta_i)\}_{i=1}^N$ such that each player maximizes its expected utility given its beliefs about the others' types:

$$\mathbb{E}_{\theta_{-i}}\left[u_i(s_i^\star(\theta_i), s_{-i}^\star(\theta_{-i}); \theta_i, \theta_{-i}) \,\middle|\, \theta_i\right] \geq \mathbb{E}_{\theta_{-i}}\left[u_i(s_i, s_{-i}^\star(\theta_{-i}); \theta_i, \theta_{-i}) \,\middle|\, \theta_i\right], \tag{1.17}$$

for all $s_i \in \mathcal{S}_i$, $\forall \theta_i \in \Theta_i$. At a Bayesian Nash equilibrium, each agent plays optimally with respect to its own private information and probabilistic expectations about the others. This concept is particularly relevant in adversarial and security scenarios, where defenders and attackers make decisions under asymmetric and uncertain information [50], [51]. It thus provides a rigorous framework for analyzing strategic uncertainty in cyber–physical and communication systems.

Classes of Games. Game-theoretic models can be categorized according to interaction structure, time scale, and information availability:

- *Cooperative vs. Non–Cooperative Games:* Cooperative games allow agents to form coalitions and share payoffs based on joint strategies, emphasizing collaboration [18]. Non–cooperative games model independent strategic actions, typically analyzed through Nash or Bayesian Nash equilibria [50].
- *Static vs. Dynamic Games:* Static (one-shot) games capture simultaneous decisions, while dynamic games model interactions evolving over time, often governed by system dynamics or learning processes [52]. Dynamic formulations are particularly useful in cyber–physical and communication systems, where feedback, delay, and adaptation shape strategic evolution.
- *Zero–Sum vs. General–Sum Games:* In zero–sum games, one player's gain equals another's loss, a structure often used in security and attack–defense modeling [50]. General–sum games capture more complex interactions where competition and cooperation coexist, as in distributed resource allocation or consensus [18].
- *Complete vs. Incomplete Information:* Complete-information games assume full knowledge of all utilities and strategies, while incomplete-information games (such as Bayesian games) model uncertainty through beliefs or probability distributions [52].

Game theory provides a unifying framework for understanding interaction, cooperation, and competition in complex systems. It enables the analysis of distributed decision-making, resource allocation, and negotiation in multi-agent settings, as well as the modeling of adversarial or uncertain environments [18], [51]. Applications range from economic markets and networked control to communication systems, social networks, and cybersecurity [50]. By identifying equilibrium strategies and quantifying trade-offs among agents, game-theoretic reasoning supports the design of algorithms and architectures that are both efficient and resilient in the presence of strategic behavior.

From a control perspective, game theory generalizes classical optimal control by introducing multiple autonomous agents whose decisions jointly determine system evolution [52]. While decision theory focuses on optimization under uncertainty for a single decision-maker, game theory embeds strategic interdependence and potential conflict among several. In dynamic settings, the resulting feedback equilibria naturally extend principles of stability, optimality, and adaptation to distributed and interactive contexts [18]. This synthesis between control and game theory is particularly relevant for modern distributed and cyber–physical systems, where autonomy, limited information, and strategic interaction coexist within complex and uncertain environments [50], [51].

1.2.7 Information Freshness

In distributed and networked control systems, decision-making and estimation rely not only on the *accuracy* of the available information but also on its *timeliness*. Delays, packet

losses, and asynchronous updates in communication networks can cause the information used by controllers or estimators to become stale, degrading performance and, in extreme cases, leading to instability [12], [53]. To explicitly quantify and analyze the temporal relevance of data in such systems, the concept of *information freshness* has been introduced as a unifying metric for the quality of information over time. It captures how recent or outdated the data available at a given node is with respect to the true system state, bridging the gap between communication theory and control performance [52], [54].

Age of Information (AoI). The *Age of Information* (AoI) is a fundamental measure of information freshness [54]. It quantifies, at any given time, how long it has been since the most recently received update was generated at the source. Consider an information source j that periodically produces timestamped updates $\{s_k^{(j)}\}$ and transmits them over a network to a receiving node or agent i. If $s_{ij}(t)$ denotes the generation time of the most recently received update at time t, the instantaneous AoI is defined as:

$$\Delta_{ij}(t) = t - s_{ij}(t). \tag{1.18}$$

The AoI increases linearly with time as data ages, and it resets (or drops) whenever a new update is successfully received. It thus captures both communication delays and inter-arrival times of updates in a single, time-varying quantity [53], [55].

In a discrete-time formulation, communication between nodes i and j can be described by:

$$\Delta_{ij}[k+1] = \begin{cases} 0, & \text{if a packet from } j \text{ is received by } i \text{ at time } k+1, \\ \Delta_{ij}[k] + 1, & \text{otherwise.} \end{cases} \tag{1.19}$$

The average AoI, denoted $\bar{\Delta}_{ij}$, provides an aggregate measure of how up-to-date the information held by i about j remains over time [54]. Smaller AoI values indicate that data is fresh and quickly updated, whereas larger values reflect outdated or delayed information, which can degrade estimation and control accuracy [12], [56].

Age of Incorrect Information (AoII). Although AoI effectively measures temporal staleness, it does not account for the *fidelity* of the information—whether it correctly represents the current system state. In practical scenarios, especially in estimation or monitoring tasks, timeliness and accuracy must be considered jointly. To this end, the *Age of Incorrect Information* (AoII) extends the AoI framework by introducing an error-dependent component that quantifies both delay and informational deviation [52].

For a true system state $\boldsymbol{x}_k^{(j)}$ and its estimate $\hat{x}_k^{(j|i)}$ available at node i, the AoII can be expressed as a function:

$$\Delta_{ij}^{\text{AoII}}[k] = f\Big(\Delta_{ij}[k], \|\boldsymbol{x}_k^{(j)} - \hat{x}_k^{(j|i)}\|\Big), \tag{1.20}$$

where $f(\cdot)$ is a non-decreasing function that penalizes both the age of the information and the estimation error. A common and intuitive formulation is [52]:

$$\Delta_{ij}^{\text{AoII}}[k] = \Delta_{ij}[k] \cdot \|\boldsymbol{x}_k^{(j)} - \hat{x}_k^{(j|i)}\|^2,$$

which jointly captures how estimation inaccuracy compounds with information staleness. In this way, AoII represents a richer and more realistic measure of informational quality, reflecting both the temporal delay and the correctness of the data.

In distributed estimation, for example, the AoI can guide update scheduling policies, ensuring that agents prioritize sending information that most improves global situational awareness [12], [53]. In control systems, AoI-aware or AoII-aware strategies can adapt sampling, communication, or control rates to maintain desired performance under varying network conditions [56]. Such formulations are particularly relevant in adversarial settings,

where attackers can intentionally increase information age through denial-of-service or delay injection, thereby degrading estimation and situational awareness [57], [58].

More broadly, information freshness bridges communication theory and control theory, establishing a quantitative link between the dynamics of information exchange and the achievable accuracy and robustness of networked decision-making [52]. It thus serves as a foundation for developing control and estimation architectures that are not only accurate but also temporally aware, adaptive, and resilient in the face of real-world communication constraints [12], [53], [56].

1.2.8 Evidence Theory

In many real-world systems—particularly those involving complex interactions, incomplete data, or expert judgment—uncertainty cannot be fully captured through classical probability theory. Traditional probabilistic models assume complete knowledge of likelihood distributions, which is often unrealistic when information is scarce, conflicting, or only partially reliable. In such cases, *Evidence Theory*, also known as the *Dempster–Shafer Theory of Belief Functions*, provides a powerful alternative framework for representing and reasoning under uncertainty [59]. It enables the explicit modeling of ignorance, partial knowledge, and conflicting evidence, offering a richer description of epistemic uncertainty compared to purely probabilistic approaches [60].

Frame of Discernment. The starting point of evidence theory is the *frame of discernment*, denoted as

$$\Theta = \{\theta_1, \theta_2, \ldots, \theta_n\},$$

which represents the finite set of all possible and mutually exclusive hypotheses concerning the state of a system or the occurrence of an event. While probability theory assigns likelihoods directly to single hypotheses, evidence theory allows assigning degrees of belief to *subsets* of Θ, reflecting uncertainty about which specific hypothesis within a subset is true [61]. This key feature enables the expression of imprecise knowledge—for example, when evidence supports that the true state lies somewhere within a subset of hypotheses, but without specifying exactly which one.

Basic Probability Assignment (BPA). A *basic probability assignment* (BPA) is a function

$$m : 2^{\Theta} \to [0, 1], \tag{1.21}$$

that satisfies

$$\sum_{A \subseteq \Theta} m(A) = 1, \qquad m(\emptyset) = 0.$$

Each subset $A \subseteq \Theta$ such that $m(A) > 0$ is called a *focal element*, representing a proposition supported by some available evidence [11]. The value $m(A)$ quantifies the amount of belief assigned *exactly* to A, expressing the confidence that the true hypothesis belongs to A without being able to specify a more precise subset. Hence, larger focal elements indicate a greater degree of uncertainty or vagueness in the available information.

Belief and Plausibility Functions. From the BPA, two fundamental functions—*belief* and *plausibility*—are derived to represent lower and upper bounds on the confidence that the true state belongs to a given subset $A \subseteq \Theta$:

$$\begin{aligned} \mathrm{Bel}(A) &= \sum_{B \subseteq A} m(B), \\ \mathrm{Pl}(A) &= \sum_{B \cap A \neq \emptyset} m(B). \end{aligned} \tag{1.22}$$

The *belief* $\mathrm{Bel}(A)$ represents the total support that evidence provides directly in favor of A, while the *plausibility* $\mathrm{Pl}(A)$ measures the degree to which the evidence does not

contradict A. The interval $[\mathrm{Bel}(A), \mathrm{Pl}(A)]$ therefore captures the range of confidence that the true state lies within A, explicitly modeling both certainty and uncertainty [60]. When $\mathrm{Bel}(A) = \mathrm{Pl}(A)$ for all subsets A, the model reduces to a conventional probability distribution.

Combination of Evidence. In practice, multiple information sources—such as sensors, models, or expert assessments—may provide distinct pieces of evidence over the same frame Θ. Evidence theory provides a systematic way to combine them through *Dempster's rule of combination*. Given two BPAs m_1 and m_2, representing independent sources of evidence, their combined BPA $m = m_1 \oplus m_2$ is defined as:

$$m(A) = \frac{1}{1-K} \sum_{B \cap C = A} m_1(B)\, m_2(C), \tag{1.23}$$

where

$$K = \sum_{B \cap C = \emptyset} m_1(B)\, m_2(C)$$

is the *conflict coefficient*, quantifying the degree of inconsistency between the two evidence sources. This rule fuses information by reinforcing mutually consistent beliefs and discounting conflicting ones, yielding a normalized and coherent representation of combined knowledge [11], [18].

Relation to Possibility and Fuzzy Frameworks. Evidence theory generalizes several other models of uncertainty, including fuzzy logic and possibility theory. When focal elements are nested subsets of Θ, the resulting belief and plausibility functions can be expressed in terms of a *possibility distribution* $\pi(\theta)$ satisfying:

$$\mathrm{Pl}(A) = \max_{\theta \in A} \pi(\theta).$$

This correspondence allows hybrid formulations that combine probabilistic, possibilistic, and evidential reasoning—particularly useful when handling both quantitative data and qualitative or linguistic assessments [59]. In the context of cyber–physical resilience analysis, such hybrid models have been used in agent-based simulators like CISIApro to capture epistemic and aleatoric uncertainty across interdependent infrastructures [60]. Similarly, in distributed trust and consensus applications, evidence-theoretic fusion provides a principled mechanism for combining heterogeneous and uncertain information among agents while maintaining robustness to conflicting or deceptive sources [11], [18].

1.2.9 Holistic and Reductionistic Approaches

The analysis and design of complex engineering systems can follow two complementary paradigms: the *reductionistic* and the *holistic* approach. Understanding their conceptual differences is essential for properly modeling modern cyber–physical systems (CPSs) and large-scale industrial infrastructures, which combine physical, digital, and human components interacting across multiple scales [59], [60], [62].

Reductionistic Perspective. The reductionistic approach seeks to understand a system by decomposing it into its elementary components and studying the behavior of each part in isolation. This methodology has been the foundation of classical engineering and control theory, where the global behavior is reconstructed by aggregating analytical models of subsystems. In control design, for instance, reductionism manifests through linearization, model abstraction, and modular decomposition, allowing tractable analysis and optimization.

Reductionism provides mathematical clarity, scalability of computation, and interpretability. However, it inherently assumes that the properties of the whole system can be fully inferred from the properties of its parts and their pairwise interactions. In

highly interconnected CPSs, this assumption may fail: emergent behaviors, nonlinear dependencies, and cross-domain feedback loops can produce global effects that cannot be captured by examining components in isolation [59], [60].

Holistic Perspective. The holistic approach, by contrast, emphasizes the system as an integrated entity whose properties arise from the interaction of its components and their environment. It focuses on emergent dynamics, collective patterns, and systemic resilience rather than on isolated performance metrics. In the context of networked control or infrastructure interdependency modeling, holistic thinking leads to the study of aggregate phenomena—such as synchronization, cascading failures, or cooperative stability—that emerge only at the system level [60], [62].

A holistic model often integrates multiple domains (cyber, physical, organizational) and leverages distributed or multi-layered representations. Rather than seeking detailed accuracy for each component, it aims to preserve structural and causal relationships that drive collective behavior. This perspective is indispensable when analyzing resilience, as vulnerabilities often emerge from global interconnections rather than from individual subsystem faults [59], [60].

Toward a Mixed Approach. In practice, neither a purely reductionistic nor a purely holistic view is sufficient for engineering complex CPSs. A *mixed* approach combines the analytical rigor of reductionism with the systemic awareness of holism. At the local level, reductionistic models enable precise estimation, control, and attack detection for individual subsystems or agents [11], [18]. At the global level, holistic models capture the propagation of failures, dependencies, and collective responses that shape overall system resilience [60].

In this thesis, both perspectives are systematically integrated. Reductionistic methods underpin the development of distributed estimators and control algorithms—each operating with localized models and limited information—while holistic simulation frameworks provide a systemic view of interdependencies and cascading risks across interconnected infrastructures [59]. This dual paradigm supports the broader objective of designing distributed yet coordinated security architectures that remain scalable, interpretable, and robust in the face of uncertainty and adversarial perturbations.

1.3 State of the Art

This section reviews the evolution of adversarial modeling and resilience analysis across the different scales that compose a modern cyber–physical ecosystem. From single control loops to networked multi-agent coordination and large-scale interdependent infrastructures, the literature converges toward a unified vision: *security must be formulated as a dynamical property, embedded in feedback, information flow, and interdependence.* The following sections progressively build this perspective, showing how control-theoretic, information-theoretic, and systemic approaches form the conceptual and methodological foundation of the adversarial framework developed in this thesis.

1.3.1 Control-Theoretic Foundations of CPS Security

Cyber–Physical Systems integrate computation, communication, and control within closed feedback architectures where information and energy co-evolve [1], [2]. In these systems, sensors measure physical states, controllers compute corrective actions, and networked communication ensures coordination across spatially distributed subsystems [63]. Because cyber and physical components are interdependent, even localized anomalies—packet losses, signal tampering, or latency—propagate through feedback loops and affect stability and safety [31], [32]. Consequently, CPS security must be analyzed as a *dynamical* problem rather than a purely logical one.

The most general control-theoretic representation of a CPS describes the continuous interaction between the physical process and the cyber layer through nonlinear dynamics:

$$\dot{\mathbf{x}}(t) = f\big(\mathbf{x}(t), \mathbf{u}(t)\big) + \mathbf{d}(t), \tag{1.24}$$

$$\mathbf{y}(t) = h\big(\mathbf{x}(t)\big) + \mathbf{v}(t), \tag{1.25}$$

where $\mathbf{x}(t)$ denotes the physical state, $\mathbf{u}(t)$ the control input, and $\mathbf{y}(t)$ the measured output. The functions $f(\cdot)$ and $h(\cdot)$ describe, respectively, the system's internal dynamics and sensing process, while $\mathbf{d}(t)$ and $\mathbf{v}(t)$ account for process disturbances and measurement noise. This nonlinear and often hybrid formulation is necessary to capture the coexistence of continuous-time physical processes with discrete-time control and communication events [64]. Sampling, quantization, and asynchronous updates introduce nonlinearities and discontinuities that couple cyber phenomena with the underlying plant, making it increasingly difficult to separate computational and physical effects [36], [43].

Although such a nonlinear description reflects the true complexity of CPS operation, it is often intractable for rigorous analysis of security and resilience. To enable analytical treatment, a discrete-time linear approximation around an operating point is typically adopted:

$$\mathbf{x}_{k+1} = F\mathbf{x}_k + B\mathbf{u}_k + \mathbf{w}_k, \quad \mathbf{y}_k = H\mathbf{x}_k + \mathbf{v}_k, \tag{1.26}$$

where $\mathbf{w}_k$ and $\mathbf{v}_k$ represent bounded process and measurement noise. This linear time-invariant (LTI) formalism preserves the essential feedback structure of the CPS while allowing tractable reasoning about key system-theoretic properties such as stability, observability, detectability, and controllability [33], [38], [39], [42], [43], [44], [45], [65]. Linearization does not diminish the generality of the security analysis: most adversarial effects—such as corrupted measurements, packet drops, or delays—manifest as structured perturbations to the nominal feedback dynamics, which can be modeled and studied effectively in the LTI setting.

Within the classical CIA triad, confidentiality is typically addressed through cryptographic mechanisms, whereas *integrity* and *availability* violations are those that directly alter the dynamical evolution of the system. These classes of attacks, which modify sensor data, control actions, or their timing, are naturally embedded in the state-space formalism by treating them as additional dynamical inputs. This perspective establishes a direct bridge between control theory and security analysis, enabling the characterization of adversarial behavior in terms of system properties and laying the groundwork for unified modeling and detection frameworks.

1.3.2 Dynamical Modeling of Adversarial Behavior

The control-theoretic formulation of cyber–physical systems offers a natural language to describe adversarial behavior through its effects on the feedback signals that sustain closed-loop operation. Rather than being characterized by the specific intrusion vector—such as malware, firmware manipulation, or protocol exploitation—the attack is modeled in terms of its *functional impact* on information flow and actuation. This abstraction, first formalized in [66], captures malicious interference as an additional input to the system dynamics, directly linking it to structural properties such as stability, observability, and controllability. By describing cyberattacks as external excitations of the system model, heterogeneous mechanisms of disruption can be analyzed within a unified dynamical framework.

In discrete time, let $\mathbf{u}_k$ and $\mathbf{y}_k$ denote, respectively, the nominal control and measurement signals at instant k. An adversary can alter their *content* (data integrity) or their *timing* (information availability and freshness). Denoting by $\mathbf{u}_k^a$ and $\mathbf{y}_k^a$ the compromised signals, a compact representation reads:

$$\mathbf{u}_k^a = \delta_k^u \mathbf{u}_k + \mathbf{a}_k^u, \qquad \mathbf{y}_k^a = \delta_k^y \mathbf{y}_k + \mathbf{a}_k^y, \tag{1.27}$$

where $\delta_k^u, \delta_k^y \in \{0,1\}$ model packet delivery (availability), and $\mathbf{a}_k^u, \mathbf{a}_k^y$ encode integrity violations. This formulation unifies integrity and availability attacks within a single analytical description, while naturally extending to include temporal effects such as induced delays and replay phenomena.

Integrity Attacks. Integrity attacks act on the content of transmitted data while preserving their apparent timing, aiming to mislead estimation or control while remaining undetected. They include false-data injection, replay, bias, and stealthy manipulations [32], [34], [35]. The notion of *zero-dynamics attacks* introduced in [34] showed that if injected signals excite only the unobservable modes of (F, H), the attack becomes completely stealthy to residual-based detectors. This concept was extended in [35] to distributed settings, showing how coordinated bias injections across communication channels can preserve stealth over time. Further analyses in [32], [36] confirmed that such attacks exploit intrinsic system-theoretic vulnerabilities associated with unobservable subspaces.

Effective deception typically requires partial or full knowledge of the plant model and feedback structure. To counteract this, watermarking-based defense schemes inject small, random perturbations into control commands to decorrelate nominal and replayed data, so that falsified measurements disrupt the expected input–output relationship [67], [68]. Other countermeasures rely on robust observers [38], [39], sliding-mode filters [15], [40], or sparse estimation frameworks [17], [44], all designed to reconstruct the true state under partial corruption. Within the model in (1.27), deception attacks correspond to nonzero $\mathbf{a}_k^u$ and $\mathbf{a}_k^y$ with $\delta_k^u = \delta_k^y = 1$, modifying signal content but not availability.

Availability Attacks and Temporal Degradation. Availability attacks target the temporal structure of communication, disrupting the timing and synchronization of feedback loops. They operate by suppressing, delaying, or reordering transmissions, forcing the controller or estimator to operate in open loop during information gaps. In (1.27), these effects appear as intermittent zeros in δ_k^u, δ_k^y, corresponding to missing packets or blocked updates. Classical analyses in [55], [69] quantify stability degradation under stochastic packet loss, deriving explicit dwell-time and loss-rate conditions for boundedness. More recent results reveal that even short, strategically timed DoS bursts can destabilize marginally stable systems or induce large transient errors [45], [56]. In addition to simple dropouts, adversaries may introduce controlled *communication delays*—by buffering packets, altering timestamps, or exploiting routing dynamics—so that data arrive correctly but too late to be useful. This class of timing-based attacks includes *delay injection*, *time-shifted replay*, and *queue saturation* [43], [57], which degrade synchronization and effectively age the information available to estimators and controllers.

The notion of informational *age*—formalized as the *Age of Information* (AoI) [54]—provides a quantitative measure of this temporal degradation. AoI defines how outdated an agent's knowledge is with respect to the true system state, capturing not only losses but also delay-induced obsolescence. Under DoS and jamming, the AoI of sensor data grows as updates are delayed or dropped, forcing the controller to rely on increasingly stale estimates. Even when no packets are falsified, excessive AoI can render the control loop effectively blind, as the system operates on temporally inconsistent information. Recent extensions, such as the *Age of Incorrect Information* (AoII) framework [52], explicitly combine timing and correctness, recognizing that attacks can produce data that are syntactically valid but semantically outdated—an effect particularly severe under replay or delay injection.

Availability attacks therefore encompass not only outright denial of data but also the subtler manipulation of their arrival times, producing *temporal desynchronization*. This temporal distortion may appear benign from a communication viewpoint, as data are successfully received, yet it undermines estimation accuracy and residual-based detection by breaking the temporal alignment between prediction and measurement. In delayed feedback loops, even minor timing offsets accumulate, causing phase lag, false residual correlations, or apparent system drift, which attackers can exploit to hide ongoing

manipulations. As discussed in [12], [17], [53], mitigation strategies must therefore account jointly for availability and freshness: event-triggered and adaptive observers dynamically regulate transmission and estimation policies to maintain bounded error growth despite irregular or delayed updates.

Unified Dynamical Abstraction. Both integrity and availability attacks—including delay-based and replay phenomena—can be embedded within the system dynamics as additive and multiplicative perturbations:

$$\begin{aligned} \mathbf{x}_{k+1} &= F\mathbf{x}_k + H\mathbf{u}_k + E\mathbf{a}_k^u + \mathbf{w}_k, \\ \mathbf{y}_k &= H\mathbf{x}_k + F\mathbf{a}_k^y + \mathbf{v}_k, \end{aligned} \tag{1.28}$$

where E and F identify the channels through which adversarial signals influence the dynamics. In delay-induced attacks, equivalent perturbations can be represented by state-augmentation or transport-delay operators acting on $\mathbf{y}_k$ and $\mathbf{u}_k$, effectively embedding freshness decay into the system model. This hybrid formulation bridges the logical and physical dimensions of adversarial behavior, enabling unified analysis of detectability, stealth, and resilience properties under simultaneous data corruption and timing manipulation.

By explicitly including timing as a dynamical variable, this framework captures the fundamental coupling between integrity, availability, and freshness that underlies cyber–physical resilience. It thus forms the analytical foundation upon which the subsequent layers of this thesis are built—observer-based detection for spatial integrity, freshness-aware communication for temporal coherence, and secure consensus for distributed coordination.

1.3.3 Observer-Based Detection as the Local Defense Layer

At the foundation of adversarial resilience lies the principle of *observer-based detection*, which establishes analytical redundancy by continuously comparing model-predicted outputs with real sensor measurements. By generating residuals that quantify deviations between expected and observed behavior, observers enable the detection of anomalies arising from faults, communication delays, or malicious interference. Unlike purely data-driven detectors, observer-based schemes retain a physical interpretation of anomalies and provide formal guarantees on stability, convergence, and robustness [31], [32], [36]. They thus represent the most rigorous and interpretable local mechanism for ensuring integrity and situational awareness in cyber–physical systems .

From an architectural viewpoint, most observer-based detection frameworks remain inherently *centralized*. A supervisory estimator aggregates sensor and actuator data from the entire plant, executes a global state estimation algorithm, and computes residuals for anomaly detection [33], [38], [39]. This configuration simplifies analytical derivations and allows formal observability-based guarantees, but it introduces structural vulnerabilities: a single point of failure, limited scalability, and sensitivity to communication bottlenecks or denial-of-service attacks. As modern CPS become more decentralized and interconnected, these limitations motivate the adoption of *distributed and hierarchical* designs, where local observers operate autonomously while exchanging selective information with neighbors to enhance coverage and fault tolerance [15], [19], [45]. Within such architectures, observer-based detection forms the fundamental *local defense layer*—a self-contained analytical mechanism that maintains integrity at the subsystem level and feeds reliable information to higher coordination layers.

Resilience under Denial-of-Service Attacks. DoS attacks undermine availability by disrupting the timely exchange of sensor or actuator data, effectively decoupling estimation and control. Their impact depends on the targeted channel: sensor-side DoS reduces observability and estimation accuracy, whereas actuation-side disruptions directly affect closed-loop stability. Early research treated DoS as a stochastic packet-loss process,

adapting Kalman and H_∞ filters to maintain mean-square stability despite intermittent data [46], [53]. Packet dropouts were modeled as Bernoulli or Markov chains, and open-loop prediction compensated missing updates through covariance inflation. Adaptive and gain-scheduled variants [12], [70], [71], [72] refined this approach by adjusting filter gains in response to network reliability. For example, the relaxing-Kalman scheme in [73] adapts to temporally correlated losses, while the "memory-less" filter of [47] preserves estimation under bursty outages. These methods provide continuity and bounded performance in centralized monitoring, but their reliance on statistical independence limits robustness against structured or adversarially timed dropouts.

Deterministic observers grounded in set-membership theory overcome this limitation by representing uncertainty geometrically rather than probabilistically. Zonotopic, ellipsoidal, and interval observers [37], [43], [74], [75] propagate reachable sets enclosing all feasible trajectories consistent with available data and dynamics. During communication outages, these sets expand to reflect increased uncertainty, contracting again when data flow resumes. This approach guarantees bounded estimation error independently of loss statistics and has proven effective in safety-critical CPS such as microgrids and process networks. Hybrid stochastic–deterministic observers [38], [39] further merge probabilistic adaptability with worst-case rigor, achieving resilient estimation under mixed stochastic and adversarial dropout conditions. The main drawback of these deterministic formulations lies in their computational complexity and conservatism, which can obscure subtle anomalies in high-dimensional systems.

Sliding-mode observers (SMOs) provide an alternative route to DoS resilience through discontinuous correction dynamics. Their finite-time convergence properties ensure that estimation errors remain bounded even during prolonged data unavailability. When sensor data are lost, SMOs maintain state prediction and rapidly resynchronize upon recovery [40], [41], [45]; when actuation commands are interrupted, they can reconstruct missing inputs and restore closed-loop consistency [42], [76]. Adaptive and super-twisting formulations [48], [49] enhance responsiveness and noise tolerance, enabling robust operation under bursty or uncertain network conditions. Distributed SMO architectures [45] extend these benefits to multi-agent CPS by allowing localized estimation and inter-node compensation.

Structural resilience can also be achieved via *unknown-input observers* (UIOs) and descriptor-based formulations, which algebraically decouple unavailable signals from estimation dynamics [77], [78], [79]. By isolating the influence of missing measurements or control commands, these observers maintain stable estimation even during sustained outages. Descriptor extensions [44], [71] further generalize this capability to hybrid scenarios involving simultaneous sensor–actuator loss. While UIOs offer formal guarantees, they require precise model knowledge and restrictive rank conditions, which limit scalability; nevertheless, privacy-preserving decentralized variants [78] suggest promising pathways for modular architectures.

A parallel research line emphasizes *co-design of estimation and communication*. Event-triggered and asynchronous observers [17], [33], [40], [80] regulate transmissions based on residual thresholds, sending updates only when prediction errors exceed adaptive bounds. This approach limits bandwidth usage and reduces exposure to jamming, while unexpected silence can itself serve as an anomaly indicator. Consensus-based observers [12], [15], [19], [81] leverage inter-agent redundancy to reconstruct missing information, preserving network-wide observability under partial connectivity. Extended-state observers (ESOs) [82] push this integration further by estimating both system states and unmodeled DoS disturbances, while discrete-event observers [83] couple logical and continuous monitoring to detect switching anomalies. Complementary mechanisms such as packet-timing analysis [58], [84] and watermarking [56], [85] add active verification of data delivery integrity. By combining predictive estimation with communication-level diagnostics, these hybrid designs achieve multi-layered resilience against availability attacks.

Resilience against Deception Attacks. Integrity attacks manipulate signal content rather than timing, creating the illusion of nominal operation while subtly degrading estimation or control. False-data injection, replay, and stealth attacks fall under this class. Observer-based detection here pursues three intertwined goals: *detection*, to expose inconsistencies between measurements and model predictions; *isolation*, to identify compromised channels; and *reconstruction*, to recover or compensate for corrupted data.

Centralized stochastic observers remain the analytical backbone for deception detection. Kalman and H_∞ filters exploit residual statistics to identify deviations from nominal patterns [34], [35]. When attacks are uncorrelated or small, innovation-based thresholds detect anomalies efficiently [37]; however, stealth attacks aligned with unobservable modes evade such detectors [86]. Adaptive Kalman variants [87], [88], [89], [90] and H_∞ formulations [17], [79], [91] extend robustness to time-varying deception and bounded adversarial disturbances, while hybrid implementations embed digital-twin or moving-horizon estimation to counter model manipulation [92], [93]. Actuator-level deception has been addressed through resilient estimator–controller co-design [38], [39], which enforces input–output consistency using invariant sets.

Set-based observers guarantee bounded integrity by replacing statistical thresholds with geometric consistency checks. Ellipsoidal, zonotopic, and interval formulations [36], [37], [43], [68], [75], [94] detect deception when measurements exit the feasible region defined by known inputs and dynamics. These deterministic schemes ensure worst-case robustness and are favored in safety-critical domains, though they incur higher computational cost. Hybrid probabilistic–deterministic designs [38], [39] alleviate this trade-off, achieving a balance between bounded guarantees and sensitivity.

Sliding-mode observers extend detection to *reconstruction*, exploiting finite-time convergence to extract the injected attack signals from the equivalent output error [40], [41], [42], [48], [95]. Adaptive and sparse SMO variants [44], [96], [97] identify multiple concurrent intrusions, while distributed implementations [45], [76] combine finite-time convergence with cooperative estimation. UIOs and descriptor observers [43], [77], [98] achieve algebraic decoupling of malicious inputs, offering guaranteed stability under partial compromise, whereas hybrid UIO–Kalman designs [38] broaden applicability to mixed stochastic–deterministic deception.

Finally, fuzzy and neural observers extend resilience to nonlinear or uncertain dynamics. Fuzzy logic estimators integrate rule-based adaptation with residual evaluation [15], [95], [99], while neural and hybrid observers [100] learn residual patterns directly from data. These learning-augmented schemes enhance adaptability but require safeguards against data poisoning and overfitting; accordingly, they are often coupled with Observer-Based observers that enforce physical consistency [101], [102]. In large-scale and networked settings, distributed observers [12], [17], [33], [56], [103] merge these paradigms through consensus-based fusion and trust-weighted cooperation, providing scalable deception resilience across multiple agents and communication layers.

1.3.4 Information Freshness as a Systemic Security Variable

Beyond integrity and availability, the resilience of cyber–physical systems critically depends on the *freshness* of information circulating across sensing, estimation, and control loops. Even when data remain accurate and communication channels functional, delays, buffering, or replayed transmissions may render information *temporally obsolete*, compromising situational awareness and coordination. Modern adversaries increasingly exploit this dimension of time, introducing artificial latency or selectively blocking updates to erode the real-time coupling between cyber and physical layers [57], [58], [84]. Hence, information freshness emerges as a fundamental *security variable*—a temporal counterpart to integrity—governing the validity and responsiveness of the control system.

The concept of information age originates in networked control and communication theory, where the *Age of Information* quantifies the time elapsed since the last successfully received update was generated [54]. Let t_k denote the generation time of the most recent successfully received measurement and t the current time; then $\Delta(t) = t - t_k$

defines the instantaneous AoI. A smaller AoI indicates fresher data, while larger values correspond to increasing informational staleness. In benign environments, minimizing AoI balances communication cost and estimation accuracy through adaptive sampling, queue management, or event-triggered transmission [53], [55]. However, in adversarial contexts, AoI alone is insufficient: an update may be recent but incorrect (e.g., falsified or replayed), or accurate but outdated due to deliberate delay. To address this, the *Age of Incorrect Information* (AoII) extends the metric to jointly capture *freshness* and *correctness* of information [52]. AoII measures the time since the receiver last possessed a *correct* representation of the system, effectively quantifying how long an estimator or controller has been misinformed.

This adversarial generalization enables a dynamical characterization of deception and DoS within a unified temporal framework. Replay attacks [32], [34] increase AoII by re-injecting previously valid data, thus maintaining apparent consistency while freezing the perceived system evolution. Similarly, communication delays and jamming attacks [55], [57] prolong AoI by postponing updates, causing estimators to rely on outdated predictions. The interplay of AoI and AoII therefore bridges physical and cyber effects: while integrity attacks corrupt the value of information, availability attacks degrade its temporal validity. In both cases, the result is informational aging, where the effective delay between the physical process and its cyber representation grows until control or estimation performance collapses.

Recent studies conceptualize freshness not as a passive performance indicator but as a controllable system state [52]. In this view, both defenders and attackers operate within a dynamic game where update policies, transmission intervals, and attack timing form competing strategies. Defenders seek to minimize their own AoII—preserving timely and correct information—while maximizing the adversary's, i.e., keeping the attacker's perception of the system outdated. This formulation captures the temporal nature of cyber conflict and connects information freshness to control theory through dynamic optimization and differential games. Similarly, [52] shows that the trade-off between communication energy and AoII reduction defines a *freshness budget*, analogous to the control energy in robust control design. Under this paradigm, freshness becomes an active design variable, regulated through event-triggered policies, redundancy scheduling, or cross-layer control of communication protocols [12], [53], [56], [73].

In systems exposed to adversarial latency or replay, maintaining bounded AoII requires both predictive estimation and temporal supervision. Predictive observers—such as the event-triggered and adaptive filters discussed in [17], [40], [53]—propagate state estimates during communication gaps, effectively resetting the information clock once new data arrive. Complementary mechanisms detect abnormal inter-arrival times [58], [84] or exploit watermarking signals [68], [85] to verify the freshness of received packets. These techniques transform temporal irregularities into observable variables, allowing residual-based detectors to integrate freshness anomalies into their decision logic. When integrated within distributed architectures, as in [12], [15], freshness management extends beyond individual links to network-level coordination, ensuring that consensus and synchronization remain robust to asymmetric delays or missing updates.

From a systemic standpoint, freshness directly modulates the effectiveness of observer-based detection and control. Residual generation assumes timely synchronization between prediction and measurement; when updates are delayed or replayed, the residuals may remain small even under significant physical deviation, enabling stealth attacks [34], [35], [104]. Maintaining bounded AoII guarantees that observers operate on temporally consistent data, preserving the interpretability of residuals and the reliability of distributed fusion. In this sense, information freshness forms the *temporal layer* of the broader adversarial resilience architecture: observer-based estimation safeguards spatial and structural integrity, while freshness regulation ensures temporal coherence across the cyber–physical loop.

The control-theoretic treatment of freshness thus aligns with the principles of dynamic security developed in previous sections. Just as observers estimate the system state under corrupted signals, freshness controllers regulate the temporal distance between the physical

system and its cyber representation. Both mechanisms share the same mathematical foundation—a dynamical feedback model—and the same security objective: maintaining bounded deviation between the true and perceived system trajectories. By embedding freshness as an explicit state variable, cyber–physical systems can autonomously adjust communication and estimation policies to sustain awareness, stability, and coordination even under persistent delay, replay, or blocking attacks.

Extending the notion of freshness to networked and multi-agent settings reveals its systemic importance. In consensus-based control and distributed estimation, asynchronous updates and variable delays can fragment shared situational awareness [15], [19]. Within such settings, temporal misalignment between agents degrades estimation accuracy and may even induce instability when outdated information propagates through feedback loops. Recent studies have introduced delay-tolerant and event-triggered consensus mechanisms that implicitly regulate information freshness by adapting communication and estimation rates to network conditions [12]. Complementary approaches exploit trust-weighted and redundancy-based update rules to mitigate the impact of delayed or inconsistent data on distributed decision-making [56], [103]. Embedding freshness awareness into cooperative estimation thus enhances both temporal coherence and resilience, ensuring that collective decisions are based on timely and trustworthy information. This perspective completes the multi-layered resilience architecture developed throughout the thesis: structural integrity through observer-based estimation, temporal coherence through freshness regulation, and coordinated decision-making through secure consensus. Together, these layers constitute a coherent dynamical defense framework capable of preserving awareness and stability in adversarial cyber–physical environments.

1.3.5 Adversarial Resilience in Networked Consensus

At the network scale, cyber–physical systems extend into distributed multi-agent architectures where nodes exchange information to achieve coordination or agreement. Consensus protocols represent the mathematical foundation of this cooperative behavior: through iterative communication, each agent updates its local state based on those of its neighbors until a global objective—such as average agreement, formation synchronization, or collective estimation—is reached [8], [9], [14]. These protocols underpin a wide range of applications, from distributed sensing and formation control to smart-grid regulation and industrial IoT coordination [105], [106]. However, the same interconnectivity that enables cooperation also creates new surfaces for adversarial interference. Attacks that manipulate exchanged data, disrupt communication, or compromise privacy can undermine convergence, destabilize coordination, or leak sensitive information [10], [20].

In nominal operation, consensus evolution over a connected graph $\mathcal{G}$ follows

$$\mathbf{x}_{k+1} = (I - \alpha L)\,\mathbf{x}_k,$$

where L is the Laplacian of $\mathcal{G}$ and $\alpha > 0$ ensures stability [13], [107]. Adversarial behavior alters this dynamics either by corrupting the information exchanged (*deception attacks*) or by interrupting it (*denial-of-service attacks*). Deception manifests through malicious nodes (Byzantine agents) that transmit biased, inconsistent, or falsified values to distort the network estimate [10], [18]. Denial acts by selectively blocking links, introducing communication delays, or desynchronizing updates, effectively degrading the network connectivity and convergence rate [13], [14]. These manipulations can result in biased equilibria, oscillatory behaviors, or complete loss of agreement. Moreover, unlike centralized architectures, distributed consensus lacks a single monitoring point—making early detection and isolation particularly challenging.

To counter these threats, a rich body of research focuses on secure and resilient consensus mechanisms capable of maintaining stability and convergence under data corruption or network degradation [8], [10]. Resilience is achieved by embedding redundancy, validation, and adaptive weighting directly into the local update rules, so that compromised information can be detected and its influence mitigated. Recent advances

generalize this concept to cooperative negotiation and game-theoretic coordination frameworks [18]. Geometric and set-theoretic formulations [11] characterize consensus safety in terms of invariant regions of admissible disagreement, providing analytical guarantees against coordinated manipulation. Dynamic coupling strategies extend this framework to nonlinear or chaotic synchronization problems, ensuring stable agreement even when agents follow complex oscillatory dynamics [108], [109], [110].

Hybrid observer–consensus schemes further improve resilience by fusing estimation and coordination within a unified architecture. In Kalman–consensus filtering [16], local observers exchange state estimates rather than raw data, reducing vulnerability to direct falsification. Subsequent works generalized this paradigm to nonlinear or quantized systems, embedding local detection through residual comparison among neighbors [12], [15], [17]. These designs ensure that even under partial compromise, consistent subgroups of agents can maintain correct consensus, isolating corrupted nodes through local residual inconsistencies.

In parallel with resilience, privacy has emerged as a central requirement for distributed systems operating over shared or untrusted networks. While consensus requires iterative exchange of information, such communication can inadvertently reveal private states or operational data. A variety of privacy-preserving protocols have been proposed to protect sensitive information while maintaining convergence guarantees [20], [21], [111]. Differentially private algorithms [21] inject stochastic perturbations calibrated to balance privacy and accuracy, achieving quantifiable privacy guarantees under statistical attacks. Alternative approaches employ *homomorphic encryption*, allowing agents to perform arithmetic operations on encrypted data without decryption [22], [23]. The seminal *ratio consensus* scheme in [22] achieves exact average consensus while preserving the confidentiality of local values, and has since been extended to asynchronous and quantized scenarios [25], [26]. State-decomposition and perturbation-based techniques [24], [28], [112] achieve similar goals through distributed obfuscation of transmitted signals, enabling privacy without heavy cryptographic computation.

Beyond stochastic and cryptographic techniques, *chaotic encryption* has attracted renewed attention as a lightweight mechanism for secure consensus in resource-constrained CPS. Building on classical results on chaotic synchronization and communication [113], [114], [115], recent works exploit chaotic oscillators to encode and mask transmitted signals in distributed settings [27], [116], [117]. These approaches leverage the sensitivity and unpredictability of chaotic dynamics to provide encryption-like confidentiality with minimal computational overhead, enabling secure synchronization and consensus among agents with limited processing power.

A complementary line of research interprets consensus from the viewpoint of social or cooperative decision-making, where agents possess heterogeneous reliability, influence, or trust. Opinion dynamics models [118], [119] capture how agents adjust their beliefs under bounded confidence and dynamic influence networks. Recent results extend this perspective to resilient coordination, formulating consensus as a distributed negotiation process among possibly reluctant or misaligned agents [18]. In these frameworks, trust coefficients evolve dynamically based on interaction history, allowing the network to attenuate the impact of unreliable nodes while preserving convergence among cooperative ones. Such adaptive trust mechanisms bridge the gap between purely topological resilience and semantic integrity, offering a human- and decision-aware dimension to multi-agent security.

Resilient and privacy-preserving consensus plays a pivotal role in large-scale infrastructures such as smart grids and industrial IoT networks [29], [30], [105]. In these contexts, cyberattacks may simultaneously target physical control signals and distributed estimation processes. Federated and encrypted consensus frameworks [29], [30] enable local autonomy and privacy while supporting global coordination and anomaly detection across heterogeneous industrial nodes. These approaches demonstrate that security and privacy are not antagonistic but complementary: confidentiality mechanisms enhance resilience by reducing exploitable information, while resilient protocols maintain stability under adversarial interference.

Within the multi-layered adversarial modeling framework developed in this thesis, consensus constitutes the *mesoscopic layer* of resilience. While observer-based detection provides local awareness and control integrity, and freshness regulation ensures temporal coherence, secure consensus extends these principles to cooperative networks. Each agent acts simultaneously as an estimator, validator, and communicator, contributing to distributed decision-making that remains robust, confidential, and synchronized even under deception, denial, or infiltration. This systemic view unifies control-theoretic resilience, network synchronization, and information security, revealing consensus not only as a coordination mechanism but as a cornerstone of collective defense in modern cyber–physical systems.

1.3.6 Modelling Cascading Effects in Interdependent Critical Infrastructures

The resilience of Critical Infrastructures (CIs)—including energy, water, transport, and telecommunications—emerges not only from the robustness of each individual system but, more fundamentally, from the network of interdependencies that binds them together. The disruption of one sector can propagate through physical, cyber, informational, or geographic connections, producing systemic or cascading effects that may amplify an initially local disturbance into a large-scale failure. The seminal framework proposed in [120] first articulated the nature of these interdependencies, identifying four main categories—physical, cyber, geographic, and logical—and establishing a foundation for analysing their systemic interactions. Subsequent works, notably [5], [121], [122], [123], refined this conceptualisation, illustrating how infrastructures behave as complex adaptive systems whose interconnections can act both as conduits for functionality and as channels for risk propagation. Early research largely considered passive dependencies, focusing on random failures or natural hazards; however, as infrastructures have become increasingly digitalised, automated, and interlinked through cyber–physical architectures, the very nature of interdependence has evolved, making it both dynamic and potentially adversarial.

Over the past two decades, numerous modelling approaches have been developed to describe and simulate these cascading phenomena. The first generation of models relied on static graph representations and percolation theory to study the structural robustness of interconnected networks, assessing critical nodes, link failures, and percolation thresholds [5], [122]. Although useful for topological analysis, such models often oversimplify dynamic and control-layer interactions, providing only a static depiction of resilience. Flow-based and coupled cyber–physical models subsequently emerged to capture the transfer of resources—such as electricity, water, or data—across interdependent systems. These models integrate physical process dynamics and communication feedback, allowing the study of cascading effects arising from control-layer disruptions or cyber intrusions. For instance, [124] developed a joint simulation framework to evaluate the impact of cyber-attacks on power grids, demonstrating that data manipulation in supervisory control systems can trigger secondary failures in physical assets. Similarly, [125] and [126] explored how delays, data loss, or misinformation in communication networks can amplify physical perturbations across interconnected domains.

At a higher level of abstraction, system-dynamics and input–output models express cross-sector dependencies through interdependency matrices and feedback loops, supporting macro-scale analyses of cascading failures and recovery trajectories [6], [122]. These models are computationally efficient and suitable for policy or risk assessments but often rely on aggregated data and linear approximations. In contrast, agent-based modelling (ABM) introduces a bottom-up perspective, representing infrastructures as autonomous entities—operators, assets, or sectors—that interact through dependency rules and local decision mechanisms. As demonstrated in [127], [128], agent-based approaches can reproduce emergent system-wide behaviours, such as escalation, coordination breakdowns, or adaptive recovery, thereby providing a dynamic representation of interdependencies consistent with real-world complexity.

This methodological diversification has been accompanied by the development of specialised simulation platforms for interdependent CIs. The Australian government's Critical Infrastructure Program for Modelling and Analysis (CIPMA) [129] was among the earliest national initiatives to model multi-sector dependencies and assess systemic vulnerabilities, combining economic, physical, and geographic perspectives. In Europe, major research programmes such as IRRIIS and DIESIS pioneered federated simulation frameworks that allow heterogeneous simulators—each representing a sector—to interact through a common High-Level Architecture (HLA) environment [6], [130]. These efforts were subsequently expanded in projects such as CIPRNet and EPIC, which introduced advanced decision-support functionalities and crisis-management integration. The I2Sim platform, developed at the University of British Columbia, represents infrastructures as interacting resource-exchange agents, enabling quantitative analysis of performance degradation and recovery in interdependent systems. Similarly, EPANET [131] has been widely used in the water sector to analyse pressure-driven dependencies, often coupled with electrical and communication models to study water–energy interactions. Other frameworks such as CyberGRID [132] and EMRALD [133] integrate cyber–physical components for security testing, vulnerability propagation, and resilience evaluation under adversarial conditions. The InfraRisk platform [134] provides asset-level analysis of failure propagation using imprecise dependency information, reflecting the increasing emphasis on uncertainty quantification. Among the most comprehensive European contributions, CISIApro [59], [60], [61] combines agent-based interdependency modelling, cyber-attack propagation, and risk analysis into a unified simulation environment, enabling the study of both accidental and deliberate cascades within complex multi-layered infrastructures.

While traditional dependency models were primarily concerned with stochastic or accidental faults, recent attention has turned toward adversarial and strategic cascades—intentional disruptions exploiting interconnection vulnerabilities. Studies such as [135], [136], [137] emphasise that modern cyber–physical infrastructures can be subject to targeted, coordinated attacks where cyber intrusions trigger physical or economic consequences. The Colonial Pipeline incident, extensively discussed in [3], [4], demonstrated how a cyber-attack can propagate from digital control systems to tangible operational and market effects, highlighting the tight coupling between cyber and physical domains. Adversarial modelling frameworks extend traditional interdependency analysis by introducing decision-making entities—attackers and defenders—whose objectives and strategies shape the evolution of cascading dynamics. Game-theoretic and multi-agent approaches such as those proposed in [50], [51] capture these strategic interactions, while simulation architectures like CISIApro [61] and EMRALD [133] operationalise them through scenario-based experimentation and impact quantification.

Quantifying cascade effects requires appropriate performance metrics capable of capturing both the extent and the dynamics of failure propagation. Early studies relied on topological metrics, such as the size of the largest connected component or the proportion of failed nodes, to estimate system robustness. More recent approaches introduce dynamic resilience indicators—including recovery time, service loss duration, and resilience margin—that capture both degradation and adaptation phases [134], [138]. In adversarial contexts, additional metrics such as stealth, detection delay, and information degradation become essential to assess how intentional actions exploit timing, misinformation, or latency to maintain persistence and amplify cascading consequences. This evolution of metrics reflects a broader shift from reliability-based assessment toward resilience-oriented analysis, where the capacity to anticipate, absorb, and adapt to both accidental and intentional disruptions defines overall systemic stability.

Despite substantial advances, several open challenges remain. The scarcity of high-resolution interdependency data continues to limit model calibration and validation [5]. Many models still assume static topologies or homogeneous dependencies, neglecting the dynamic evolution of modern infrastructures [138]. The integration of cyber–physical control mechanisms, although improving, often lacks fidelity in representing real communication protocols or control logic [124]. Moreover, few frameworks incorporate truly strategic, adaptive adversaries [135], despite the growing relevance of such threats.

Temporal coherence—how delays, asynchrony, and data freshness affect the evolution of interdependent dynamics—remains another underexplored dimension [139]. Finally, large-scale interoperability among simulators, such as those achieved in DIESIS or CIPRNet, is computationally demanding and still constrained by standardisation gaps and limited cross-domain data availability.

1.3.7 Unified Perspective

The state of the art reviewed in this chapter converges toward a unified and hierarchical understanding of adversarial resilience in cyber–physical ecosystems. Across all scales—ranging from local control loops to distributed consensus networks and interdependent critical infrastructures—a common theoretical structure emerges: security, stability, and resilience are not static properties but *dynamic phenomena* arising from the continuous interaction of feedback, information flow, and adaptation. This convergence allows the integration of control theory, communication theory, and network science within a single dynamical framework for modelling, analyzing, and reinforcing the resilience of complex cyber–physical systems.

At the microscopic level, control theory provides the mathematical substrate for resilience. Within the state–space representation of a cyber–physical process, observer-based estimation establishes analytical redundancy by comparing predicted and measured outputs, thereby revealing inconsistencies induced by faults, communication loss, or adversarial manipulation. Security thus manifests as a feedback property: the system continuously regulates itself not only against physical disturbances but also against informational ones. This perspective transforms the notion of intrusion detection into a problem of dynamic observability—detecting whether the deviation between predicted and measured trajectories exceeds what can be explained by modelled uncertainties. Robust and adaptive observers, sliding-mode filters, and set-theoretic estimators all embody this principle, translating system-theoretic concepts such as observability, detectability, and boundedness into quantifiable guarantees of local integrity.

At the mesoscopic scale, the concept of *information freshness* extends this reasoning into the temporal domain. Information exchanged within a cyber–physical feedback loop has both semantic correctness and temporal validity, jointly determining how accurately and promptly the cyber representation mirrors the physical process. The metrics of Age of Information (AoI) and Age of Incorrect Information (AoII) formalize this temporal dimension, linking communication dynamics to estimation accuracy and control stability. Freshness becomes a state variable in its own right—a quantity that evolves over time, is affected by attacks or delays, and can be actively regulated through scheduling, redundancy, or predictive estimation. Under this paradigm, security is understood not only as maintaining correct information but also as ensuring its timeliness. Temporal desynchronization, whether caused by denial-of-service, replay, or delay injection, therefore represents a direct attack on the feedback structure itself, degrading situational awareness and undermining the capacity of the system to respond coherently to change.

At the distributed level, secure and resilient consensus generalizes local detection and freshness regulation to cooperative multi-agent settings. Each node in a networked control system becomes simultaneously an estimator, a verifier, and a communicator. Resilient consensus algorithms such as W-MSR and its derivatives embed validation, isolation, and redundancy into their update dynamics, ensuring convergence despite faulty or malicious participants. Consensus, when viewed through the lens of control theory, constitutes a global feedback mechanism: agents collectively regulate disagreement, synchronize internal models, and maintain bounded divergence even under incomplete, delayed, or deceptive information. The same mechanisms that underpin local estimation—residual generation, weighting, and adaptation—reappear here as distributed validation and trust management. As a result, resilience is no longer confined to the boundaries of an individual controller but becomes an emergent property of cooperative networks.

At the macroscopic scale, interdependency modelling extends this multi-agent resilience to large-scale critical infrastructures. Here, infrastructures are treated as coupled adaptive

systems whose physical, cyber, and organizational dependencies transmit not only resources but also disturbances and risks. Simulation platforms such as CIPMA, DIESIS, CIPRNet, InfraRisk, and CISIApro instantiate this systemic viewpoint, integrating agent-based models, cyber–physical coupling, and adversarial dynamics into federated co-simulation environments. Within these models, the principles identified at lower scales—feedback integrity, freshness regulation, and distributed coordination—manifest as systemic resilience mechanisms that mitigate cascading failures and coordinated multi-sector attacks. In this sense, interdependency modelling provides the structural closure of the framework: it demonstrates how local feedback processes and networked decision-making interact to produce emergent, system-wide stability.

Taken together, these layers reveal a unifying principle: *resilience arises from the recursive embedding of feedback across scales.* At the lowest level, feedback governs physical stabilization and detection; at the network level, it governs synchronization and trust; and at the infrastructure level, it governs adaptive interdependence and systemic recovery. Information freshness ensures temporal coherence among these layers, while consensus dynamics ensure cooperative alignment. Adversaries, conversely, exploit the same mechanisms in reverse—injecting bias into feedback, introducing delay into information flow, and leveraging interdependencies to amplify localized disruptions. Thus, defending a cyber–physical ecosystem requires the joint stabilization of its feedback structures in space, time, and topology.

In this unified perspective:

- *Security* is a *feedback property*, continuously maintained through estimation, correction, and adaptation rather than imposed through static barriers.
- *Information freshness* functions as a *temporal state variable*, whose regulation ensures coherence and responsiveness across communication layers.
- *Resilience* emerges from the hierarchical interplay of detection, synchronization, and adaptation across scales—from local observers to distributed consensus and interdependent infrastructures.

This synthesis provides the conceptual foundation for the adversarial modelling and simulation framework developed in the subsequent chapters of this thesis. By explicitly linking observer-based detection, freshness-aware communication, secure consensus, and interdependency dynamics under a single mathematical and conceptual lens, the unified framework establishes a continuous spectrum of resilience—from micro-scale control integrity to macro-scale systemic stability—capable of addressing the evolving challenges of adversarial cyber–physical environments.

1.4 Thesis Overview and Contributions

The overarching goal of this dissertation is to advance the *resilience, security, and trustworthiness* of modern cyber–physical systems by developing *Observer-Based*, *distributed*, and *uncertainty-aware* methods for monitoring, estimation, coordination, and systemic risk assessment. The research unfolds across four complementary yet tightly interconnected directions, each addressing a specific limitation identified in the state of the art and collectively contributing to a unified theory and practice of *distributed adversarial resilience* in industrial and critical infrastructures.

1. *Observer-Based Attack Detection and Reconstruction.* The first research axis addresses the need for scalable and reliable estimation mechanisms that minimize dependence on centralized supervision. Building on the theory of observers and Kalman filtering, the thesis exploits the sparsity and spatial decay of cross-correlations in large-scale systems to design partitioned estimation architectures. This insight leads to the development of the *Multirate Interlaced Kalman Filter* (MIKF), a distributed estimation framework capable of handling asynchronous

updates, network-induced delays, and information aging. Complementary nonlinear approaches based on *Sliding Mode Observers* (SMOs) are also proposed to achieve robust detection, isolation, and reconstruction of sensor and actuator attacks under bounded uncertainty. Together, these methods extend the paradigm of *Observer-Based Anomaly Detection* (OBAD) toward distributed, scalable, and consistency-preserving implementations suitable for real-time operation in industrial CPSs.

2. *Information Freshness and Temporal Resilience in Adversarial Networks.* The second research axis investigates how information timeliness and correctness jointly determine the stability and situational awareness of networked control systems. Building upon the communication-theoretic notion of the *Age of Information* (AoI), the work introduces and formalizes the *Age of Incorrect Information* (AoII) as a quantitative indicator capturing both the temporal and semantic reliability of exchanged data. Within this framework, the thesis systematically studies the impact of common adversarial mechanisms—such as *false-data injection*, *denial-of-service*, and *data modification* attacks—on the evolution of AoII and on the resulting estimation and control performance. By modeling the interaction between legitimate transmitters and adversaries via a game-theoretic formulation, equilibrium conditions are derived that balance update frequency, estimation accuracy, and resilience under communication constraints and targeted disruptions. This formulation provides a principled foundation for the co-design of estimation, communication, and scheduling policies that explicitly regulate information freshness and correctness in adversarial and resource-limited environments.
3. *Secure and Resilient Consensus in Multi-Agent Systems.* The third research direction extends the previous results from local estimation to cooperative networked control. Traditional consensus protocols guarantee synchronization and agreement among agents but rely on assumptions of reliable communication and honest participation—conditions rarely satisfied in adversarial settings. To overcome these limitations, the thesis develops a suite of *secure and resilient consensus protocols* that maintain convergence and stability under false-data injection, substitution, or denial-of-service attacks. The proposed framework integrates trust-weighted update rules, evidence-theoretic reasoning, and dynamic isolation mechanisms that allow agents to identify and attenuate the influence of compromised peers. Moreover, *privacy-preserving extensions* are introduced using nonlinear synchronization and chaotic masking, ensuring that cooperative behavior does not expose confidential information. Collectively, these contributions establish a rigorous and practical basis for resilient, privacy-aware cooperative control in distributed CPSs.
4. *Modeling and Simulation of Critical Infrastructure Interdependencies.* The fourth and final research axis elevates the analysis from component-level defense to systemic interdependence among critical infrastructures. Building upon and extending the *CISIApro 2.0* simulation environment, the thesis introduces fuzzy and possibilistic extensions for representing the propagation of cyber–physical disturbances across coupled sectors. These extensions generalize the classical *Confidentiality, Integrity, and Availability* (CIA) dimensions to include epistemic and aleatoric uncertainty, enabling hybrid quantitative–qualitative modeling of cascading effects. The enhanced CISIApro framework supports the analysis of cross-sector dependencies, uncertainty-driven resilience assessment, and scenario-based evaluation of coordinated attacks and defense strategies. This contribution bridges analytical modeling and decision support, offering tools for policy-makers and operators to evaluate national-scale resilience and preparedness.

Together, these four axes form a coherent research trajectory spanning multiple scales of cyber–physical resilience: from local observer-based estimation (microscopic layer), through freshness-aware and consensus-driven coordination (mesoscopic layer), to the systemic modeling of interdependent infrastructures (macroscopic layer). Each layer builds

upon and reinforces the others, collectively demonstrating that security and resilience emerge not from isolated mechanisms but from the structured interplay of *feedback, information freshness, and interdependence*.

The thesis thus both theoretical advances and practical frameworks that support the realization of *resilient and trustworthy autonomy*—the capacity of cyber–physical systems to detect, withstand, and recover from adversarial disruptions while maintaining stability, integrity, and availability across spatial and temporal scales. This unifying perspective provides the conceptual and methodological foundation for the multi-layer adversarial modeling framework developed in the subsequent chapters.

1.5 Thesis Organization

The dissertation is organized into *four main parts*, each addressing a complementary aspect of cyber–physical system security and resilience. The structure follows a logical flow—from local observer-based estimation and attack detection, through the temporal reliability of information, to secure multi-agent consensus, and finally to the modeling of interdependent critical infrastructures. Each part builds on the previous one, contributing to the unified goal of developing a model-based framework for resilient cyber–physical systems.

To guide the reader through the structure of the dissertation, Table 1.2 summarizes the main focus and contributions of each part and chapter. In particular, it relates the thesis chapters to the corresponding scientific publications authored by the candidate during the Ph.D. period.

Part I – Observer-Based Attack Detection and Isolation (Chapters 2–5) establishes the theoretical and methodological foundations of observer-based security. It focuses on observer design, estimation robustness, and the reconstruction of attacks in distributed control architectures. *Chapter 2* introduces the fundamentals of observer-based fault and attack detection. It provides a comprehensive review of existing model-based approaches—such as Luenberger observers, Kalman filters, and unknown input observers—highlighting their strengths and limitations in the context of cyber–physical systems. The chapter concludes by outlining the main research gaps that motivate the contributions of this thesis. *Chapter 3* explores the exploitation of sparsity in Kalman filtering for large-scale interconnected systems. It presents the mathematical framework for spatial decay analysis, demonstrating how estimation error correlations diminish with distance in networked settings. This property enables computationally efficient and scalable implementations of distributed Kalman filters. *Chapter 4* develops the proposed *Multirate Interlaced Kalman Filter* , a novel architecture for distributed attack detection under asynchronous operation. The chapter derives stability and consistency conditions for multirate estimation, and validates the method through simulation studies that show improved robustness against sensor attacks and data dropouts. *Chapter 5* introduces a *Sliding Mode Observer*-based framework for distributed attack reconstruction. By combining modular decomposition with nonlinear estimation, it enables the detection and isolation of sensor attacks in networked control systems. The chapter presents both theoretical results and numerical experiments, illustrating the observer's capability to localize and quantify attack signals in real time.

Part II – Information Freshness in Adversarial Multi-Agent Systems (Chapters 6–10) focuses on the temporal reliability and correctness of information exchanged in adversarial communication networks. It introduces the concepts of *Age of Information* and *Age of Incorrect Information* as quantitative metrics that link communication dynamics with estimation accuracy. *Chapter 6* provides an overview of information freshness in networked control systems, reviewing existing metrics and models under adversarial conditions. It introduces the AoI and AoII frameworks used throughout the thesis and formalizes the modeling assumptions for the subsequent game-theoretic analysis. *Chapter 7* investigates *False Data Injection* attacks using a static game-theoretic formulation under complete

Table 1.2: Summary of thesis structure and main contributions

Part / Chapters	Focus Area	Main Contributions
Part I		
Ch. 2–5	Observer-Based Attack Detection	Distributed Model-based detection and reconstruction of attacks based on Kalman filtering ([C2], [C5], [J4]). Sliding-mode observer design for distributed attack isolation ([C16], [J9]).
Part II		
Ch. 6–10	Information Freshness in Adversarial Networks	Analysis of *Age of Information (AoI)* and *Age of Incorrect Information (AoII)* ([J2], [J3], [J7]). Game-theoretic modeling of false data injection, denial-of-service, and data substitution attacks ([C4], [C10], [C13], [C14]). Comparative study of equilibrium strategies and resilience trade-offs.
Part III		
Ch. 11–13	Secure and Resilient Consensus	Design of trust-weighted and privacy-preserving consensus mechanisms ([C1], [J1]). Integration of trust reasoning and nonlinear synchronization for secure coordination. Demonstration of resilience and privacy guarantees through simulations.
Part IV		
Ch. 14–17	Critical Infrastructure Interdependencies	Modeling of systemic resilience using the *CISIApro 2.0* simulator ([C6], [C12]). Simulation of cascading failures and uncertainty using possibilistic and fuzzy reasoning. Applications to the national cybersecurity perimeter and interdependent infrastructures ([C11]).

information. The interaction between attacker and defender is modeled as a strategic optimization problem, and Nash equilibria are derived to characterize optimal attack and defense strategies. *Chapter 8* extends the analysis to *Denial-of-Service* attacks, where data transmission is intermittently blocked. The chapter introduces a stochastic game formulation capturing random communication failures and quantifies their effect on information freshness and system performance. *Chapter 9* addresses *Data Substitution* attacks, where outdated or falsified information replaces legitimate updates. Analytical results are derived to determine equilibrium conditions, providing insight into when such attacks can persist undetected. *Chapter 10* presents a comparative analysis of all three adversarial scenarios, synthesizing results on equilibrium properties, defense efficiency, and the trade-off between accuracy, timeliness, and resilience in multi-agent communication.

Part III – Secure and Resilient Consensus in Networked Multi-Agent Systems (Chapters 11–13) extends the investigation from estimation to distributed coordination, addressing the problem of achieving reliable consensus among autonomous agents in adversarial environments. *Chapter 11* introduces the fundamental principles of resilient consensus and surveys the main strategies proposed in the literature to ensure stability under malicious influence. It positions the thesis contributions within this broader context, emphasizing the integration of trust and uncertainty reasoning into consensus protocols. *Chapter 12* proposes a *trust-weighted consensus mechanism* in which agents dynamically

adjust their confidence in neighbors based on observed behavior. A reputation assignment and weight correction strategy is developed, allowing the network to maintain agreement even when some agents are compromised. The approach is validated through simulation results demonstrating its resilience. *Chapter 13* presents a *privacy-preserving consensus* framework based on nonlinear synchronization and chaotic masking. This scheme enables agents to achieve accurate consensus without revealing their internal states, ensuring both cooperation and confidentiality. Analytical and numerical results confirm the robustness and privacy guarantees of the proposed approach.

Part IV – Critical Infrastructure Interdependencies Modeling (Chapters 14–17) broadens the perspective from distributed control to systemic resilience, focusing on the modeling and simulation of interdependent critical infrastructures. It integrates cyber–physical reasoning with uncertainty modeling to capture cascading effects across large-scale systems. *Chapter 14* reviews the state of the art on interdependency modeling and cascading failures, outlining conceptual, regulatory, and computational frameworks. It introduces the *CISIApro 2.0* agent-based simulator used throughout this part. *Chapter 15* applies the CISIApro 2.0 framework to the analysis of the *national cybersecurity perimeter*, simulating cross-sector impacts of cyber incidents across electric, telecommunication, transport, and governmental infrastructures. The results highlight how local disruptions can propagate through interdependent networks. *Chapter 16* extends CISIApro 2.0 with a *possibilistic* reasoning layer to explicitly model epistemic uncertainty. This extension allows the simulation of incomplete or ambiguous knowledge and improves the robustness of risk assessment. *Chapter 17* integrates *fuzzy modeling* into the CISIApro 2.0 framework, introducing hybrid possibilistic–fuzzy reasoning for a more expressive representation of uncertainty. The chapter presents application results that demonstrate the advantages of this hybrid approach in systemic resilience analysis.

Finally, the concluding chapter synthesizes the main contributions of the thesis, emphasizing how observer-based detection, information freshness analysis, resilient consensus, and interdependency modeling converge into a unified framework for trustworthy and resilient cyber–physical systems.

References — Introduction

[1] Wolf, W., "Cyber-physical systems," *Computer*, vol. 42, no. 03, pp. 88–89, 2009.

[2] H. Lee, R. Nuqui, and J. Hong, "Cyber intrusion detection and sensor failure diagnosis using statistical hypothesis-based correlation: Application to FACTS AC filter banks," in *2024 IEEE/PES Transmission and Distribution Conference and Exposition (T&D)*, Journal Abbreviation: 2024 IEEE/PES Transmission and Distribution Conference and Exposition (T&D), May 6, 2024, pp. 1–5.

[3] Hobbs, A., "The colonial pipeline hack: Exposing vulnerabilities in us cybersecurity," in *SAGE Business Cases*, SAGE Publications: SAGE Business Cases Originals, 2021.

[4] Goodell, J. W. and Corbet, S., "Commodity market exposure to energy-firm distress: Evidence from the colonial pipeline ransomware attack," *Finance Research Letters*, vol. 51, p. 103 329, 2023.

[5] Ouyang, M., "Review on modeling and simulation of interdependent critical infrastructure systems," *Reliability engineering & System safety*, vol. 121, pp. 43–60, 2014.

[6] Setola, R., Rosato, V., Kyriakides, E., and Rome, E., *Managing the complexity of critical infrastructures: A modelling and simulation approach*. Springer Nature, 2016.

[7] Boyd, S., Boyd, S. P., and Vandenberghe, L., *Convex optimization*. Cambridge university press, 2004.

[8] Olfati-Saber, R., Fax, J. A., and Murray, R. M., "Consensus and cooperation in networked multi-agent systems," *Proceedings of the IEEE*, vol. 95, no. 1, pp. 215–233, 2007.

[9] Ren, W., "Consensus based formation control strategies for multi-vehicle systems," in *2006 American Control Conference*, IEEE, 2006, 6–pp.

[10] Pasqualetti, F., Dörfler, F., and Bullo, F., "Attack detection and identification in cyber-physical systems," *IEEE transactions on automatic control*, vol. 58, no. 11, pp. 2715–2729, 2013.

[11] Fioravanti, C., Oliva, G., Faramondi, L., and Hadjicostis, C. N., "A geometrical approach for consensus security," *Systems & Control Letters*, Submitted.

[12] Weng, P., Chen, B., and Wang, R., "Secure fusion estimation against FDI sensor attacks in cyber–physical systems," *Franklin Open*, vol. 7, 2024. DOI: 10.1016/j.fraope.2024.100108. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85212601982&doi=10.1016%2fj.fraope.2024.100108&partnerID=40&md5=e2cc6cceade2ff9487294b998ff4ea39.

[13] Olfati-Saber, R. and Murray, R. M., "Consensus problems in networks of agents with switching topology and time-delays," *IEEE Transactions on automatic control*, vol. 49, no. 9, pp. 1520–1533, 2004.

[14] Sabattini, L., Secchi, C., Chopra, N., and Gasparri, A., "Distributed control of multirobot systems with global connectivity maintenance," *IEEE Transactions on Robotics*, vol. 29, no. 5, pp. 1326–1332, 2013.

[15] Li, J., Yang, D., and Su, Q., "Sliding mode resilient control and application based on intermediate variable observer in smart grid," *International Journal of Control, Automation and Systems*, vol. 21, no. 6, pp. 1803–1815, 2023. DOI: 10.1007/s12555-021-0916-0. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85156221639&doi=10.1007%2fs12555-021-0916-0&partnerID=40&md5=5e9f10d1f575b855a6dab889f9a75cdd.

[16] Olfati-Saber, R., "Kalman-consensus filter: Optimality, stability, and performance," in *Proceedings of the 48h IEEE Conference on Decision and Control (CDC) held jointly with 2009 28th Chinese Control Conference*, Ieee, 2009, pp. 7036–7042.

[17] Chen, G., Zhang, Y., Gu, S., and Hu, W., "Resilient state estimation and control of cyber-physical systems against false data injection attacks on both actuator and sensors," *IEEE Transactions on Control of Network Systems*, vol. 9, no. 1, pp. 500–510, 2022. DOI: 10.1109/TCNS.2021.3113265. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85115126198&doi=10.1109%2fTCNS.2021.3113265&partnerID=40&md5=c70f88330325683ae416e838d0f90786.

[18] Oliva, G., Rikos, A. I., Gasparri, A., and Hadjicostis, C. N., "Distributed negotiation for reaching agreement among reluctant players in cooperative multi-agent systems," *IEEE Transactions on Automatic Control*, 2022.

[19] Chattopadhyay, A. and Mitra, U., "Security against false data-injection attack in cyber-physical systems," *IEEE Transactions on Control of Network Systems*, vol. 7, no. 2, pp. 1015–1027, 2020. DOI: 10.1109/TCNS.2019.2927594. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85069439477&doi=10.1109%2fTCNS.2019.2927594&partnerID=40&md5=c1f7af7cb157c776bb7e05c26d346374.

[20] Mo, Y. and Murray, R. M., "Privacy preserving average consensus," *IEEE Transactions on Automatic Control*, vol. 62, no. 2, pp. 753–765, 2016.

[21] Nozari, E., Tallapragada, P., and Cortés, J., "Differentially private average consensus: Obstructions, trade-offs, and optimal algorithm design," *Automatica*, vol. 81, pp. 221–231, 2017.

[22] Hadjicostis, C. N. and Domínguez-García, A. D., "Privacy-preserving distributed averaging via homomorphically encrypted ratio consensus," *IEEE Transactions on Automatic Control*, vol. 65, no. 9, pp. 3887–3894, 2020.

[23] Fang, W., Zamani, M., and Chen, Z., "Secure and privacy preserving consensus for second-order systems based on paillier encryption," *Systems & Control Letters*, vol. 148, p. 104 869, 2021.

[24] Calis, M., Heusdens, R., and Hendriks, R. C., "A privacy-preserving asynchronous averaging algorithm based on state decomposition," in *2020 28th European Signal Processing Conference (EUSIPCO)*, IEEE, 2021, pp. 2115–2119.

[25] Kishida, M., "Encrypted average consensus with quantized control law," in *2018 IEEE Conference on Decision and Control (CDC)*, IEEE, 2018, pp. 5850–5856.

[26] Rikos, A. I., Charalambous, T., Johansson, K. H., and Hadjicostis, C. N., "Distributed event-triggered algorithms for finite-time privacy-preserving quantized average consensus," *IEEE Transactions on Control of Network Systems*, 2022.

[27] Fioravanti, C., Oliva, G., Panzieri, S., and Hadjicostis, C. N., "Private consensus using chaotic oscillator-based encryption," in *2022 30th Mediterranean Conference on Control and Automation (MED)*, IEEE, 2022, pp. 927–932.

[28] Xiong, Y. and Li, Z., "Privacy preserving average consensus by adding edge-based perturbation signals," in *2020 IEEE Conference on Control Technology and Applications (CCTA)*, IEEE, 2020, pp. 712–717.

[29] Boopalan, P. et al., "Fusion of federated learning and industrial internet of things: A survey," *Computer Networks*, p. 109 048, 2022.

[30] Abdel-Basset, M., Moustafa, N., and Hawash, H., "Privacy-preserved generative network for trustworthy anomaly detection in smart grids: A federated semisupervised approach," *IEEE transactions on industrial informatics*, vol. 19, no. 1, pp. 995–1005, 2022.

[31] Teixeira, A., Sandberg, H., and Johansson, K., "Networked control systems under cyber attacks with applications to power networks," presented at the Proceedings of the 2010 American Control Conference, ACC 2010, 2010, pp. 3690–3696. DOI: 10.1109/acc.2010.5530638. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-77957775842&doi=10.1109%2facc.2010.5530638&partnerID=40&md5=5097e13dfcc486051b09d9ced37a0cbd.

[32] Chowdhury, N., Belikov, J., Baimel, D., and Levron, Y., "Observer-based detection and identification of sensor attacks in networked CPSs," *Automatica*, vol. 121, 2020. DOI: 10.1016/j.automatica.2020.109166. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85088921633&doi=10.1016%2fj.automatica.2020.109166&partnerID=40&md5=d01180439d1af8f1e18094afbfefc239.

[33] Ahmad, S. and Ahmed, H., "Robust intrusion detection for resilience enhancement of industrial control systems: An extended state observer approach," *IEEE Transactions on Industry Applications*, vol. 59, no. 6, pp. 7735–7743, 2023. DOI: 10.1109/TIA.2023.3305361. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85168281024&doi=10.1109%2fTIA.2023.3305361&partnerID=40&md5=0aadf0e4dddca4fcbd379c3cf8c5faf2.

[34] Y. Mo, R. Chabukswar, and B. Sinopoli, "Detecting integrity attacks on SCADA systems," *IEEE Transactions on Control Systems Technology*, vol. 22, no. 4, pp. 1396–1407, Jul. 2014. DOI: 10.1109/TCST.2013.2280899.

[35] K. Manandhar, X. Cao, F. Hu, and Y. Liu, "Detection of faults and attacks including false data injection attack in smart grid using kalman filter," *IEEE Transactions on Control of Network Systems*, vol. 1, no. 4, pp. 370–379, Dec. 2014. DOI: 10.1109/TCNS.2014.2357531.

[36] Zhang, H., He, L., Wang, Z., Gao, S., and Yan, H., "The detection mechanism for false data injection attack via the ellipsoidal set-membership approach," *Asian Journal of Control*, vol. 25, no. 6, pp. 4853–4863, 2023. DOI: 10.1002/asjc.3166. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85164323819&doi=10.1002%2fasjc.3166&partnerID=40&md5=e0d42f40a1398ce244db93233ac19307.

[37] Li, J., Wang, Z., Shen, Y., and Xie, L., "Attack detection for cyber-physical systems: A zonotopic approach," *IEEE Transactions on Automatic Control*, vol. 68, no. 11, pp. 6828–6835, 2023. DOI: 10.1109/TAC.2023.3240383. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85148462989&doi=10.1109%2fTAC.2023.3240383&partnerID=40&md5=0ec1842e79fa72b3241e8601a323c516.

[38] Forti, N., Battistelli, G., Chisci, L., and Sinopoli, B., "Joint attack detection and secure state estimation of cyber-physical systems," *International Journal of Robust and Nonlinear Control*, vol. 30, no. 11, pp. 4303–4330, 2020. DOI: 10.1002/rnc.4724. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85071986226&doi=10.1002%2frnc.4724&partnerID=40&md5=c64ef74d02845a7ffa03b43962dd0eff.

[39] Leko, D. and Vasak, M., "Minimum robust invariant sets and kalman filtering in cyber attacking and defending," presented at the Proceedings of the American Control Conference, vol. 2022-June, 2022, pp. 2924–2931. DOI: 10.23919/ACC53348.2022.9867272. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85138492583&doi=10.23919%2fACC53348.2022.9867272&partnerID=40&md5=5cb4fd3346b494777218c47a362a5266.

[40] Su, Q., Li, S., Gao, Y., Huang, X., and Li, J., "Observer-based detection and reconstruction of dynamic load altering attack in smart grid," *Journal of the Franklin Institute*, vol. 358, no. 7, pp. 4013–4027, 2021. DOI: 10.1016/j.jfranklin.2021.02.008. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85102077025&doi=10.1016%2fj.jfranklin.2021.02.008&partnerID=40&md5=f3f5033ebd2f30654f193b26053deaca.

[41] Li, J., Yang, D., and Su, Q., "Reliable control strategy based on sliding mode observer against FDI attacks in smart grid," *Asian Journal of Control*, vol. 25, no. 2, pp. 910–920, 2023. DOI: 10.1002/asjc.2839. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85129260072&doi=10.1002%2fasjc.2839&partnerID=40&md5=8034b5a649f3cebf8d0c518b7e03cc82.

[42] He, K., Long, Y., and Li, T., "Attack detection and secure correction control for cyber-physical systems subject to actuator deception attacks," presented at the Proceedings of the 35th Chinese Control and Decision Conference, CCDC 2023, 2023, pp. 3356–3361. DOI: 10.1109/CCDC58219.2023.10326906. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85181806979&doi=10.1109%2fCCDC58219.2023.10326906&partnerID=40&md5=5d913316bb236d39ab0aa3204eb35405.

[43] Liu, H., Li, Y., Han, Q.-L., Raissi, T., and Chai, T., "Secure estimation, attack isolation, and reconstruction based on zonotopic unknown input observer," *IEEE Transactions on Automatic Control*, vol. 68, no. 12, pp. 7312–7325, 2023. DOI: 10.1109/TAC.2023.3275965. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85160221136&doi=10.1109%2fTAC.2023.3275965&partnerID=40&md5=3dfe8540d677cbb541e9a028c2633f16.

[44] Yang, H. and Ma, R., "Sparse attack reconstruction for cyber–physical systems via descriptor reduced-order observer," *Nonlinear Analysis: Hybrid Systems*, vol. 50, 2023. DOI: 10.1016/j.nahs.2023.101380. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85166240334&doi=10.1016%2fj.nahs.2023.101380&partnerID=40&md5=6c4f534825b72b07cac859cf96bb0986.

[45] Kachhwaha, M., Modi, H., Nehra, M., and Fulwani, D., "Robust observer-based defense strategy against actuator and sensor cyber-attacks in DCMGs," *IEEE Transactions on Industrial Informatics*, vol. 20, no. 10, pp. 11 687–11 696, 2024. DOI: 10.1109/TII.2024.3412211. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85206516903&doi=10.1109%2fTII.2024.3412211&partnerID=40&md5=d0e7331a2ad7d1793579a40eb1e27078.

[46] Y. Mo and B. Sinopoli, "On the performance degradation of cyber-physical systems under stealthy integrity attacks," *IEEE Transactions on Automatic Control*, vol. 61, no. 9, pp. 2618–2624, Sep. 2016. DOI: 10.1109/TAC.2015.2498708.

[47] Bonczek, P. and Bezzo, N., "Memoryless cumulative sign detector for stealthy CPS sensor attacks," presented at the IFAC-PapersOnLine, vol. 53, 2020, pp. 838–844. DOI: 10.1016/j.ifacol.2020.12.840. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85107537377&doi=10.1016%2fj.ifacol.2020.12.840&partnerID=40&md5=f3fcaf93fd070d910adca02236b651c7.

[48] Chen, Y., Liang, Z., Zhao, Z., Wang, H., and Dong, J., "Super-twisting nonsingular terminal sliding mode control for cyber physical system under FDI attacks," *Nonlinear Dynamics*, vol. 113, no. 3, pp. 2545–2558, 2025. DOI: 10.1007/s11071-024-10358-8. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85205535663&doi=10.1007%2fs11071-024-10358-8&partnerID=40&md5=27f3b222d0a7ef7cbbbb7d26809fd71e.

[49] Nateghi, S., Shtessel, Y., and Edwards, C., "Resilient control of cyber-physical systems under sensor and actuator attacks driven by adaptive sliding mode observer," *International Journal of Robust and Nonlinear Control*, vol. 31, no. 15, pp. 7425–7443, 2021. DOI: 10.1002/rnc.5694. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85110016740&doi=10.1002%2frnc.5694&partnerID=40&md5=76e7a5942040f084365b2a59a6513c47.

[50] Attiah, A., Chatterjee, M., and Zou, C. C., "A game theoretic approach to model cyber attack and defense strategies," in *2018 IEEE International Conference on Communications (ICC)*, IEEE, 2018, pp. 1–7.

[51] Kotenko, I., Saenko, I., and Lauta, O., "Modeling the impact of cyber attacks," *Cyber Resilience of Systems and Networks*, pp. 135–169, 2019.

[52] Banerjee, S. and Ulukus, S., "The freshness game: Timely communications in the presence of an adversary," *IEEE/ACM Transactions on Networking*, vol. 32, no. 5, pp. 4067–4084, 2024. DOI: 10.1109/TNET.2024.3410228.

[53] Liu, H., Wang, S., and Li, Y., "Event-triggered control and proactive defense for cyber-physical systems," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 52, no. 10, pp. 6305–6313, 2022. DOI: 10.1109/TSMC.2022.3144337. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85124075239&doi=10.1109%2fTSMC.2022.3144337&partnerID=40&md5=023f58b4b0c7cb5e1636d8b04ee52e37.

[54] Yates, R. D., Sun, Y., Brown, D. R., Kaul, S. K., Modiano, E., and Ulukus, S., "Age of information: An introduction and survey," *IEEE Journal on Selected Areas in Communications*, vol. 39, no. 5, pp. 1183–1210, May 2021, Publisher: Institute of Electrical and Electronics Engineers (IEEE). DOI: 10.1109/jsac.2021.3065072. Accessed: Jul. 22, 2025.

[55] Zhang, H., Cheng, P., Shi, L., and Chen, J., "Optimal denial-of-service attack scheduling with energy constraint," *IEEE Transactions on Automatic Control*, vol. 60, no. 11, pp. 3023–3028, 2015. DOI: 10.1109/TAC.2015.2409905. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-84945315428&doi=10.1109%2fTAC.2015.2409905&partnerID=40&md5=715aacd03ce9a4c2a096d193d6ad790f.

[56] Li, G., Wang, Y., Min, Z., and Zhang, D., "Faults and attacks collaborative detection for cyber-physical systems with mixed network attacks," presented at the Chinese Control Conference, CCC, 2024, pp. 4859–4864. DOI: 10.23919/CCC63176.2024.10661834. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85205509751&doi=10.23919%2fCCC63176.2024.10661834&partnerID=40&md5=a2aa3e0dc924d40489d3e3eae0ddfa79.

[57] J. Xiao and M. Feroskhan, "Cyber attack detection and isolation for a quadrotor UAV with modified sliding innovation sequences," *IEEE Transactions on Vehicular Technology*, vol. 71, no. 7, pp. 7202–7214, Jul. 2022. DOI: 10.1109/TVT.2022.3170725.

[58] Piao, X. and Lin, C., "Sensor attack and state estimation for cyber-physical systems with state delay via reduced-order observer," presented at the Chinese Control Conference, CCC, vol. 2023-July, 2023, pp. 5158–5162. DOI: 10.23919/CCC58697.2023.10240583. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85175578064&doi=10.23919%2fCCC58697.2023.10240583&partnerID=40&md5=bfaf5b418aa8373844a692baa649992d.

[59] Foglietta, C. and Panzieri, S., "Resilience in critical infrastructures: The role of modelling and simulation," in *Issues on Risk Analysis for Critical Infrastructure Protection*, IntechOpen, 2020.

[60] Bernardini, E., Foglietta, C., and Panzieri, S., "Modeling telecommunications infrastructures using the cisiapro 2.0 simulator," in *International Conference on Critical Infrastructure Protection*, Springer, 2020, pp. 325–348.

[61] Foglietta, C., Palazzo, C., Santini, R., and Panzieri, S., "Assessing cyber risk using the cisiapro simulator," in *International Conference on Critical Infrastructure Protection*, Springer, 2015, pp. 315–331.

[62] Foglietta, C. et al., "The italian national security perimeter for cyber: Assessing and modeling cyber threats in interdependent critical infrastructures with the cisiapro simulator," in *Seventeenth Annual IFIP WG 11.10 International Conference on Critical Infrastructure Protection*, 2023.

[63] Cao, L., Jiang, X., Zhao, Y., Wang, S., You, D., and Xu, X., "A survey of network attacks on cyber-physical systems," *IEEE Access*, vol. 8, pp. 44 219–44 227, 2020.

[64] Kordestani, M. and Saif, M., "Observer-based attack detection and mitigation for cyberphysical systems: A review," *IEEE Systems, Man, and Cybernetics Magazine*, vol. 7, no. 2, pp. 35–60, 2021.

[65] X. Huang, J. Li, and Q. Su, "An observer with cooperative interaction structure for biasing attack detection and secure control," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 53, no. 4, pp. 2543–2553, Apr. 2023. DOI: 10.1109/TSMC.2022.3213516.

[66] Teixeira, A., Sou, K. C., Sandberg, H., and Johansson, K. H., "Secure control systems: A quantitative risk management approach," *IEEE Control Systems Magazine*, vol. 35, no. 1, pp. 24–45, 2015.

[67] G. Lu and D. Feng, "Network security situation awareness for industrial control system under integrity attacks," in *2018 21st International Conference on Information Fusion (FUSION)*, Journal Abbreviation: 2018 21st International Conference on Information Fusion (FUSION), Jul. 10, 2018, pp. 1808–1815. DOI: 10.23919/ICIF.2018.8455208.

[68] Ren, G., Liu, H., Zhang, Y., and Li, Y., "Attack detection based on encoding-decoding approach for cyber-physical systems," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 53, no. 10, pp. 6293–6302, 2023. DOI: 10.1109/TSMC.2023.3283414. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85162858130&doi=10.1109%2fTSMC.2023.3283414&partnerID=40&md5=7fc527c6f9b8b94d096dcd9868f1dfab.

[69] Khorrami, L. and Afshar, A., "Attack detection in active queue management within large-scale networks control system with information of network and physical system," presented at the 2016 24th Iranian Conference on Electrical Engineering, ICEE 2016, 2016, pp. 714–719. DOI: 10.1109/IranianCEE.2016.7585614. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-84994085527&doi=10.1109%2fIranianCEE.2016.7585614&partnerID=40&md5=920b9a99ee3553e3b1e5d592968e1f81.

[70] Zheng, J., Liu, D., Hua, J., and Ning, X., "Optimal two-channel switching false data injection attacks against remote state estimation of the unmanned aerial vehicle cyber-physical system," *Defence Technology*, vol. 47, pp. 319–332, 2025. DOI: 10.1016/j.dt.2024.12.025. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85215929359&doi=10.1016%2fj.dt.2024.12.025&partnerID=40&md5=23042d033060c30901ab22d750b45114.

[71] Zhuang, X., Wang, Z., and He, W., "Secure remote state estimation against false data injection attacks based on improved kalman filtering algorithm," presented at the CFIMA 2024 - Proceedings of 2024 2nd International Conference on Frontiers of Intelligent Manufacturing and Automation, 2025, pp. 396–400. DOI: 10.1145/3704558.3707075. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85217840049&doi=10.1145%2f3704558.3707075&partnerID=40&md5=ebc0941ae445fe7997066487c0108294.

[72] Sun, D., Cho, M., and Hwang, I., "On attack detection and identification for the cyber–physical system using lifted system model," *Automatica*, vol. 176, 2025. DOI: 10.1016/j.automatica.2025.112243. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-105000842080&doi=10.1016%2fj.automatica.2025.112243&partnerID=40&md5=1ca057402feab68dd00c03218dcb52d6.

[73] I. Jovanov and M. Pajic, "Relaxing integrity requirements for attack-resilient cyber-physical systems," *IEEE Transactions on Automatic Control*, vol. 64, no. 12, pp. 4843–4858, Dec. 2019. DOI: 10.1109/TAC.2019.2898510.

[74] Fan, X.-Y. and Lin, W.-J., "Reachable set control for cyber-physical systems with false data injection attacks," presented at the Proceedings - 2023 IEEE 6th International Conference on Industrial Cyber-Physical Systems, ICPS 2023, 2023. DOI: 10.1109/ICPS58381.2023.10127989. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85163058876&doi=10.1109%2fICPS58381.2023.10127989&partnerID=40&md5=aed699ad761792c4420ac427ed32e081.

[75] Qiu, S. and Liu, H., "A new zonotope-based attack detection method for UAV," presented at the Chinese Control Conference, CCC, vol. 2022-July, 2022, pp. 4276–4280. DOI: 10.23919/CCC55666.2022.9902124. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85140482486&doi=10.23919%2fCCC55666.2022.9902124&partnerID=40&md5=9bd3f93303cc73d974a6f6e05a5e12c4.

[76] Wang, J. et al., "Covert attack detection and resilient control of quadcopter," *IEEE Transactions on Industrial Informatics*, vol. 21, no. 5, pp. 3924–3933, 2025. DOI: 10.1109/TII.2025.3528579. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85218769026&doi=10.1109%2fTII.2025.3528579&partnerID=40&md5=61e799c8e3383266419e2df8cf94152b.

[77] Eslami, A. and Khorasani, K., "Cyber-attack detection and isolation in event-based cyber-physical systems," presented at the Proceedings of the American Control Conference, 2024, pp. 3460–3466. DOI: 10.23919/ACC60939.2024.10644303. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85204468219&doi=10.23919%2fACC60939.2024.10644303&partnerID=40&md5=5ffadeedb440bd91d84c5410e5dd3a8d.

[78] Bahrami, M. and Jafarnejadsani, H., "Privacy-preserving stealthy attack detection in multi-agent control systems," presented at the Proceedings of the IEEE Conference on Decision and Control, vol. 2021-December, 2021, pp. 4194–4199. DOI: 10.1109/CDC45484.2021.9683427. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85125691405&doi=10.1109%2fCDC45484.2021.9683427&partnerID=40&md5=92a261c8ba4acb98043cd749ddd967eb.

[79] Q. Liu, Y. Long, T. Li, and C. L. P. Chen, "Attack tolerant fault detection for CPSs: An unknown input interval observer approach," *IEEE Transactions on Automation Science and Engineering*, vol. 22, pp. 1163–1172, 2025.

[80] Wang, J., Zheng, Y., Ding, J., Xie, X., and Zhang, W., "Multiasynchronous extended dissipative sliding mode control of LC circuits in grid-connected system under actuator attacks," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 72, no. 4, pp. 1609–1620, 2025. DOI: 10.1109/TCSI.2024.3517703. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-105001685778&doi=10.1109%2fTCSI.2024.3517703&partnerID=40&md5=06fbc2750170491b32fb7e942b49a47c.

[81] Y. Yan, P. Antsaklis, and V. Gupta, "A resilient design for cyber physical systems under attack," in *2017 American Control Conference (ACC)*, Journal Abbreviation: 2017 American Control Conference (ACC), May 24, 2017, pp. 4418–4423. DOI: 10.23919/ACC.2017.7963635.

[82] Ahmad, S. and Ahmed, H., "Robust intrusion detection for resilience enhancement of industrial control systems: An extended state observer approach," presented at the 2022 IEEE Texas Power and Energy Conference, TPEC 2022, 2022. DOI: 10.1109/TPEC54980.2022.9750751. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85128775901&doi=10.1109%2fTPEC54980.2022.9750751&partnerID=40&md5=3842adab1600e49b6a8c1d612c000eb4.

[83] Beaudet, A., Sicard, F., Escudero, C., and Zamai, E., "Process-aware model-based intrusion detection system on filtering approach: Further investigations," presented at the Proceedings of the IEEE International Conference on Industrial Technology, vol. 2020-February, 2020, pp. 310–315. DOI: 10.1109/ICIT45562.2020.9067195. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85084119915&doi=10.1109%2fICIT45562.2020.9067195&partnerID=40&md5=e0fc305e39cc9087e1d2fac7f1cc0862.

[84] Liu, H., Zhang, Y., Li, Y., and Niu, B., "Proactive attack detection scheme based on watermarking and moving target defense," *Automatica*, vol. 155, 2023. DOI: 10.1016/j.automatica.2023.111163. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85163519547&doi=10.1016%2fj.automatica.2023.111163&partnerID=40&md5=e53ad9c7782d7717a52ddc44030d9d4c.

[85] T. Rhouma, J. Y. Keller, K. Chabir, D. Sauter, and M. N. Abdelkrim, "Coding control signals and switching LQG controller for secure fault-tolerant control against stealthy false data injection," in *2016 3rd Conference on Control and Fault-Tolerant Systems (SysTol)*, Journal Abbreviation: 2016 3rd Conference on Control and Fault-Tolerant Systems (SysTol), Sep. 7, 2016, pp. 750–755. DOI: 10.1109/SYSTOL.2016.7739838.

[86] Y. Chen, S. Kar, and J. M. F. Moura, "Optimal attack strategies subject to detection constraints against cyber-physical systems," *IEEE Transactions on Control of Network Systems*, vol. 5, no. 3, pp. 1157–1168, Sep. 2018. DOI: 10.1109/TCNS.2017.2690399.

[87] Pei, C., Xiao, Y., Liang, W., and Han, X., "A deviation-based detection method against false data injection attacks in smart grid," *IEEE Access*, vol. 9, pp. 15 499–15 509, 2021. DOI: 10.1109/access.2021.3051155. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85099594437&doi=10.1109%2faccess.2021.3051155&partnerID=40&md5=265dd819e2aac2707de79a68e40c8ee5.

[88] Li, J. and Zhang, Y., "Detection of false data injection attack in CPS by adaptive unscented kalman filter," presented at the IECON Proceedings (Industrial Electronics Conference), 2023. DOI: 10.1109/IECON51785.2023.10312054. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85179524773&doi=10.1109%2fIECON51785.2023.10312054&partnerID=40&md5=7c3d49e3ce97d0bf72db7a4987a8ec21.

[89] Liu, X., Dong, S., Huang, J., Yang, L., and Tong, X., "Dynamic state estimation of smart grid based on CKF under false data injection attacks," presented at the Proceedings of the 16th IEEE Conference on Industrial Electronics and Applications, ICIEA 2021, 2021, pp. 1883–1888. DOI: 10.1109/ICIEA51954.2021.9516062. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85115448688&doi=10.1109%2fICIEA51954.2021.9516062&partnerID=40&md5=6a0d6f169fa0dbe97c8567dcf02e6702.

[90] Nanda, S. et al., "Gaussian filtering with false data injection and randomly delayed measurements," *IEEE Access*, vol. 11, pp. 88 637–88 648, 2023. DOI: 10.1109/ACCESS.2023.3305288. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85168275859&doi=10.1109%2fACCESS.2023.3305288&partnerID=40&md5=83840b80780cfdcbfcff017e05be5ef8.

[91] S. Khan, I. Hwang, and J. Goppert, "Robust state estimation in the presence of stealthy cyberattacks," in *2022 American Control Conference (ACC)*, Journal Abbreviation: 2022 American Control Conference (ACC), Jun. 8, 2022, pp. 304–309. DOI: 10.23919/ACC53348.2022.9867774.

[92] J. Pisani, G. Cavone, L. Giarré, and F. Pascucci, "Industrial control systems attack detection by hybrid digital twin," in *2023 European Control Conference (ECC)*, Journal Abbreviation: 2023 European Control Conference (ECC), Jun. 13, 2023, pp. 1–6. DOI: 10.23919/ECC57647.2023.10178183.

[93] Á. M. Gómez, L. Navarro-Hilfiker, and R. Wiśniewski, "Model-based detection of data-injection cyber-attacks on wind turbine controllers," in *2024 10th International Conference on Control, Decision and Information Technologies (CoDIT)*, Journal Abbreviation: 2024 10th International Conference on Control, Decision and Information Technologies (CoDIT), Jul. 1, 2024, pp. 1780–1785. DOI: 10.1109/CoDIT62066.2024.10708323.

[94] Huang, X., Ding, D.-W., Ren, Y., and Xie, X., "Attack detection of complex cyber-physical networks via set-membership theory," *IEEE Transactions on Network Science and Engineering*, vol. 10, no. 4, pp. 2283–2293, 2023. DOI: 10.1109/TNSE.2023.3245214. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85149414924&doi=10.1109%2fTNSE.2023.3245214&partnerID=40&md5=01edab9a98e6b02f6721bfe411831439.

[95] Wang, P., Ren, X., and Wang, D., "Nonlinear cyber-physical system security control under false data injection attack," presented at the Chinese Control Conference, CCC, vol. 2022-July, 2022, pp. 4311–4316. DOI: 10.23919/CCC55666.2022.9902555. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85140481263&doi=10.23919%2fCCC55666.2022.9902555&partnerID=40&md5=e0191709514598d09e012ecedfdbae8e.

[96] Yang, H., Yin, S., Han, H., and Sun, H., "Sparse actuator and sensor attacks reconstruction for linear cyber-physical systems with sliding mode observer," *IEEE Transactions on Industrial Informatics*, vol. 18, no. 6, pp. 3873–3884, 2022. DOI: 10.1109/TII.2021.3111221. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85114735782&doi=10.1109%2fTII.2021.3111221&partnerID=40&md5=253061fb0d55528459bb27f4e2c0e662.

[97] L. Bin Bashar et al., "Observer-based attack and disturbance mitigation for heavy-duty vehicles in an autonomous docking application," *IEEE Access*, vol. 12, pp. 142 488–142 504, 2024. DOI: 10.1109/ACCESS.2024.3467263.

[98] X. Wang, X. Luo, M. Zhang, Z. Jiang, and X. Guan, "Detection and isolation of false data injection attacks in smart grid via unknown input interval observer," *IEEE Internet of Things Journal*, vol. 7, no. 4, pp. 3214–3229, Apr. 2020. DOI: 10.1109/JIOT.2020.2966221.

[99] Yang, H., Han, H., Yin, S., Han, H., and Wang, P., "Sliding mode-based adaptive resilient control for markovian jump cyber–physical systems in face of simultaneous actuator and sensor attacks," *Automatica*, vol. 142, 2022. DOI: 10.1016/j.automatica.2022.110345. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85129516600&doi=10.1016%2fj.automatica.2022.110345&partnerID=40&md5=e3bdbcb9b072d221358879e9dd319460.

[100] Fattore, G. and Valcher, M. E., "A data-driven approach to UIO-based fault diagnosis," en, in *2024 IEEE 63rd Conference on Decision and Control (CDC)*, Milan, Italy: IEEE, Dec. 2024, pp. 6831–6836. DOI: 10.1109/CDC56724.2024.10886303. Accessed: Jul. 2, 2025. [Online]. Available: https://ieeexplore.ieee.org/document/10886303/.

[101] Yang, G., Herrera, L., and Yao, X., "False Data Injection Attack Detection in DC Microgrids Based on Data-Driven Unknown Input Observers," *IEEE Journal of Emerging and Selected Topics in Power Electronics*, 2025. DOI: 10.1109/JESTPE.2025.3539958. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85217523555&doi=10.1109%2fJESTPE.2025.3539958&partnerID=40&md5=d26fb059848670a429ec371032343010.

[102] Wang, X., Wang, X., Luo, X., Guan, X., and Wang, S., "Novel cyber-physical collaborative detection and localization method against dynamic load altering attacks in smart energy grids," *Global Energy Interconnection*, vol. 7, no. 3, pp. 362–376, 2024. DOI: 10.1016/j.gloei.2024.06.003. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85197504950&doi=10.1016%2fj.gloei.2024.06.003&partnerID=40&md5=cc5d027f9f498e13b1ca396961f04a50.

[103] Gao, B., Bu, B., and Wang, X., "A comprehensive resilient control strategy for CBTC systems through train-to-train communications under malicious attacks," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 11, pp. 21 015–21 033, 2022. DOI: 10.1109/TITS.2022.3192510. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-

85135756590&doi=10.1109%2fTITS.2022.3192510&partnerID=40&md5=f7efcb044835ada297b0139fd389b045.

[104] Menelaou, C., Zhang, K., Timotheou, S., Panayiotou, C. G., and Parisini, T., "Stealthy attack detection of controlled ramp meters in freeway networks," in *2024 IEEE 63rd Conference on Decision and Control (CDC)*, IEEE, 2024, pp. 1289–1294.

[105] Kumar, P., Lin, Y., Bai, G., Paverd, A., Dong, J. S., and Martin, A., "Smart grid metering networks: A survey on security, privacy and open research issues," *IEEE Communications Surveys & Tutorials*, vol. 21, no. 3, pp. 2886–2927, 2019.

[106] Gebremichael, T. et al., "Security and privacy in the industrial internet of things: Current standards and future challenges," *IEEE Access*, vol. 8, pp. 152 351–152 366, 2020.

[107] Godsil, C. and Royle, G. F., *Algebraic graph theory*. Springer Science & Business Media, 2001, vol. 207.

[108] Wu, C. W. and Chua, L. O., "A unified framework for synchronization and control of dynamical systems," *International Journal of Bifurcation and Chaos*, vol. 4, no. 04, pp. 979–998, 1994.

[109] Wu, C. W., *Synchronization in complex networks of nonlinear dynamical systems*. World scientific, 2007.

[110] Liu, X., Li, C., Ge, S. S., and Li, D., "Time-synchronized control of chaotic systems in secure communication," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 69, no. 9, pp. 3748–3761, 2022.

[111] Fiore, D. and Russo, G., "Resilient consensus for multi-agent systems subject to differential privacy requirements," *Automatica*, vol. 106, pp. 18–26, 2019.

[112] Rezazadeh, N. and Kia, S. S., "Privacy preservation in a continuous-time static average consensus algorithm over directed graphs," in *2018 Annual American Control Conference (ACC)*, IEEE, 2018, pp. 5890–5895.

[113] Matthews, R., "On the derivation of a "chaotic" encryption algorithm," *Cryptologia*, vol. 13, no. 1, pp. 29–42, 1989.

[114] Schmitz, R., "Use of chaotic dynamical systems in cryptography," *Journal of the Franklin Institute*, vol. 338, no. 4, pp. 429–441, 2001.

[115] De la Hoz, M. Z., Acho, L., and Vidal, Y., "A modified chua chaotic oscillator and its application to secure communications," *Applied Mathematics and Computation*, vol. 247, pp. 712–722, 2014.

[116] Loan, N. A., Hurrah, N. N., Parah, S. A., Lee, J. W., Sheikh, J. A., and Bhat, G. M., "Secure and robust digital image watermarking using coefficient differencing and chaotic encryption," *IEEE Access*, vol. 6, pp. 19 876–19 897, 2018.

[117] Wang, M., Wang, X., Zhang, Y., and Gao, Z., "A novel chaotic encryption scheme based on image segmentation and multiple diffusion models," *Optics & Laser Technology*, vol. 108, pp. 558–573, 2018.

[118] Rainer, H. and Krause, U., "Opinion dynamics and bounded confidence: Models, analysis and simulation," 2002.

[119] Jia, P., MirTabatabaei, A., Friedkin, N. E., and Bullo, F., "Opinion dynamics and the evolution of social power in influence networks," *SIAM review*, vol. 57, no. 3, pp. 367–397, 2015.

[120] Rinaldi, S., Peerenboom, J., and Kelly, T., "Identifying, understanding, and analyzing critical infrastructure interdependencies," *IEEE Control Systems Magazine*, vol. 21, no. 6, pp. 11–25, 2001.

[121] Pederson, P., Dudenhoeffer, D., Hartley, S., and Permann, M., "Critical infrastructure interdependency modeling: A survey of us and international research," *Idaho National Laboratory*, vol. 25, p. 27, 2006.

[122] Satumtira, G. and Dueñas-Osorio, L., "Synthesis of modeling and simulation methods on critical infrastructure interdependencies research," *Sustainable and Resilient Critical Infrastructure Systems: Simulation, Modeling, and Intelligent Engineering*, pp. 1–51, 2010.

[123] Brown, T., Beyeler, W., and Barton, D., "Assessing infrastructure interdependencies: The challenge of risk analysis for complex adaptive systems," *International Journal of Critical Infrastructures*, vol. 1, no. 1, pp. 108–117, 2004.

[124] Liu, R. and Srivastava, A., "Integrated simulation to analyze the impact of cyber-attacks on the power grid," in *2015 Workshop on Modeling and Simulation of Cyber-Physical Energy Systems (MSCPES)*, IEEE, 2015, pp. 1–6.

[125] Drias, Z., Serhrouchni, A., and Vogel, O., "Analysis of cyber security for industrial control systems," in *2015 international conference on cyber security of smart cities, industrial control system and communications (ssic)*, IEEE, 2015, pp. 1–8.

[126] Bhamare, D., Zolanvari, M., Erbad, A., Jain, R., Khan, K., and Meskin, N., "Cybersecurity for industrial control systems: A survey," *computers & security*, vol. 89, p. 101 677, 2020.

[127] Masucci, D., Palazzo, C., Foglietta, C., and Panzieri, S., "Enhancing decision support with interdependency modeling," in *International Conference on Critical Infrastructure Protection*, Springer, 2016, pp. 169–183.

[128] Oliva, G., Panzieri, S., and Setola, R., "Modeling and simulation of critical infrastructures," *WIT Transactions on State-of-the-art in Science and Engineering*, vol. 54, 2012.

[129] Government, A., *Attorney-General's Department: CIPMA: Critical Infrastructure, Program for Modelling and Analysis, 20 April 2015*, https://www.tisn.gov.au/Documents/CIPMA-flyer.PDF, 2015.

[130] Amélie, G., Aurélia, B., Emmanuel, L., Mohamed, E., and Gilles, D., "The challenge of critical infrastructure dependency modelling and simulation for emergency management and decision making by the civil security authorities," in *International Conference on Critical Information Infrastructures Security*, Springer, 2016, pp. 255–258.

[131] Pathirana, A., "Epanet2 desktop application for pressure driven demand modeling," *Water Distribution Systems Analysis*, vol. 2010, pp. 65–74, 2010.

[132] Chen, J., "Cybergrid: Teaching and experiencing international teamwork through virtual environment," 2017.

[133] Ficco, M., Choraś, M., and Kozik, R., "Simulation platform for cyber-security and vulnerability analysis of critical infrastructures," *Journal of Computational Science*, vol. 22, pp. 179–186, 2017. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1877750317303605.

[134] Balakrishnan, S. and Zhang, Z., "Modeling interdependent effects of infrastructure failures using imprecise dependency information," *Sustainable and Resilient Infrastructure*, vol. 7, no. 2, pp. 153–169, 2022.

[135] Kovacevic, A. and Nikolic, D., "Cyber attacks on critical infrastructure: Review and challenges," *Handbook of Research on Digital Crime, Cyberspace Security, and Information Assurance*, pp. 1–18, 2015.

[136] Johnson, T. A., *Cybersecurity: Protecting critical infrastructures from cyber attack and cyber warfare*. CRC Press, 2015.

[137] Lehto, M., "Cyber-attacks against critical infrastructure," in *Cyber security: Critical infrastructure protection*, Springer, 2022, pp. 3–42.

[138] Sun, W., Bocchini, P., and Davison, B. D., "Overview of interdependency models of critical infrastructure for resilience assessment," *Natural Hazards Review*, vol. 23, no. 1, p. 04 021 058, 2022.

[139] Reilly, A. C., Baroud, H., Flage, R., and Gerst, M. D., "Sources of uncertainty in interdependent infrastructure and their implications," *Reliability Engineering & System Safety*, vol. 213, p. 107 756, 2021.

==========

Observer Based Anomaly Detection and Isolation

Chapter 2

Introduction to Observer-Based Attack Detection and Isolation

The increasing integration of computation, communication, and control in modern industrial and critical systems has given rise to complex *cyber–physical systems* that are both powerful and vulnerable. While this tight coupling enables advanced automation, distributed intelligence, and real-time adaptation, it also exposes new attack surfaces: adversaries can now manipulate sensors, actuators, or communication channels to induce incorrect perceptions of the system state and thereby compromise control decisions. Detecting such manipulations as they manifest in the system dynamics is a foundational challenge in ensuring trustworthy autonomy.

Observer–Based Attack Detection and Isolation addresses this challenge by embedding physics–based knowledge of the system into the monitoring architecture. An observer acts as a dynamic replica of the process, generating state estimates and predicted outputs that serve as a formal *reference of truth*. By comparing these predictions with the actual sensor measurements, Observer–Based Attack Detection and Isolation produces *residuals* that encode discrepancies between expected and observed behavior. Under normal operating conditions, residuals remain consistent with known uncertainty bounds arising from process noise or modeling errors. When malicious manipulation occurs—whether through sensor falsification, actuator interference, or communication disruption—these residuals deviate systematically, revealing the presence and, under suitable conditions, the origin of the attack.

Compared with purely data-driven attack detection methods, observer-based approaches offer distinct advantages. First, they are inherently *interpretable*: every residual has a physical meaning and can be traced to specific states or interconnections, facilitating diagnosis and validation. Second, they allow for *formal guarantees*: detectability, isolability, and robustness can be analyzed rigorously through observability, rank, and invariant subspace conditions. Third, they *integrate seamlessly* with estimation and control loops: observers such as Kalman filters, Luenberger observers, Sliding Mode Observers, and Unknown Input Observers can be augmented to detect attacks, identify compromised components, and even reconstruct the injected signals for compensation or recovery.

Scalability is a critical requirement in the era of large-scale and networked control systems. Centralized monitoring—where all measurements are collected at a single location—is often impractical due to bandwidth constraints, delays, privacy concerns, and the risk of single points of failure. Observer-based detection naturally extends to *distributed and decentralized architectures*, where each subsystem maintains its own local observer and exchanges selective information with its neighbors. The structural properties of many engineered systems—such as sparsity, locality of interactions, and the decay of spatial correlations—can be systematically exploited to reduce computational and communication overheads while maintaining diagnostic performance. These principles underpin the scalable architectures developed later in this thesis.

Beyond simple attack flagging, modern Observer–Based Attack Detection and Isolation aims for *isolation* (localizing the compromised components) and *reconstruction* (estimating the attack signals themselves). These capabilities transform attack detection into a functional component of resilience: once the malicious influence is identified and reconstructed, the control system can reconfigure its operation to contain or neutralize the disturbance, maintaining safety and continuity of service. In this sense, Observer–Based Attack Detection and Isolation serves not merely as a detection layer, but as an active interface between monitoring, estimation, and control—a bridge between security and autonomy.

This chapter introduces the theoretical and methodological foundations of observer-based attack detection and isolation as adopted in this thesis. We first review the principal system models and attack abstractions used in the literature, focusing on linear time-invariant representations with stochastic or bounded uncertainties. Then, we discuss the main categories of attacks relevant to this framework—particularly *deception* and *denial-of-service* —and how they manifest dynamically in the measurement and actuation channels. Subsequently, we survey the major classes of observer-based detection architectures, from centralized to fully distributed, emphasizing their objectives, trade-offs, and performance guarantees. Finally, we position the original contributions of this thesis within this context, showing how structural sparsity, distributed estimation, and robust observer design enable scalable, explainable, and reconstructive security.

Relation to the Author's Publications. The state-of-the-art review presented in this chapter originates from the comprehensive survey currently under preparation in [J6], which systematically analyzes the evolution of *observer-based attack detection* techniques in cyber–physical systems, with particular emphasis on model-based architectures, attack taxonomies, and the interplay between estimation, detection, and resilience. Although still under preparation, this work provides the conceptual foundation for the literature synthesis developed in the present chapter.

Building upon this foundation, the methodological developments on *observer-based detection using Kalman filtering* derive from the investigations reported in [J4] and [J8]. The study presented in [J4], currently under review, investigates the emergence of *covariance sparsity* in large-scale dynamical systems and formalizes how spatial decay properties can be exploited to design scalable and locality-preserving estimation architectures. Complementarily, the work in [J8], presently under preparation, introduces a *clustering-based interpretation* of estimation interactions, showing how structural grouping of states and measurements enables modular and distributed observer design. Together, these contributions establish the theoretical foundations for *sparsity-aware* and *scalable* Kalman-based detection architectures.

Further extensions and implementation strategies were developed in the conference contributions [C2], [C5], and [C9], which introduced and analyzed the *Multirate Interlaced Kalman Filter* framework, providing scalable and stable solutions for distributed attack detection in cyber–physical systems.

The chapter of this part devoted to *robust and reconstructive detection via Sliding Mode Observers* originates from two complementary research efforts. First, the theoretical framework and distributed reconstruction method for sensor cyber-attacks were introduced in [C16], an accepted conference paper focusing on distributed SMO design and validation. Second, these ideas are being extended and generalized in the forthcoming journal paper [J9], currently under preparation, which formalizes the analytical reconstruction guarantees and scalability analysis for large-scale distributed Sliding Mode Observer architectures.

Organization of the Chapter. The remainder of this chapter is organized to provide both a conceptual overview and a methodological synthesis of observer-based attack detection and isolation (OBAD) in cyber–physical systems.

- **Section 2.1** introduces the theoretical underpinnings of model-based attack detection and isolation. It first reviews the system models and attack abstractions commonly adopted in the literature, emphasizing how deception and denial-of-service behaviors manifest in system dynamics. The section then surveys major observer-based detection paradigms—including statistical, set-based, sliding-mode, unknown-input, data-driven, and hybrid observers—highlighting their fundamental principles, design objectives, and resilience properties under adversarial conditions.
- **Section 2.2** contextualizes the original contributions of this thesis within the broader research landscape. It discusses how the proposed methodologies leverage structural

sparsity, distributed estimation, and robust observer design to achieve scalable, explainable, and reconstructive detection and isolation. The section also outlines how these ideas are developed in subsequent chapters on sparsity-aware Kalman architectures and sliding-mode reconstruction frameworks.

2.1 Literature review on Model Based Attack Detection and Isolation

In observer–based attack detection for cyber–physical systems, the *system model* serves as the fundamental anchor that formalizes nominal behavior and supports residual generation [46], [47], [140]. Throughout this chapter, we adopt the notation and modeling conventions introduced in Section 1.2, and, unless stated otherwise, we consider the discrete–time linear time–invariant model defined therein. linear time–invariant models are the de facto standard in observer–based attack detection because they combine analytical tractability, a mature theoretical foundation, and efficient computational implementations—enabling Kalman filtering, Luenberger observers, and residual–based tests with strong nominal guarantees [141]. A stochastic LTI system with process and measurement noise is given by (1.28).

Nonlinear and hybrid system descriptions are also relevant, for instance in robotics, aerospace, and mode–switching processes [142], [143]. However, their use in OBAD is often limited by identification complexity, computational overhead, and weaker general guarantees (e.g., nonconvexity and limited observability results) [144]. In practice, many such systems admit local linearizations or piecewise–affine surrogates that allow leveraging linear–system tools. For example, [144] shows that a Kalman filter designed on a linearized model can reveal false–data injection attacks on a nonlinear hybrid plant, while [143] learns affine hybrid automata from input–output data and analyzes them using linear techniques.

For these reasons, and consistently with the notation of Section 1.2, the survey and subsequent developments focus on linear models as the core paradigm for attack detection—acknowledging nonlinear and hybrid extensions only when they can be reduced to, or interfaced with, the linear framework in order to preserve scalability and real–time deployability.

Characterization of Cyberattacks for Observer-Based Attack Detection Deception and Denial-of-Service attacks represent the two most prevalent categories addressed by Observer-Based attack detection techniques [145], [146], [147]. Rather than classifying cyberattacks by their entry point—such as malware injection, firmware corruption, or physical manipulation—OBAD approaches focus on the attacks' observable effects on the system's dynamics. This system-theoretic perspective enables abstraction into two primary functional categories: deception attacks, which compromise the integrity of sensor readings or actuator commands [94], [145], [146], and DoS attacks, which compromise the availability of information or control signals [57], [84], [148].

Stealthy deception attacks are particularly dangerous, as they can be designed to remain within the system's expected behavior bounds—escaping detection while still degrading system performance over time [76], [94], [104]. These attacks often exploit knowledge of the system model to introduce malicious inputs that avoid triggering residual thresholds. On the other hand, DoS attacks disrupt communication by blocking or delaying the transmission of control or measurement data, which can severely impact control performance and observability [53], [80], [149].

This functional classification is central to OBAD design, as it directly informs the structure of residual generators and observers. By abstracting cyber threats based on their impact on system dynamics, rather than their implementation layer, it becomes possible to generalize detection strategies across a wide range of CPS architectures and attack vectors.

This abstraction reflects a central insight in OBAD: regardless of whether an attacker penetrates the system through the software stack, injects malicious code in a PLC, or manipulates physical sensors, what ultimately matters is how the attack alters the input-output behavior of the system [31], [67], [150]. This behavioral focus is a key enabler for system-theoretic detection methods, which rely on residuals, observers, or estimators to detect deviations from nominal dynamics [17], [33], [47].

The proposed taxonomy distinguishes attacks only by their functional impact—deception or denial—because this distinction is directly relevant to Observer-Based detection frameworks. Other forms of attack, such as those aimed at *confidentiality* (e.g., eavesdropping or information leakage), do not produce an observable deviation in the system dynamics, and are therefore not amenable to detection via OBAD techniques. These privacy-focused attacks are more effectively studied using discrete event system (DES) models and symbolic methods [151].

Additional attack surfaces may exist within a CPS, such as compromising the controller logic itself. For example, an attacker might exploit a vulnerability in the firmware of a PLC to alter control laws or gain unauthorized access to system variables [152]. However, even in these cases, the practical manifestation of the attack typically affects either the sensor network (by falsifying measurements) or the actuator commands (by injecting incorrect control signals). Consequently, such controller-level attacks still reduce, from a modeling standpoint, to deception or DoS patterns. This reinforces the usefulness of the two-class taxonomy in capturing the vast majority of attack scenarios relevant to Observer-Based security [153], [154].

Deception Attacks: Deception attacks aim to subtly alter the information used for decision-making by injecting false sensor data or modifying control inputs. The goal is to mislead the estimator or controller into perceiving an incorrect system state, without raising alarms. These attacks are often carefully crafted to remain consistent with the system dynamics and expected noise levels, making them difficult to detect [47], [104], [155].

Mathematically, a deception attack can be described by modifying the sensor output or the control input:

$$\boldsymbol{y}_k^a = \boldsymbol{y}_k + a_k^y, \tag{2.1}$$
$$\boldsymbol{u}_k^a = \boldsymbol{u}_k + a_k^u, \tag{2.2}$$

where a_k^y and a_k^u represent the additive attack signals on the sensor and actuator channels, respectively. These can be deterministic, stochastic, or designed adversarially to satisfy stealth conditions [94].

Replay attacks correspond to a special case of deception where historical data is re-injected into the system:

$$\boldsymbol{y}_k^a = \boldsymbol{y}_{k-T}, \tag{2.3}$$

with T representing the replay delay [37]. Bias injection is another subclass where the attacker introduces a constant or slowly varying offset:

$$\boldsymbol{y}_k^a = \boldsymbol{y}_k + b, \quad b \in \mathbb{R}^n. \tag{2.4}$$

Stealthy deception attacks are those that satisfy:

$$|\boldsymbol{r}_k| < \eta \quad \text{for all } k, \tag{2.5}$$

where $\boldsymbol{r}_k$ is the detection residual and η is the detection threshold, meaning the attack remains undetected while still influencing system behavior [46], [78].

These strategies can be extremely effective when the adversary possesses knowledge of the system model, but their success hinges on precise alignment with system dynamics. Imperfect knowledge can cause detectable anomalies, reducing the attack's stealth.

Denial-of-Service Attacks: DoS attacks obstruct the transmission of sensor or actuator data, leading to partial or complete information loss. These attacks are typically modeled as multiplicative processes that disable the normal signal:

$$\boldsymbol{y}_k^a = \delta_k^y \boldsymbol{y}_k, \tag{2.6}$$
$$\boldsymbol{u}_k^a = \delta_k^u \boldsymbol{u}_k, \tag{2.7}$$

where $\delta_k^y, \delta_k^u \in 0, 1$ are binary indicators representing successful transmission (1) or loss (0). This abstraction captures both deterministic dropouts (e.g., periodic jamming) and stochastic packet loss (e.g., Bernoulli or Markovian processes) [12], [53], [149].

More generally, DoS attacks can also be described using additive disruption models:

$$\boldsymbol{y}_k^a = \boldsymbol{y}_k + \eta_k^y, \tag{2.8}$$
$$\boldsymbol{u}_k^a = \boldsymbol{u}_k + \eta_k^u, \tag{2.9}$$

where η_k^y, η_k^u represent injected noise or irrelevant data intended to overwhelm the receiver or obscure the true signal [55], [156].

Compared to deception attacks, DoS strategies are less dependent on detailed system knowledge and are thus easier to implement in practice. However, they often have a more abrupt impact on system performance, potentially making them easier to detect through residual analysis or degradation metrics [103], [157].

Detection Architectures in Cyber-Physical Systems Security The architectural design of detection mechanisms in Cyber-Physical Systems plays a pivotal role in shaping the performance, scalability, and resilience of security solutions. Broadly, three main architectural paradigms can be identified in the literature: centralized, decentralized, and distributed. Each of these paradigms exhibits unique advantages and limitations, which are often reflected in the types of systems they are applied to, and in the nature of attacks they are designed to detect.

In a *centralized* architecture, a single monitoring entity aggregates system-wide data to perform attack detection. This configuration provides full observability and lends itself well to mathematically rigorous solutions, but it introduces bottlenecks in scalability and creates a single point of failure. In contrast, *decentralized* architectures rely on individual subsystems to perform detection using only local data. While this preserves privacy and improves robustness to localized faults, it limits global observability and may fail to detect coordinated multi-point attacks. A middle ground is offered by *distributed* detection architectures, in which local detectors exchange information with their peers to build a cooperative detection strategy. This approach balances autonomy and observability and is gaining traction in large-scale CPS applications.

Centralized Detection Centralized architectures collect system-wide data at a single location for analysis. This global perspective allows for high-fidelity detection of subtle anomalies and the use of powerful estimation tools. For instance, [158] propose an observer-based detection method against actuator attacks, exploiting full model knowledge for precise fault localization, albeit with limited scalability. Similarly, [159] develop a secure estimation framework against FDI attacks using a central observer, offering high accuracy but depending on reliable communication links.

In smart grid applications, centralized methods are used for wide-area monitoring. [35], [160] investigate joint attack and fault detection via centralized residual generation, achieving robust performance but suffering from high data acquisition costs. SCADA systems are addressed in [161], which proposes detection algorithms based on the deviation of system dynamics, and [34], where stealthy data injection attacks are identified via optimization-based criteria.

Centralized methods also feature in industrial control systems. [33], [82] propose robust observers for secure monitoring, which are effective in clean environments but may struggle under model uncertainties. [36], [58] focus on state reconstruction under sensor deception attacks using Kalman filter enhancements, improving accuracy but assuming

a precise system model. [38] presents a joint detection scheme combining control and estimation to improve performance against stealthy threats.

In actuator-level protection, [92] applies centralized observers for industrial plants, and [162] uses sliding-mode observers for resilience, both showing effective detection but increased sensitivity to measurement noise. Code-based attack detection is discussed in [85], while [163] integrate feedback control and attack detection for robustness.

Other centralized approaches include [164] (fuzzy observer-based detection), [165] (sensor fault diagnosis), [166] (adaptive fault-tolerant control), and [167] (switching observer designs). Despite strong theoretical foundations, these methods typically suffer from high computational burden and limited scalability.

Decentralized Detection Decentralized detection shifts the analysis to local subsystems, allowing for autonomous operation and reduced communication load. This architecture is ideal for privacy-preserving and communication-limited applications. For example, [78] propose decentralized attack detection for multi-agent systems using information-theoretic privacy metrics. The benefit is local autonomy, though coordination is absent.

In [77], a disconnection-aware event-triggered scheme is developed, detecting actuator attacks with limited data sharing. [72] develop a decentralized detection mechanism for stealthy sensor attacks in large-scale systems, favoring speed over accuracy. [168] address detection in partially disconnected networks, tolerating communication faults at the expense of decreased detection rates. [157] explores decentralized observer synthesis for systems with bounded network disruptions.

[169] focus on stealthy attack identification using distributed residual analysis, showing low detection delay under certain attack models but requiring tightly synchronized event triggers.

The key limitation across these works is the inability to capture correlated or coordinated attacks, which remain invisible without system-wide data correlation.

Distributed Detection: Distributed architectures enable each subsystem to process local data while exchanging selective information with neighbors. This fosters scalability and robustness, especially in large, interconnected CPSs.

[65] propose a distributed observer that combines local filters with sparse communication, allowing for resilient estimation against joint sensor-actuator attacks. [73] focus on relaxing communication requirements via asynchronous message-passing, reducing bandwidth but requiring conservative detection thresholds.

Smart grid applications include [45], which designs a distributed estimator with event-based communication to detect sensor attacks efficiently, and [103], which surveys secure architectures for microgrids, emphasizing detection modularity and fault tolerance.

Fault-tolerant strategies are further developed in [15], [56], [170], with local observers exchanging residuals to localize faults. [37], [44], [68] introduce sparse recovery techniques to identify attack vectors using networked estimators. [40] proposes distributed estimation strategies under partial communication failures.

Dynamic attack detection in smart grids is presented in [171], and [75] propose attack detectors that adapt to system configuration changes. Robustness to communication attacks is addressed in [17], [147], which rely on redundancy to mitigate false information propagation. [172] addresses covert attack detection using multi-agent interaction.

In vehicular networks, [173], [174] explore distributed Kalman filters for attack detection. Stealthy attacks are handled in [175], [176], which combine Observer-Based and data-driven filters to balance performance and robustness.

Collaborative estimation schemes, such as [19], [70], develop secure consensus algorithms, while broader architectural designs are presented in [81], [177], [178], [179]. Large-scale architectures are surveyed in [2], [180], and adaptive detection strategies are addressed in [181], [182], [183], [184], [185], [186], [187].

While distributed detection offers a compelling trade-off between local autonomy and global insight, it depends heavily on reliable communication and synchronization. Furthermore, inter-node communication may itself be a vulnerability, requiring secure communication protocols and authentication mechanisms.

Objectives of Observer-Based Security Strategies Observer-based strategies in cyber-physical systems are developed to meet a range of security objectives, from detecting the presence of an attack to mitigating its impact through robust control. Each objective brings unique benefits, but also entails trade-offs in terms of complexity, timeliness, and robustness.

Detection: Detection is the fundamental entry point for any security architecture. It involves identifying anomalies between predicted and actual system behavior, often using residuals from an observer. The main advantage of detection is its relative simplicity and fast response: it enables early awareness without requiring full knowledge of the attack model. However, detection alone cannot determine the source, nature, or severity of the attack, and may generate false alarms in the presence of noise or modeling uncertainties.

Several observer-based detection methods target deception attacks. For instance, [155] proposes a finite-time observer that enables rapid detection by quickly estimating the system output and flagging deviations. This approach is beneficial for time-critical systems, but its finite-time design can be sensitive to model inaccuracies. [34], [46] introduce statistically grounded residual-based detectors, offering robustness to noise but requiring a priori knowledge of system dynamics and statistical thresholds. [76] designs a nonlinear observer for UAVs that can detect covert attacks by capturing subtle disturbances, which improves stealth detection but may be computationally demanding. Similarly, [17] proposes switching-based observers that adapt to changes in system dynamics, improving detection under varying conditions, though at the cost of increased complexity. [33], [94] further explore robust observer designs that improve detection under model uncertainties.

DoS detection typically focuses on identifying missing or delayed data. [57] uses sliding-mode innovation sequences to detect communication failures in UAVs. The approach is robust to noise but may introduce chattering in estimates. [47] offers a memoryless observer that does not rely on past measurements, enabling fast response during severe packet loss, although with limited performance under high-frequency disturbances. [148] extends this to design robust observers that maintain performance despite packet losses.

General-purpose detectors such as those by [35], [160] apply statistical methods to distinguish between faults and attacks, ensuring broad applicability but possibly increasing false positives. [161] applies detection to smart grids, but its applicability depends on system observability assumptions. [73] alleviates such assumptions, improving generalizability but requiring more complex system modeling. Finally, [146], [147] introduce active coding and learning-enhanced techniques, injecting excitation signals or leveraging system behavior to improve detectability. While this enhances sensitivity to attacks, it may affect system performance if not carefully tuned.

Identification: Identification enhances detection by classifying the type of attack or identifying the subsystem affected. Its primary advantage is that it allows tailoring mitigation strategies to the specific threat. However, it is often more computationally intensive and may rely on detailed system knowledge or residual pattern matching, which limits its applicability in large-scale or highly nonlinear systems.

[75] introduces a general framework that classifies attacks based on residual features. This improves flexibility but may suffer from ambiguity when multiple attacks produce similar residuals. [188] applies a Luenberger observer to vehicle platooning, using the spatial distribution of residuals to distinguish between sensor and actuator faults. This spatial sensitivity is valuable for distributed systems but requires synchronized communication. [189] uses switching logic to identify attack modes in switched systems, which is effective for hybrid systems but less so in continuous or adaptive contexts. [36] extends classification to incorporate detection confidence levels, helping prioritize responses.

In the DoS context, [10] proposes the use of attack graphs and structural observability analysis. This approach provides rigorous guarantees and supports Observer-Based classification, but its reliance on accurate topology information can be a limitation in real-time or partially observable scenarios.

While identification improves system awareness and supports precision defense, its implementation can be complex and may lag behind fast-evolving attacks if based solely

on residual signatures.

Isolation and Reconstruction: Isolation focuses on pinpointing the location of an attack within the system, whereas reconstruction aims to estimate the actual injected signal. These objectives support fine-grained responses and enable compensation or forensic analysis. Their strength lies in the depth of information they provide; however, they are typically more demanding in terms of observer design and often require redundant sensing or modeling assumptions.

[43] uses a bank of subsystem-specific observers to isolate stealthy FDI attacks. This modular structure enhances scalability and fault localization, but it assumes non-overlapping faults and may be sensitive to inter-subsystem coupling. [98] adopts a similar approach for smart grids, using redundancy and residual comparison to isolate attacks. While effective, the need for redundant observers increases the system's computational overhead. In the case of DoS, [57] adapts innovation sequence analysis to localize compromised channels, a low-cost yet limited solution in highly interconnected systems.

Reconstruction techniques provide even more detailed insights. [65] reconstructs biasing attacks by estimating their magnitude along with the system state, enabling real-time compensation. However, reconstruction performance may degrade in the presence of strong noise or multiple overlapping attacks. [159] enhances this by using hybrid observers that can handle intermittent observability, suitable for systems under sporadic DoS attacks, though requiring precise timing models. [41], [49] extend resilient estimation by incorporating attack reconstruction under sensor faults, enabling robust control but demanding careful tuning of observer gains. [165], [187] address networked systems by dynamically adjusting observer gains to compensate for packet loss, though such adaptability may add instability if over-reactive.

General approaches such as those by [44], [96] use sparse optimization to recover sparse attack signals, offering robustness to multi-point attacks, yet their real-time applicability remains limited due to computational complexity. [99] introduces a sliding-mode observer that offers robustness to high-frequency attack signals, though susceptible to chattering. [97] combines state estimation and reconstruction in a unified framework, reducing design complexity but possibly reducing modularity. [155] achieves finite-time reconstruction, offering fast recovery at the expense of higher sensitivity to modeling errors. [157] further integrates observer-based estimation with model-driven compensation strategies.

In summary, isolation and reconstruction offer powerful tools for attack diagnosis and recovery, but require careful trade-offs between complexity, speed, and robustness.

Mitigation, Secure Estimation, and Secure Control: Mitigation represents the most advanced objective: to preserve system functionality in the presence of attacks. This can involve reconfiguring control, adapting observers, or filtering corrupted data. The advantage is clear: mitigation maintains performance even under sustained attacks. However, these approaches often require co-design of observers and controllers, and their success depends on the quality of detection and reconstruction layers.

[164] presents a switching-based control system that dynamically adapts to attacks. This strategy provides agility in real-time, though it risks instability due to mode-switching delays. [97] proposes a co-design framework that tightly integrates estimation and control, enabling coordinated response to attacks. This reduces latency and enhances robustness but may limit modular upgrades to estimation or control subsystems.

Secure estimation forms the backbone of mitigation. [43], [146] design observers that remain accurate even under stealthy attacks. Their resilience improves state reliability but may introduce estimation delay. [178] offers a finite-state observer resistant to intermittent attacks, especially suitable for discrete-time systems but less flexible for dynamic ones. [12], [148] provide robust observers for DoS conditions, maintaining performance despite packet losses. However, performance may degrade if the DoS duration exceeds the observer's horizon.

Secure control extends these ideas to the control loop. [65] leverages reconstructed attacks for feedback compensation, which enhances robustness but requires high-fidelity reconstruction. [76] implements a covert-attack-aware controller for UAVs, improving stealth detection response but requiring accurate dynamics models. [99] integrates

sliding-mode estimation and control to handle high-frequency perturbations, providing strong robustness but limited smoothness. [41] fuses estimation with control in an integrated framework that resists sensor attacks. [190] introduces predefined-time control, ensuring recovery within a guaranteed time frame, ideal for safety-critical systems, though challenging to tune conservatively. [186] contributes additional strategies for resilience in feedback control under structured attack scenarios.

Overall, mitigation strategies embody the shift from reactive to proactive security. While powerful, their effectiveness relies on the accuracy and timeliness of detection, estimation, and reconstruction processes. Designing such systems requires balancing responsiveness, robustness, and implementation overhead.

2.1.1 Observer-Based Detection Strategies

In this section, we review and discuss the main families of observer-based approaches proposed in the literature, highlighting the underlying estimation philosophies that guide their formulation. Particular attention is given to how these observers differ in terms of their model assumptions, estimation mechanisms, and suitability for various types of attacks such as *detection*, *deception*, and *denial-of-service* (DoS).

Statistical Observers

Statistical observers—exemplified by the Kalman filter and its numerous extensions—represent one of the most mature and widely adopted estimation paradigms in cyber–physical systems. Their success lies in strong theoretical foundations: under linear Gaussian assumptions, they yield minimum-variance, unbiased estimates, thereby defining an "optimal" solution in a stochastic sense. Their recursive prediction–correction architecture ensures computational efficiency and analytical tractability, supporting real-time deployment in safety-critical infrastructures. Within this framework, residual-based monitoring naturally emerges as a powerful tool for identifying *detection*-type attacks, particularly integrity violations and false data injection (FDI) attacks, through the statistical analysis of innovation sequences [34], [35], [191], [192].

However, this theoretical elegance conceals several structural fragilities. First, the underlying linear–Gaussian hypothesis assumes that disturbances are random, memoryless, and independent—an assumption incompatible with the strategic and correlated nature of cyberattacks. When facing stealthy or *deception*-type attacks that emulate nominal noise statistics, such as those described in [175], [193], the residual remains statistically indistinguishable from normal operation, leading to complete detection failure. This limitation is intrinsic: a purely stochastic detector cannot distinguish adversarial manipulation from legitimate randomness when their statistical footprints coincide.

A second, equally fundamental limitation concerns model dependency. The optimality of the Kalman filter collapses under model mismatch, parameter drift, or unmodeled nonlinearities—conditions ubiquitous in large-scale or evolving CPS. In [192] and [19], it is shown that even moderate inaccuracies in system matrices or covariance estimates can trigger false alarms or render the residual statistically meaningless. These vulnerabilities become acute in distributed or networked architectures, where delays, packet drops, and heterogeneous sensors further disrupt the statistical consistency required for reliable estimation.

Efforts to enhance robustness have led to hybrid formulations that blend stochastic estimation with deterministic correction mechanisms. For instance, adaptive and unscented Kalman filters have been proposed to counter nonlinearities and time-varying disturbances [12], [88], while distributed and fusion-based variants attempt to contain attack propagation across networked nodes [194], [195]. Nonetheless, even these enhanced designs remain bounded by their dependence on an accurate nominal model and the validity of Gaussian noise assumptions. In [17] and [87], hybrid Kalman formulations demonstrate improved resilience against actuator and sensor FDI attacks, yet their performance deteriorates as the proportion of compromised measurements increases.

A distinct weakness emerges under *denial-of-service* conditions. Statistical observers rely on continuous measurement updates for correction. When packet loss or communication interruptions occur, the prediction covariance grows unbounded, ultimately leading to estimator divergence. In such cases, detection itself is trivial—the system simply recognizes missing data—but maintaining a secure and bounded estimate becomes impossible. This fragility is well documented in [55], [180], which highlight how Kalman filters cannot ensure safe estimation continuity during sustained DoS events. Even when augmented with adaptive or reinitialization mechanisms, these filters can only delay, not prevent, divergence.

Conceptually, statistical observers reflect an estimation philosophy focused on probabilistic optimality rather than resilience. They assume benign stochasticity, not strategic manipulation. This paradigm is appropriate in classical control or estimation theory but misaligned with adversarial environments where uncertainty is intentional and coordinated. Studies such as [86], [175] further emphasize that adversaries can optimize their attack sequences to remain undetected by Kalman-based residual thresholds, exploiting the filter's reliance on Gaussian innovations.

From a broader security perspective, the Kalman family's strengths—analytical tractability, optimality, and simplicity—become liabilities under adversarial uncertainty. They are reactive tools: capable of signaling statistically evident anomalies, but incapable of maintaining situational awareness when data integrity or availability collapses. Consequently, while Kalman filters remain indispensable for real-time monitoring and *detection*-type attack identification, they are structurally inadequate for active *deception* or prolonged *DoS* scenarios. Ensuring estimation continuity and control safety under such conditions requires hybridization with robust, set-based, or sliding-mode methodologies that extend beyond purely statistical paradigms [43], [57], [100], [101].

Set-Based Observers

Set-based observers embody a deterministic estimation paradigm in which all possible system trajectories consistent with the dynamics, inputs, and bounded uncertainties are enclosed within a well-defined geometric set. Unlike statistical observers that rely on stochastic assumptions, these approaches provide guaranteed estimation bounds under unknown-but-bounded (UBB) noise conditions. This philosophy of estimation is particularly suitable for adversarial environments, where uncertainties are deliberate rather than random, and the goal is to ensure *guaranteed safety* rather than probabilistic optimality.

Recent developments in the literature reveal three main classes of set-based formulations: zonotopic, ellipsoidal, and interval-hull observers. Each class offers distinct trade-offs between computational efficiency, conservatism, and robustness. Zonotopic observers have emerged as one of the most influential methods, owing to their closure properties under linear transformations and Minkowski sums, enabling efficient over-approximation of reachable states in real time. In [37], a zonotopic observer is employed to detect false data injection attacks by continuously computing tight overapproximations of the system's reachable set. This work demonstrates that the violation of set consistency between predicted and measured outputs is a strong indicator of integrity breaches, even in the presence of bounded disturbances. Similarly, [43] integrates a zonotopic unknown input observer to simultaneously achieve secure estimation, attack isolation, and reconstruction under sensor and actuator attacks, showing that the set-based framework can support not only detection but also post-attack recovery.

Alternative geometric representations have also gained traction. The ellipsoidal set-membership observer proposed in [36] formulates false data injection detection as an inclusion test within predicted confidence ellipsoids. This method leverages convex optimization and linear matrix inequalities to achieve computational tractability, but may produce overly conservative bounds in systems with strong cross-correlations. Interval-hull approaches, as developed in [94], address this issue by adopting a more flexible bounding strategy suitable for large-scale or complex networked CPS. The interval formulation

facilitates guaranteed enclosures of state trajectories while enabling efficient consistency checks for detecting deception and stealthy manipulations, as illustrated in their moving-target defense scheme.

Recent studies have expanded the applicability of set-based estimation beyond simple detection tasks. The encoding–decoding approach in [68] combines set-membership theory with information-theoretic encoding to improve resilience against stealthy attacks in nonlinear CPS. By encoding sensor outputs within bounded uncertainty envelopes, the observer reconstructs attack-free estimates at the decoder, effectively extending detection into a distributed architecture. The reachable set control paradigm proposed in [74] further illustrates how set-based reasoning can be extended from estimation to proactive safety control: through Lyapunov-based reachable set computation, the system ensures bounded behavior despite ongoing false data injection attacks, shifting the objective from detection to guaranteed containment.

Several other contributions emphasize the integration of set-based reasoning with robust control and hybrid estimation. For instance, [39] introduces invariant set detectors that combine Kalman filtering with robust invariant set theory, providing a bridge between stochastic and deterministic paradigms. Similarly, [75] applies a zonotope-based method for UAV systems to detect sensor deception attacks, demonstrating the suitability of these approaches for fast, embedded architectures where probabilistic computation is infeasible. Earlier works such as [38] and [140] established the theoretical foundations for combining Bayesian estimation and worst-case set analysis, showing how the reachable set formulation can quantify the maximum undetectable deviation under stealthy attack strategies. These studies collectively underscore that set-based estimation inherently provides robustness to deception attacks—by construction, any adversarial manipulation that drives measurements outside the predicted set becomes detectable.

Nevertheless, this robustness comes at the cost of computational intensity and potential conservatism. The propagation and intersection of high-dimensional zonotopes or ellipsoids can be computationally prohibitive, especially in real-time or large-scale systems. Over-approximations accumulate across iterations, potentially masking small attacks or generating false positives. Scalability issues are particularly prominent in networked CPS, where dimensionality grows with the number of agents or sensors. Moreover, under denial-of-service conditions, the absence of measurement updates prevents the contraction of the reachable set, leading to progressive uncertainty expansion. This limitation is reflected in [196], where decentralized resilient observers maintain bounded stability during packet losses but cannot guarantee accurate estimation without intermittent data recovery.

Overall, set-based observers offer a conceptually powerful and mathematically rigorous framework for secure state estimation. Their deterministic nature ensures guaranteed error bounds and makes them particularly suited for *deception* and *integrity* attack scenarios, where statistical detectors typically fail. However, their high computational complexity, conservatism, and sensitivity to prolonged data loss pose significant challenges for scalability and real-time deployment. Future research is moving toward hybrid architectures—combining zonotopic or ellipsoidal observers with adaptive and data-driven techniques—to balance deterministic robustness with the adaptability required by modern, distributed CPS environments.

Sliding Mode Observers

Sliding Mode Observers exploit discontinuous correction terms to ensure finite-time convergence of the estimation error, achieving robustness against matched uncertainties and bounded disturbances. Their fundamental principle is the enforcement of a sliding manifold along which estimation errors are driven to zero in finite time. This property makes SMOs particularly effective for reconstructing *deception*-type attacks, such as actuator and sensor manipulations, by simultaneously estimating both system states and attack inputs [40], [41], [45], [76], [96], [159].

Classical formulations, such as the Sliding Mode Integral Observer (SMIO) proposed in [197], laid the groundwork for fault detection and isolation in nonlinear systems through equivalent output error injection signals. Subsequent advancements have generalized these ideas to cyber–physical contexts, where SMOs are used not only for fault tolerance but also for attack estimation and control resilience. In [49], an adaptive SMO enables online reconstruction of actuator and sensor attacks in power systems, exploiting equivalent control principles to maintain stability despite uncertain disturbances. Similar adaptive and descriptor-based extensions are employed in [96], where sparse false data injection (FDI) attacks are reconstructed in linear CPSs through descriptor reduced-order observers, ensuring asymptotic convergence even under switched dynamics.

Recent developments have diversified SMO formulations toward improved robustness, finite-time convergence, and reduced chattering. The introduction of higher-order and terminal sliding-mode designs—such as the super-twisting nonsingular terminal SMO in [48]—addresses the classical chattering issue by replacing discontinuous switching with continuous control laws that preserve finite-time convergence. Fractional-order and nonsingular terminal observers, such as those proposed in [198], enhance resilience in nonlinear power systems under stealthy FDI attacks, while ensuring smoother convergence dynamics. Similarly, the adaptive integral SMO in [199] achieves optimal L_2-gain performance under large-scale disturbances, combining Lyapunov-based design with linear matrix inequality (LMI) constraints to balance estimation speed and noise suppression.

Hybrid architectures extend the SMO paradigm beyond single-agent systems. In [200], predefined-time distributed SMOs are employed for networked cooperative systems, enabling decentralized detection of local intrusions and communication attacks under switching topologies. The event-triggered and multiasynchronous dissipative sliding-mode designs proposed in [80] and [201] introduce asynchronous or mode-dependent switching mechanisms to reduce communication overhead in grid-connected systems, demonstrating that SMOs can adapt to time-varying network structures while preserving finite-time estimation guarantees. These distributed and event-triggered variants mark a key shift from centralized to scalable implementations, addressing one of the traditional weaknesses of classical SMOs.

In the context of power and energy systems, SMO-based secure estimation has proven particularly effective. Works such as [15], [40], [41], [170] integrate SMO-based estimation with resilient control strategies to achieve simultaneous attack reconstruction and system stabilization under FDI and dynamic load-altering attacks. These approaches highlight the dual utility of SMOs as both detection and compensation mechanisms: by reconstructing attack signals in real time, they enable immediate feedback correction, ensuring continued operational integrity even during active deception attempts.

However, SMOs are not without drawbacks. The primary limitation remains the phenomenon of *chattering*—undesirable high-frequency oscillations caused by the discontinuous control law—which can induce actuator wear and degrade estimation accuracy in high-bandwidth systems. Although higher-order and adaptive variants mitigate this issue, they often do so at the expense of increased design complexity and parameter tuning sensitivity. The observer gains must be carefully tuned to balance robustness and responsiveness; overly aggressive gains risk amplifying measurement noise, while conservative gains reduce attack sensitivity. Furthermore, in the absence of measurement updates, such as during prolonged *denial-of-service* (DoS) conditions, the enforcement of the sliding surface becomes infeasible, as noted in [202]. Under such scenarios, estimation uncertainty inevitably grows despite the intrinsic robustness of the SMO framework.

Emerging research trends seek to hybridize SMOs with adaptive, fuzzy, or neural estimation layers to further enhance robustness and mitigate structural limitations. The adaptive-fuzzy sliding control approach in [95] demonstrates how neural approximation and fuzzy compensation can effectively suppress chattering while maintaining estimation accuracy under nonlinear attack conditions. Similarly, backstepping–super-twisting hybrid architectures [203] achieve robust estimation and control under FDI attacks, combining the finite-time convergence of sliding modes with the smoothness of backstepping control.

These advances indicate a shift toward flexible, layered SMO architectures capable of handling uncertainty, nonlinearity, and cyber interference in a unified framework.

Overall, the evolution of SMOs—from early SMIO-based fault detectors [197] to recent adaptive, distributed, and higher-order formulations [45], [48], [80], [159], [200]—demonstrates their central role in bridging estimation and resilient control. Their unique ability to provide real-time attack reconstruction and finite-time convergence makes them highly suitable for *deception* and *integrity* attack scenarios. Yet, their practical deployment in large-scale CPSs still requires advances in chattering mitigation, gain adaptation, and hybridization with data-driven or model-free paradigms to ensure scalability and robustness under complex, adversarial environments.

Unknown Input Observers (UIOs)

Unknown Input Observers (UIOs) represent one of the most theoretically rigorous frameworks for secure state estimation in the presence of unmeasured disturbances, faults, or malicious inputs. The defining feature of a UIO is its ability to decouple the influence of unknown inputs from the estimation error dynamics, enabling accurate state reconstruction even when the system is subject to adversarial perturbations. This property makes UIOs particularly suitable for the detection and isolation of *deception*-type attacks, such as false data injection (FDI), covert manipulation, and replay attacks [43], [77], [204].

Classical formulations of UIOs rely on strict structural conditions—specifically, rank and observability constraints on the system matrices—that guarantee disturbance decoupling. When these conditions are satisfied, the observer yields unbiased and asymptotically stable state estimates regardless of the unknown input signal. This structure has been widely adopted in cyber–physical systems (CPSs), where maintaining estimation integrity under compromised communication is critical. For example, the zonotopic UIO proposed in [43] achieves simultaneous attack detection, isolation, and reconstruction under FDI and actuator attacks. Similarly, the event-based decentralized UIO in [77] eliminates continuous communication requirements while preserving detection performance, demonstrating the adaptability of UIOs to distributed networked architectures.

Beyond these standard designs, recent works have expanded UIO theory to encompass adaptive, stochastic, and interval-based formulations. In the smart grid domain, [171] develops a dynamic reduced-order UIO that reduces computational complexity while enhancing robustness to measurement noise and model variations. The interval-based approaches proposed in [98] and [79] replace fixed statistical thresholds with interval residuals, allowing the detection of attacks and faults with bounded uncertainty and frequency-dependent performance guarantees. These advances enable resilient fault and attack detection in fuzzy or nonlinear CPSs, where classical linear conditions may fail. Stochastic and resilient variants, such as the SUIO-based load frequency control scheme in [162], further extend applicability to systems with significant parametric uncertainty, providing probabilistic robustness under attack.

UIOs have also proven central to distributed detection architectures, where agents must infer local and neighboring states under partial information. In [183] and [205], distributed UIO banks are deployed across interconnected subsystems to detect covert or coordinated attacks in DC microgrids and power networks. By combining local observer residuals with inter-agent information, these architectures improve detection granularity and robustness to communication loss. The event-triggered scheme proposed in [182] further optimizes communication efficiency by activating estimation updates only when predefined thresholds are exceeded, thereby conserving bandwidth while retaining timely detection capability. Similarly, the privacy-preserving distributed detection method in [78] demonstrates that UIO-based estimators can support attack detection without revealing sensitive state information, a key consideration in multi-agent systems.

A different yet conceptually aligned research stream focuses on hybrid UIO-based detection and control for complex industrial and infrastructure systems. The centralized and distributed CAFDI frameworks presented in [204] integrate UIO-based detectors

into hierarchical plant–controller architectures to simultaneously identify covert, zero-dynamics, and replay attacks, both in centralized and large-scale interconnected CPSs. Similarly, in [164], a UIO embedded in the secondary frequency control loop of an isolated microgrid detects cyberattacks and uses the reconstructed attack estimate for active compensation, ensuring continued system stability. Earlier foundational works such as [161] and [31] established UIOs as a model-based diagnostic tool for SCADA and networked control systems, where their decoupling capability enables attack isolation in complex physical–cyber interactions.

Despite their theoretical elegance, UIOs are constrained by practical limitations that hinder their deployment in large-scale, uncertain, or dynamically evolving CPSs. The rank and matching conditions required for exact decoupling are easily violated in high-dimensional or nonlinear systems, leading to biased or divergent estimates. Moreover, their reliance on accurate knowledge of system and input matrices makes them vulnerable to model uncertainty—an inherent feature of modern interconnected infrastructures. Interval and reduced-order variants [79], [98], [171] partially relax these constraints but at the cost of increased conservatism or computational burden. Furthermore, under *denial-of-service* (DoS) conditions, UIOs cannot perform correction steps, resulting in progressive divergence of the estimation error, as their structure inherently assumes at least intermittent measurement updates [69].

Recent work attempts to overcome these limitations through adaptive and intelligent designs. In [206], a dual-observer architecture integrates UIOs with intelligent classification logic to distinguish between faults and cyberattacks in DC microgrids, thereby addressing the long-standing indistinguishability problem between attack-induced and naturally occurring anomalies. Covert attack detection methods such as [172] combine adaptive observers with UIO logic to estimate attack-induced control deviations while maintaining secure operation. Although promising, these adaptive and intelligent UIOs introduce new challenges in stability proof, computational cost, and susceptibility to learning-based adversarial manipulation.

UIOs thus remain one of the most conceptually robust and mathematically grounded tools for attack detection and state estimation in CPSs. Their ability to isolate unknown or malicious inputs while preserving estimation accuracy positions them as a cornerstone of secure observer design. Yet, their dependence on precise structural conditions and model fidelity limits their resilience under realistic conditions of data loss, uncertainty, and complex nonlinear coupling. Ongoing research seeks to integrate UIOs into hybrid, event-triggered, and data-driven architectures to bridge the gap between theoretical decoupling and practical resilience in adversarial cyber–physical environments

Data-Driven Observers

Data-driven observers extend the traditional observer paradigm to scenarios in which the system model is partially known, uncertain, or entirely unavailable. Instead of relying on first-principle models, these approaches infer system dynamics and residual relationships directly from data, exploiting historical or real-time measurements to perform estimation and attack detection. This paradigm has become increasingly relevant for complex cyber–physical systems , where accurate modeling is often infeasible due to nonlinearities, network-induced uncertainties, and heterogeneous components [100], [101], [102], [207].

Early data-driven formulations focused on deriving residual generators directly from input–output data, bypassing explicit model identification. In [100], a dead-beat unknown-input observer is reconstructed using historical data alone, demonstrating that under mild persistency-of-excitation conditions, it is possible to synthesize a residual generator equivalent to a model-based UIO. This approach formally bridges data-driven estimation with classical observer theory, maintaining interpretability while relaxing modeling requirements. Similarly, [101] proposes a data-driven UIO for false data injection (FDI) detection in DC microgrids. The observer is trained using simulated and experimental data, achieving effective real-time attack localization without knowledge of physical

parameters. Such works exemplify a shift from model-driven to model-free observer synthesis, combining theoretical rigor with practical scalability.

Hybrid frameworks integrating physical reasoning and machine learning are increasingly prevalent. The hybrid physics-informed neural network (HPINN) of [208] embeds neural approximators within a physics-based structure, preserving interpretability while leveraging data-driven adaptability. This architecture enables real-time detection and mitigation of sensor attacks in DC–DC converters, demonstrating strong resilience against both constant and time-varying FDI attacks. A similar philosophy is reflected in the dual security control framework in [209], where mechanistic analysis and data-driven prediction are fused under an adaptive event-triggered communication scheme for industrial CPS. Here, machine learning modules based on extreme learning machines predict control reconfigurations under DoS conditions, while robust observers maintain analytical guarantees, embodying an active–passive defense mechanism. This kind of hybridization represents a significant step toward interpretable and resource-efficient data-driven security architectures.

In the broader context of cyber–physical coordination, data-driven observers play a key role in predictive and hierarchical detection frameworks. For example, [102] introduces a collaborative cyber–physical detection architecture for dynamic load-altering attacks in smart grids. The framework combines a cyber-layer prediction model—based on radial basis function (RBF) neural networks—with a physical-layer UIO for localization, achieving enhanced responsiveness and cross-domain resilience. Likewise, [210] extends the concept to heterogeneous nonlinear multi-agent systems, integrating distributed observer-based attack isolation with data-driven resilient control to preserve synchronization under FDI attacks. These studies reveal the growing role of data-driven approaches as coordinators across layers and agents in multi-scale CPS security.

Learning-based observer structures also emerge in vehicle and energy domains. The neural–observer hybrid for autonomous electric vehicles proposed in [211] combines a conventional state observer with a neural network residual classifier, enabling the detection of integrity, replay, and DoS attacks through data consistency analysis. Similarly, [207] presents an analytical, parameter-free detection approach for grid-connected converters, capable of distinguishing cyber-induced perturbations from physical transients without model dependence or prior parameter tuning. These contributions demonstrate that even in high-dimensional or fast-dynamic systems, data-driven observers can provide lightweight, adaptive alternatives to classical filters.

Another growing research stream leverages unsupervised learning and clustering for attack detection in smart infrastructures. [212] proposes an observer-meter framework for electricity theft detection in advanced metering infrastructures (AMI), using wavelet-based feature extraction and fuzzy c-means clustering to differentiate between normal and malicious consumption patterns. The unsupervised nature of this method avoids dependence on labeled datasets, improving adaptability in real-world utilities where attack data is scarce or incomplete.

While these approaches offer unprecedented flexibility, they introduce new forms of vulnerability. Their performance is inherently tied to the representativeness and integrity of training data—exposing them to data poisoning, concept drift, and adversarial manipulation. Moreover, the black-box character of many neural or regression-based models reduces interpretability, complicating certification for safety-critical applications. Several studies attempt to address these issues through event-triggered or model-free adaptive control frameworks. For instance, [213] proposes a model-free adaptive control (MFAC) structure that uses only real-time input–output data for predictive compensation under FDI and DoS attacks. This design eliminates dependence on explicit models and demonstrates that data-driven feedback can maintain operational stability even under communication constraints. Similarly, [214] employs long short-term memory (LSTM) neural networks within an adaptive discrete event-triggered communication scheme to reconstruct lost data under high-energy DoS attacks, achieving hybrid active–passive intrusion tolerance in industrial CPS.

Despite these advances, fundamental challenges remain unresolved. The absence of

formal guarantees on stability and robustness under adversarial perturbations limits the trustworthiness of data-driven observers. Hybrid designs that embed physical priors into learning architectures (e.g., [208], [209]) partially mitigate this issue but still lack rigorous convergence analysis. Furthermore, data-driven methods often rely on extensive computational resources and high-quality datasets, constraining their real-time applicability in embedded or distributed CPS.

Data-driven observers thus embody a paradigm shift—from model-based to information-based estimation—allowing flexible detection in uncertain and dynamically evolving environments. They excel in general *detection* tasks, particularly under model uncertainty and partial observability, but remain less reliable for sustained *deception* or *DoS* scenarios where data continuity and integrity are compromised. Continued progress in hybrid learning, interpretability, and resilient adaptation will be essential to elevate data-driven observers from empirical tools to certifiable components of secure CPS architectures.

Hybrid and Adaptive Observers

Hybrid and adaptive observers represent the current frontier of resilient state estimation in cyber–physical systems (CPSs), merging analytical models with adaptive mechanisms and data-driven inference to ensure robustness under uncertainty and adversarial interference. These observers dynamically adjust their parameters or structural configurations in response to time-varying conditions such as *DoS*, *false data injection* (FDI), or intermittent communication, providing an adaptable alternative to fixed-parameter designs.

Lifted observer frameworks [72] extend the state vector to encapsulate latent attack dynamics, enabling explicit reconstruction of injected or hidden signals through augmented system representations. Intermediate and adaptive observers [15], [158], [172] exploit adaptive laws or variable gains to mitigate modeling errors and to maintain estimation convergence despite stochastic or intermittent measurement losses. Such schemes have demonstrated efficacy in smart grids and industrial CPSs, where dynamic operating conditions and communication irregularities demand flexible observer structures. Reduced-order and leader-based observers [103], [171] further improve computational tractability in large-scale networks, reducing communication overhead without compromising estimation accuracy.

Several works emphasize hybrid architectures that combine adaptive estimation with robust control. For instance, the adaptive integral sliding mode observer proposed in [199] achieves optimal L_2-gain performance under FDI attacks by continuously tuning its gain matrix according to Lyapunov-based adaptation laws. Similarly, the resilient control scheme of [49] leverages an adaptive sliding mode observer to detect and compensate sensor and actuator attacks in power networks. These adaptive mechanisms improve attack compensation speed and robustness but require careful tuning to avoid instability caused by overreactive gain adaptation.

Predictive and proactive hybrid observers integrate estimation with early warning mechanisms. The predictive attack estimation method in [58] employs model-based anticipation of sensor manipulation, while sampled-data estimation frameworks [202] use hybrid discrete–continuous observers to maintain stability under intermittent DoS conditions. Fault-tolerant observers such as those in [56] couple structural redundancy analysis with adaptive residual evaluation, achieving real-time attack isolation even under partial data availability. Cross-domain predictive control schemes [84] and lightweight distributed observers [32] further demonstrate that resilience can be enhanced through multi-layer hybridization—integrating prediction, diagnosis, and control across physical and cyber domains.

Recent advances in hybrid distributed estimation have also explored predefined-time convergence and event-triggered communication. The cooperative framework in [200] combines predefined-time switched observers with consensus algorithms to detect local and communication-layer attacks in mobile robot networks. Similarly, event-triggered distributed detection methods [182] minimize communication load while preserving

detectability in multi-agent systems. These architectures highlight the growing trend toward resource-aware adaptive observers capable of balancing performance and network efficiency.

In smart grids, adaptive observers have been employed to enhance attack detection precision. The robust adaptive observer in [215] detects and localizes bias load injection attacks (BLIA) by dynamically adjusting its thresholds to account for model linearization errors and exogenous disturbances. Likewise, reduced-order asynchronous control methods [201] combine adaptive and sliding-mode logic to stabilize nonlinear networked systems under simultaneous FDI and switching attacks. These contributions illustrate the evolution of adaptive observers toward scalable, uncertainty-tolerant architectures for real-world CPS applications.

Hybrid strategies also extend beyond deterministic designs, incorporating stochastic, interval, and data-driven components. The frequency-informed interval UIO in [79] demonstrates that hybridizing interval estimation with adaptive tuning improves both fault sensitivity and noise tolerance in fuzzy CPSs. Stochastic formulations based on Markovian switching [99] and multi-modal learning augment classical deterministic observers with probabilistic resilience. Moreover, hybrid detection frameworks [216] integrate adaptive thresholds and logic-based discrimination to differentiate between faults and replay attacks in hybrid dynamic systems.

While hybrid and adaptive observers enhance flexibility and robustness, they also introduce new challenges. Adaptive laws can themselves destabilize the estimation loop if improperly designed, especially under abrupt attack transitions or measurement dropouts. The hybridization of multiple estimation and control layers increases the complexity of stability proofs, tuning, and real-time implementation. Furthermore, the computational overhead associated with adaptation and prediction mechanisms can restrict their deployment in embedded or large-scale CPSs.

Overall, hybrid and adaptive observers embody the synthesis of resilience and intelligence in modern estimation theory. By fusing robust control, adaptive learning, and predictive reasoning, they advance the frontier of secure estimation in adversarial environments. Nevertheless, their practical adoption demands careful trade-offs among accuracy, complexity, and scalability, motivating ongoing research into lightweight, verifiable, and distributed adaptive estimation architectures.

Discussion

The analysis of observer-based detection strategies reveals a persistent trade-off between analytical tractability, robustness to adversarial interference, and adaptability to uncertain or evolving system conditions. No single observer architecture offers universal superiority; rather, each embodies a distinct balance between mathematical rigor, computational feasibility, and resilience under specific classes of attacks.

Statistical observers, such as Kalman-based designs, remain unmatched in terms of computational efficiency and optimality when their modeling assumptions hold. They are particularly effective for *detection*-type attacks manifesting as measurable deviations in sensor data, offering fast and analytically grounded attack identification [43], [57], [100], [101]. However, their reliance on Gaussian noise and accurate model structures makes them fragile under *deception* and prolonged *DoS* scenarios. In environments with limited uncertainty and high-quality models, they represent the most practical solution, but their robustness rapidly deteriorates as adversarial behavior departs from stochastic assumptions.

Set-based observers occupy the opposite end of the spectrum. Their deterministic formulation ensures formal guarantees on estimation bounds and consistency, making them exceptionally strong against *deception*-type attacks designed to mimic noise statistics [36], [37], [68], [94]. They outperform statistical filters in safety-critical contexts where guarantees are required, such as power grids and autonomous systems, but this robustness is achieved at the cost of computational scalability. Their performance advantage is thus context-dependent: they dominate under bounded-uncertainty assumptions but become

impractical for high-dimensional, fast-evolving systems or under extended data loss, where set over-approximation grows uncontrollably [43], [74].

Sliding Mode Observers (SMOs) offer the best compromise between robustness and responsiveness. Their finite-time convergence and ability to reconstruct attack signals make them superior for rapid *deception* and *actuator attack* detection [41], [97], [155], [159]. In nonlinear or high-disturbance environments, they outperform both statistical and set-based observers in terms of speed and robustness [99], [197]. Nonetheless, their practical utility is constrained by the *chattering* phenomenon and by the need for continuous measurement feedback. SMOs are therefore most effective in systems with high sampling rates and robust actuators, but less suitable under intermittent measurements or DoS conditions.

Unknown Input Observers (UIOs) provide a theoretically elegant solution for isolating and decoupling unknown or malicious inputs. When the required rank and observability conditions are satisfied, they achieve precise estimation and attack reconstruction that neither Kalman nor sliding-mode structures can guarantee [43], [77], [164]. This makes them highly advantageous for structured systems—such as microgrids or industrial controllers—where accurate models are available. However, in heterogeneous or partially known systems, these structural requirements are rarely met, and UIOs lose their theoretical robustness [79], [98], [171]. Their superiority thus lies in model-driven, well-structured environments, not in uncertain or data-limited ones.

Data-driven observers invert the traditional paradigm by inferring dynamic relationships directly from measurements. Their main strength is adaptability: they can be trained for highly nonlinear or uncertain systems where model-based observers fail [100], [101], [207], [208], [209]. They excel in broad *detection* scenarios—particularly when diverse operational data is available—but remain vulnerable to *data poisoning*, bias, and a lack of interpretability. In large-scale CPS with rich sensing and communication infrastructures, data-driven observers outperform model-based ones in adaptability and deployment speed, yet they fall short in providing formal guarantees or bounded resilience under adversarial manipulation [102], [211], [212], [214].

Hybrid and adaptive observers attempt to combine the best features of these paradigms—analytical structure, robustness, and learning adaptability—within unified frameworks. Their adaptive gain tuning, predictive logic, and cross-domain integration enable resilience under mixed *deception–DoS* conditions, outperforming fixed-parameter observers in dynamic and partially observable systems [72], [103], [158], [172]. However, they introduce new vulnerabilities related to design complexity, stability assurance, and computational overhead. When properly tuned, hybrid observers represent the most versatile option for real-world CPS, achieving balanced performance across attack classes, but their practical success hinges on careful adaptation law design and verifiable stability proofs [32], [56], [84].

Ultimately, the choice of observer depends on the dominant threat and operational constraints. Statistical and UIO-based observers remain the most efficient and theoretically sound under structured, model-rich conditions; set-based and sliding-mode approaches provide superior robustness for bounded and nonlinear uncertainties; while data-driven and hybrid observers dominate in high-complexity, data-rich environments where modeling fidelity cannot be guaranteed. Crucially, the detection of a *denial-of-service* attack is inherently trivial—the absence of packets directly reveals the disruption—but ensuring *secure estimation and control continuity* during such events remains a fundamental challenge.

The future direction of research should therefore shift from simple attack *detection* to *attack-tolerant estimation and control*, emphasizing resilience, boundedness, and self-adaptation under adversarial uncertainty. The integration of model-based guarantees with data-driven adaptability—through verifiable hybrid architectures—appears to be the most promising pathway toward truly secure and resilient CPS operation [58], [202], [216].

2.2 Thesis Contributions

The review presented in the previous section highlights several open challenges in the field of Observer-Based Attack Detection for Cyber–Physical Systems . Most existing approaches concentrate on the task of *detection*, that is, recognizing deviations from the system's nominal behavior. While essential, detection alone offers limited value in practice, as it does not provide the deeper insight required for effective response—namely, identifying which component is compromised, reconstructing the malicious input, and isolating the affected subsystem. Consequently, there is a need for frameworks that move beyond attack flagging and enable real-time *reconstruction* and *isolation* of attack signals.

A second limitation concerns the architecture of existing detection systems. Centralized designs remain predominant because of their analytical convenience, yet they suffer from scalability bottlenecks and create single points of failure. Decentralized and distributed schemes alleviate these issues, but often at the expense of accuracy or by assuming overly homogeneous subsystems and ideal communication conditions. This limits their applicability to the heterogeneous and large-scale nature of real industrial CPSs.

Furthermore, the inherent structure of large dynamical systems is often underexploited. In spatially distributed processes, the influence between distant state variables decays rapidly, creating a form of *implicit sparsity* in the system matrices. However, this structural property has rarely been quantified or used systematically to improve the efficiency of estimation and detection. Finally, most distributed methods neglect subsystem heterogeneity, an important aspect in practical infrastructures where local dynamics and parameters differ significantly.

These limitations motivate the research direction of this thesis, which advances OBAD from simple attack detection to a *scalable, structured, and reconstructive* framework for resilient CPS security. The thesis develops three main contributions addressing these challenges.

1. Quantification of Implicit Sparsity in Large-Scale Kalman Filters (Chapter 3). We first provide a theoretical characterization of the *implicit sparsity* that naturally arises in discretized dynamical systems. Although the system, gain, and covariance matrices of Kalman filters appear dense, their entries decay exponentially with spatial separation between states. We introduce formal metrics to quantify this decay and provide analytical tools to predict the effective coupling range. This result establishes the theoretical foundation for exploiting sparsity in large-scale estimation and detection problems.

2. Sparse and Scalable Attack Detection Schemes (Chapter 4). Building on this insight, we design *sparsity-aware* detection architectures that strategically truncate negligible matrix entries to localize estimation and communication tasks. We analyze the resulting bias–complexity trade-off and derive explicit bounds on estimation errors introduced by truncation. This leads to lightweight, distributed Kalman-based estimators that preserve near-optimal accuracy while achieving substantial gains in scalability and computational efficiency.

3. Distributed Reconstruction and Isolation of Sensor Attacks (Chapter 5). Finally, we propose a distributed framework based on *Sliding Mode Observers* for the real-time reconstruction and isolation of malicious sensor signals. The approach exploits the robustness of SMOs to estimate the injected attack $\delta^i(t)$ and provides analytical guarantees on the reconstruction error. A cooperative inter-node strategy combines local alarms to identify compromised subsystems and suppress false detections. Unlike many existing distributed methods, the proposed scheme accommodates heterogeneous subsystem dynamics, making it suitable for complex industrial settings.

Chapter 3

Exploiting sparsity for efficient Kalman filtering

This chapter introduces a theoretical and algorithmic framework to explain why large–scale observer-based estimators, such as the Kalman filter, remain effectively *local* despite the apparent density of their matrices after discretization. While many physical systems—power grids, industrial plants, transportation networks, or environmental monitoring infrastructures—are inherently spatial and networked, their continuous–time dynamics are typically characterized by local couplings: each subsystem interacts primarily with its immediate neighbors, and only weakly with distant ones. However, when such systems are discretized and embedded in recursive estimators, the resulting transition, covariance, and gain matrices appear dense, suggesting that centralized processing is necessary.

This chapter challenges that assumption by introducing and formalizing the concept of *implicit sparsity*. Although discrete–time matrices may appear fully populated, the numerical influence between distant states decays rapidly along the physical or topological distance of the underlying interaction graph. We show that long–range correlations exist but are systematically small, and that this decay can be precisely quantified and exploited to achieve scalable estimation.

From this observation, two complementary contributions emerge: (i) We develop a rigorous *Spatially Exponential Decay* (SED) formalism to describe how the decay of local interactions propagates through the algebraic operations defining the Kalman filter—matrix sums, products, and inversions. We prove that key steady–state quantities, such as the predicted covariance and gain–measurement products, inherit exponential decay as a function of graph distance. The decay rate is shown to depend on a compact set of parameters encoding sampling rate, process dynamics, and noise intensities. (ii) Building on this theoretical foundation, we design distance–based truncation schemes that prune negligible entries in the covariance and gain updates. We derive analytical bounds on the bias and stability introduced by such approximations, and provide interpretable performance indicators linking the truncation radius to estimation accuracy and statistical consistency. These insights enable the development of lightweight, distributed estimators that approximate the centralized Kalman filter within provable limits.

Beyond their intrinsic analytical value, these results establish the foundation for the distributed and resilient estimation methods presented in later chapters. They justify partitioning large–scale systems into weakly coupled subsystems and support interlaced multi–rate estimation schemes (such as the) and observer-based attack detection architectures that exchange only local information while preserving global coherence.

Organization of the Chapter. The remainder of this chapter is structured as follows:

- **Section 3.1** This section introduces the mathematical framework adopted for describing the considered cyber–physical estimation problem. It defines the state-space representation, key assumptions, and notation conventions that form the foundation for the subsequent theoretical developments.
- **Section 3.2** – Here, the theoretical results concerning the spatial decay properties of estimation processes are derived. The section discusses how interconnection topology and communication sparsity influence estimation performance and scalability.
- **Section 3.3.1** This section presents simulation experiments designed to validate the proposed theoretical framework. Quantitative results illustrate the convergence

behavior, accuracy, and robustness of the estimation scheme under different network configurations.

- **Section 3.4** Building on the previous analysis, this section discusses practical considerations for designing distributed and scalable Kalman filters. It highlights trade-offs between computational complexity, communication constraints, and estimation quality.
- **Section 13.5 – Concluding Remarks.** The final section summarizes the key contributions and insights of the chapter. It also outlines how the presented results connect to the broader research objectives of this thesis and motivate the developments in the following chapters.

3.1 System Model

This section summarizes the dynamical model and estimation framework that form the basis for the analysis presented in this chapter. The complete formulation of the continuous– and discrete–time systems, along with the centralized Kalman filter and its steady–state characterization, is provided in Chapter 1.2.5. Here, we recall only the essential definitions and introduce additional elements needed for the development of state observers and their role in observer-based attack detection.

3.1.1 Continuous- and Discrete-Time Models

We consider a linear time–invariant stochastic system consistent with the formalism defined in Section 1.2. In continuous time, the system evolves according to

$$\dot{\boldsymbol{x}}_t = L\,\boldsymbol{x}_t + \tilde{\boldsymbol{w}}_t, \qquad \boldsymbol{y}_t = H\,\boldsymbol{x}_t + \boldsymbol{v}_t, \tag{3.1}$$

where $\boldsymbol{x}_t$ is the system state, $\boldsymbol{y}_t$ the measurement vector, and $\tilde{\boldsymbol{w}}_t$, $\boldsymbol{v}_t$ are zero-mean Gaussian noise processes with covariances Q_c and R_c, respectively. Spatially distributed systems typically exhibit sparse matrices L and H, reflecting local couplings among neighboring subsystems.

Discretization with sampling period Δt yields the discrete-time system

$$\boldsymbol{x}_{k+1} = F\,\boldsymbol{x}_k + \boldsymbol{w}_k, \qquad \boldsymbol{y}_k = H\,\boldsymbol{x}_k + \boldsymbol{v}_k, \tag{3.2}$$

where $F = e^{\Delta t L}$ and $\boldsymbol{w}_k \sim \mathcal{N}(0, Q)$, $\boldsymbol{v}_k \sim \mathcal{N}(0, R_c)$ are independent Gaussian processes derived from Q_c and R_c. This discrete-time model serves as the foundation for all observer and filter designs discussed below.

3.1.2 Centralized Kalman Filtering

As introduced in Section 1.2.5, the Kalman filter provides the optimal linear estimator for discrete-time stochastic systems affected by Gaussian noise. For the system model defined in Section 3.1, it computes the minimum mean-square error estimate of the state by recursively updating the predicted state and its covariance through the classical *prediction–correction* cycle.

The recursive update can be compactly expressed using the matrix mappings $f_k(\cdot)$ and $f_p(\cdot)$:

$$\begin{aligned} f_k(X) &= XH^\top (HXH^\top + R_c)^{-1} H, \\ f_p(X) &= Q + F(I - f_k(X))XF^\top, \end{aligned} \tag{3.3}$$

so that

$$K_k H = f_k(P_{k+1|k}), \qquad P_{k+1|k} = f_p(P_{k|k-1}).$$

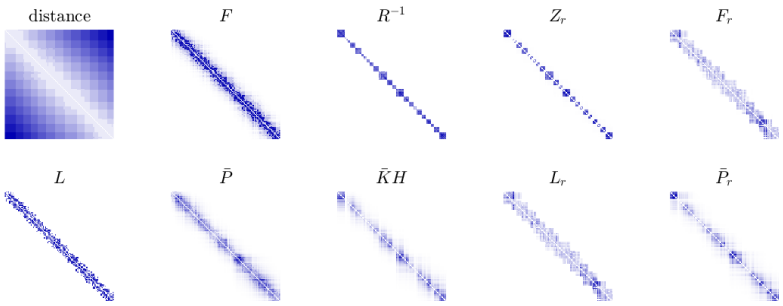


Figure 3.1: Illustration of the decaying structure in key constituent matrices of the Kalman filter. Darker colors indicate larger magnitudes, highlighting the rapid spatial decay of off-diagonal entries.

An equivalent formulation, obtained by introducing the matrix $R^{-1} = H^\top R_c^{-1} H$, is given by

$$\begin{aligned} f_k(X) &= X(I + R^{-1}X)^{-1}R^{-1}, \\ f_p(X) &= Q + FX(I + R^{-1}X)^{-1}F^\top, \end{aligned} \tag{3.4}$$

which highlights the structural dependencies among F, H, and R_c. These mappings characterize the filter's recursive dynamics in a compact algebraic form, enabling the analysis of stability, convergence, and sparsity propagation through the update steps.

In steady state, the covariance matrix converges to a fixed point $\bar{P}$ satisfying the discrete algebraic Riccati equation (DARE):

$$\bar{P} = f_p(\bar{P}), \qquad \bar{K}H = f_k(\bar{P}). \tag{3.5}$$

The corresponding steady-state Kalman gain $\bar{K}$ defines a stationary estimator that minimizes the asymptotic estimation error variance, provided that the system satisfies standard observability conditions. In particular, the Kalman filter yields the optimal state estimate—minimizing the mean quadratic error—by recursively updating predictions based on incoming measurements. To guarantee that the system state can be uniquely reconstructed from the available outputs, the following assumption is introduced.

Assumption 3.1.1

The pair (F, H) is observable. □

Assumption 3.1.1 ensures that the system's internal state is fully identifiable from its measurements, thereby enabling reliable operation of the Kalman estimator in steady-state conditions.

When the measurement matrix H is full rank (i.e., $s = n$ and R invertible), the DARE admits an equivalent, compact representation in transformed coordinates (F_r, Z_r) as discussed in Section 1.2.5. This reformulation simplifies spectral analysis and convergence proofs, providing analytical tools for later developments on distributed and structure-preserving estimation.

It is worth noting that centralized Kalman filtering requires access to all measurements $\boldsymbol{y}_k$ and the computation of global covariance updates, which becomes challenging as the system dimension increases. In large-scale or spatially distributed settings, the matrices F, H, and P_k quickly grow dense due to recursive coupling across distant states, leading to substantial communication and computational overhead. These scalability limitations motivate the exploration of sparsity-preserving and distributed estimation strategies—topics that are developed in the following sections.

3.1.3 Steady state behavior

In time-invariant systems, the state estimation error covariance matrix $P_{k+1|k}$ converges to a steady-state value $\bar{P} = \lim_{k\to\infty} P_{k+1|k}$. The convergence rate is determined by the largest eigenvalue of F. This $\bar{P}$ satisfies the discrete algebraic Riccati equation (DARE)

$$\bar{P} = f_p(\bar{P}) \,, \tag{3.6}$$

and defines the steady-state Kalman gain $\bar{K}$ via

$$\bar{K}H = f_k(\bar{P}) \,. \tag{3.7}$$

The predicted state is then updated using $\bar{K}H$

$$\hat{\boldsymbol{x}}_k = (I - \bar{K}H)F\hat{\boldsymbol{x}}_{k-1} + \bar{K}H \cdot (H^\dagger \boldsymbol{y}_k) \,. \tag{3.8}$$

The DARE can also be expressed as a quadratic matrix equality

$$\begin{bmatrix} -\bar{P} & I \end{bmatrix} \underbrace{\begin{bmatrix} F^{-\top} & F^{-\top}R^{-1} \\ QF^{-\top} & QF^{-\top}R^{-1} + F \end{bmatrix}}_{M} \begin{bmatrix} I \\ \bar{P} \end{bmatrix} = 0 \,, \tag{3.9}$$

where M is a symplectic matrix. The steady-state solution can be found iteratively through $P_{k+1|k} = f_p(P_{k|k-1})$, with $0 \preceq P_{k|k-1} \preceq P_{k+1|k} \preceq \bar{P}$ when starting from $P_{0|-1} = 0$.

3.1.4 Steady-state under the invertibility of R

When R is invertible (i.e., $s = n$ and H has rank n), the covariance matrix can be mapped to $\bar{P}_r = g_r(\bar{P})$, where $g_r(X) = I + AXA^\top$ and $A = R_c^{-\frac{1}{2}}H$. This $\bar{P}_r$ satisfies the DARE

$$\bar{P}_r = f_r(\bar{P}_r) \,, \tag{3.10}$$

where

$$\begin{aligned} f_r(X) &= Z_r + F_r(I - X^{-1})F_r^\top \\ F_r &= AFA^{-1} \\ Z_r &= g_r(Q) = I + AQA^\top = I + G_r B_r G_r^\top \\ G_r &= AGA^{-1} \\ B_r &= R_c^{-\frac{1}{2}} H Q_c H^\top R_c^{-\frac{1}{2}} \,. \end{aligned} \tag{3.11}$$

The equivalent to equation (3.9) is

$$\begin{bmatrix} -\bar{P}_r & I \end{bmatrix} \underbrace{\begin{bmatrix} 0 & F_r^{-\top} \\ -F_r & Z_r F_r^{-\top} + F_r \end{bmatrix}}_{M_r} \begin{bmatrix} I \\ \bar{P}_r \end{bmatrix} = 0 \,, \tag{3.12}$$

with M_r still symplectic. A semi-definite ordered recursion holds for $Z_{k+1} = f_r(Z_k)$, where $Z_k = g_r(P_{k|k-1})$. Starting from $Z_0 = g_r(0) = I$, we get $Z_1 = Z_r$, and $I \preceq Z_{k-1} \preceq Z_k \preceq \bar{P}_r$. This simplified form for R invertible is valuable for further investigation. For $s < n$, a compact and useful result is harder to obtain.

3.2 Decaying Properties

3.2.1 Generalities

This chapter focuses on *spatially distributed systems*, where state coupling decays with distance and the state transition matrix L is sparse. While discretization in equation (3.2) makes the discrete-time state transition matrix F formally dense, its entries, and

those of other key matrices like $\bar{P}$, exhibit rapid decay with distance. This is illustrated in Fig. 3.1, where nodes are linearly arranged with coupling limited to nearby nodes for visual clarity. Given the sparse nature of these systems, sensing is also assumed to be spatially distributed and localized, meaning H, Q, and R^{-1} also possess sparse and localized characteristics, as shown in Fig. 3.1. This phenomenon, previously explored in [217] as a spatially exponential decaying (SED) property for constituent matrices, is here generalized and strengthened.

Following [217], we construct an undirected network among the n states of $\boldsymbol{x}_k$, serving as nodes. Links are identified using the equivalent DARE equation (3.11).

Definition 3.2.1

Two nodes i and j are considered neighbors if at least one of $[L_r]_{ij}$, $[L_r]_{ji}$, or $[B_r]_{ji}$ is non-zero, where $L_r = ALA^{-1}$. This defines an adjacency matrix N for the network

$$[N]_{i,j} = \begin{cases} 1 & \text{if } [L_r]_{ij},\ [L_r]_{ji},\ \text{or } [B_r]_{ji} \text{ is active} \\ 0 & \text{otherwise}. \end{cases} \tag{3.13}$$

□

The choice of L_r is due to its role in expressing F_r and G_r via Taylor expansion, while B_r enables full tracking of matrix properties. A^{-1} also needs to be distance-one constrained for easy recovery of $\bar{P}$ from $\bar{P}_r$. The localized nature of the system and sensing ensures N is localized and sparse. This defines a symmetric distance measure $d(i,j) = d(j,i)$ between nodes, with d denoting the maximum network distance (diameter), as shown in the first block of Fig. 3.1.

3.2.2 SED formalism and its decaying properties

We generalize the findings of [217] by applying a matrix structure in terms of power series, suitable for expressing $F_r = e^{\Delta t\, L_r}$ through Taylor expansion. This implies the SED decaying property. We introduce two forms for SED matrices: a strong-SED form (like exponential functions) for constituent matrices such as F_r or Z_r, and a weak-SED form (like geometric series) more suitable for $\bar{P}_r$.

Definition 3.2.2 (*strong-SED form*)

A matrix X is in strong-SED form $X = x(I + \beta\, \mathcal{S}_0(\lambda))$, with scale parameters x, β and decay parameter λ, if

$$X = xI + x\,\beta \sum_{k\geq 1} \tfrac{1}{k!} \lambda^k U^k\,, \tag{3.14}$$

where U^k is a matrix with bounded norm $\|U^k\| \leq 1$ and active entries up to distance k. If $\beta = 1$, we use $X = x\,\mathcal{S}(\lambda)$. □

Definition 3.2.3 (*weak-SED form*)

A matrix X is in weak-SED form $X = x(I + \beta\, \mathcal{W}_0(\lambda))$, if

$$X = xI + x\,\beta \sum_{k\geq 1} \lambda^k U^k\,, \tag{3.15}$$

with the same notation as strong-SED form. If $\beta = 1$, we use $X = x\,\mathcal{W}(\lambda)$. □

The strong-SED form implies the weak-SED form with the same parameters. SED properties are related to decaying matrix entries:

Theorem 3.2.1

A strong-SED or weak-SED matrix X with parameters (x, β, λ) and $\lambda < 1$ has a

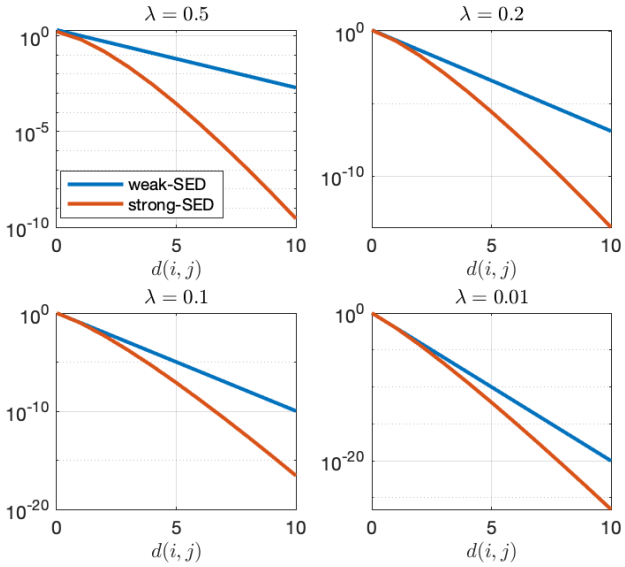


Figure 3.2: Comparison of the decaying properties of weak-SED (slower decay, in blue) and strong-SED (faster decay, in red), for selected values of λ, with $x = \beta = 1$.

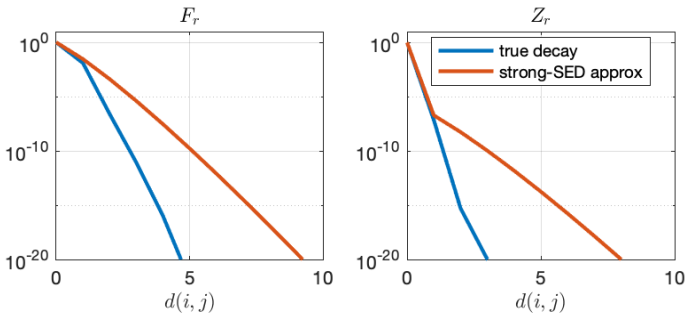


Figure 3.3: Comparison between the true decaying properties (faster decay, in blue) and the strong-SED approximation (slower decay, in red) for matrices F_r and Z_r.

decaying property

$$|X_{i,j}| \leq x\,\beta \cdot \begin{cases} \bar{\gamma}_{d(i,j)}(\lambda) & \text{strong-SED} \\ \frac{1}{1-\lambda}\,\lambda^{d(i,j)} & \text{weak-SED,} \end{cases} \tag{3.16}$$

where $d(i,j) > 0$ is the distance between i and j, and

$$\bar{\gamma}_d(\lambda) = \sum_{k \geq d} \tfrac{1}{k!}\lambda^k < \lambda^d\, e^{\lambda}\ , \tag{3.17}$$

is related to the incomplete gamma function. □

Proof. See Appendix A.1. □

Fig. 3.2 shows strong-SED matrices decay faster, especially as λ increases. Theorem 3.2.1 strengthens [217]'s decay property (which relates to the upper bound in (3.17)) and generalizes it to weak-SED matrices. The SED matrix formalism is more robust in tracking matrix properties than just the decaying formalism.

Under these definitions (see also Fig. 3.3):

Theorem 3.2.2

The system constituent matrices are in strong-SED form

$$\begin{aligned} F_r, F_r^\top &= \mathcal{S}(\lambda_f)\ , \quad \lambda_f = \Delta t\, \|L_r\| \\ Z_r &= (1+q)I + \beta_q\, \mathcal{S}_0(2\lambda_f)\ , \quad q = \tfrac{1}{2}\,\Delta t^2 \|B_r\| \\ &\qquad\qquad \beta_q = q\, \frac{2\left(\frac{e^{\lambda_f}-1}{\lambda_f}\right)^2 - 1}{e^{2\lambda_f}-1}\ . \end{aligned} \tag{3.18}$$

□

Proof. See Appendix A.2. □

We incidentally observe that q collects the ratio between the state (model) and the observation noises (see the definition of B_r in (3.11)), and, as such, it captures the overall strength of the unmodeled disturbances or stochastic perturbations affecting the system's dynamics. In the literature, it is well established that larger values of model noise covariance matrix Q_c improve the numerical stability of the Kalman filter, as they prevent the estimator from becoming overconfident in its model predictions [218]. This is because higher process noise leads the filter to rely more on measurements, enhancing robustness against modeling inaccuracies and ensuring convergence of the error covariance matrix. Consequently, increasing q typically results in improved estimation performance and avoids divergence issues in practical implementations, especially in the presence of model uncertainties and disturbances. However, since factor Δt^2 appears in the definition of q, q is expected to be a very small number. Overall, from (3.18) we get that the two key parameters identifying our system are λ_f, which identifies the decaying rate of constituent matrices, and q, or better

$$\frac{q}{\lambda_f^2} = \frac{\|B_r\|}{2\|L_r\|^2} = \frac{\|R_c^{-\frac{1}{2}} H Q_c H^\top R_c^{-\frac{1}{2}}\|}{2\|R_c^{-\frac{1}{2}} H L H^{-1} R_c^{\frac{1}{2}}\|^2}\ , \tag{3.19}$$

which summarizes the balance between state and observation noises, respectively at numerator and denominator.

3.2.3 SED properties

To fully leverage the SED formalism, we present results on SED matrix algebra. This formalism can simply capture the dynamics of sums and products, and a strong result is available for inverses. It also simplifies additive inverses, e.g., if $X = x\,\mathcal{W}(\lambda)$, then $-X = -xI + x\,\mathcal{W}_0(\lambda)$. These properties are crucial for tracking SED properties in the DARE Riccati equation (3.10).

Theorem 3.2.3

The product of two strong-SED matrices satisfies

$$\mathcal{S}(\lambda_1)\,\mathcal{S}(\lambda_2) = \mathcal{S}(\lambda_1 + \lambda_2)\ , \tag{3.20}$$

meaning decay factors simply sum. □

Proof. See Appendix A.3. □

Theorem 3.2.4

The mixed product of a strong-SED and a (truncated) weak-SED matrix satisfies

$$\mathcal{S}(\lambda_s)\,\mathcal{W}_0(\lambda_w) = \sum_{k\geq 1} \gamma_k(\tfrac{\lambda_s}{\lambda_w})\,\lambda_w^k\, U^k\ , \tag{3.21}$$

where $\gamma_k(\epsilon) = \sum_{n=0}^{k-1} \frac{1}{n!}\epsilon^n = e^\epsilon - \bar{\gamma}_k(\epsilon)$ *is related to the complementary incomplete gamma function, and* $\gamma_k(\epsilon) < \gamma_{k+1}(\epsilon) < \gamma_\infty(\epsilon) = e^\epsilon$. □

Proof. See Appendix A.4. □

Theorem 3.2.5

The inverse of a weak-SED matrix $X = I + \beta\,\mathcal{W}_0(\lambda)$*, with* $\beta > 0$ *and* $\lambda(1+\beta) < 1$*, is*

$$X^{-1} = I + \tfrac{\beta}{1+\beta}\,\mathcal{W}_0\big((1+\beta)\lambda\big)\,, \tag{3.22}$$

implying an increase in the decay parameter by $1+\beta$ *(slower decay) and a decrease in the weighting factor* $\frac{\beta}{1+\beta} < \beta$. □

Proof. See Appendix A.5. □

Theorem 3.2.5 is particularly useful for inverting the DARE Riccati equation (3.10) when X is weak-SED. The SED formalism results are more reliable than those in [217], for instance, avoiding multiplication by network size n in the product theorems and allowing for an inverse theorem. For symmetric X (like $P_{k|k-1}$ and Z_k), an even stronger result exists.

Theorem 3.2.6

The inverse of a symmetric weak-SED matrix $X = X^\top = I + \beta\,\mathcal{W}_0(\lambda)$*, with* $\beta > 0$ *and* $\lambda(1+\beta) < 1$*, is*

$$X^{-1} = \kappa_i(\beta,\lambda)\,I + \kappa_w(\beta,\lambda)\,\mathcal{W}_0\big((1+\beta)\lambda\big)\,, \tag{3.23}$$

where

$$\begin{aligned} \kappa_i(\beta,\lambda) &= 1 + \tfrac{1}{2}\,\tfrac{\beta^2\lambda^2}{1-(1+\beta)^2\lambda^2} \\ \kappa_w(\beta,\lambda) &= \tfrac{\beta}{1+\beta}\left(1 - \tfrac{1}{2}\,\tfrac{\beta\lambda}{1+(1+\beta)\lambda}\right). \end{aligned} \tag{3.24}$$

□

Proof. See Appendix A.6. □

The eigenvalue ranges for X^{-1} when $X = I + \beta\,\mathcal{W}_0(\lambda)$ is symmetric are shown in Table 3.1. Theorem 3.2.6 improves the lower bound of Theorem 3.2.5 towards the true minimum.

Table 3.1: Eigenvalue ranges for inverse theorems obtained using a symmetric matrix $X = X^\top = I + \beta\,\mathcal{W}_0(\lambda)$, under the condition $\lambda(1+\beta) < 1$.

	$\lambda_{\min} > 0$	$\lambda_{\max}$
X	$1 - \frac{\beta\lambda}{1-\lambda}$	$1 + \frac{\beta\lambda}{1-\lambda}$
Theorem 3.2.5	$1 - \frac{\lambda\beta}{1-\lambda(1+\beta)}$	$1 + \frac{\lambda\beta}{1-\lambda(1+\beta)}$
Theorem 3.2.6	$1 - \frac{\lambda\beta(1+\lambda)}{1-\lambda^2(1+\beta)^2}$	$1 + \frac{\lambda\beta}{1-\lambda(1+\beta)}$
true X^{-1}	$1 - \frac{\lambda\beta}{1-\lambda(1-\beta)}$	$1 + \frac{\lambda\beta}{1-\lambda(1+\beta)}$

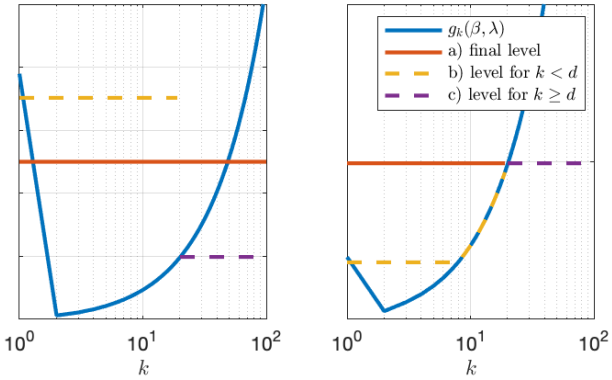


Figure 3.4: Flattening of function g_k.

3.2.4 Weak-SED behavior of $\bar{P}_r$

SED matrix algebra is a powerful tool to investigate $\bar{P}_r$'s SED properties, aiming to confirm insignificance of distant correlations. This sparsity, when mapped back to $\bar{P}$, is vital for designing distributed observers that achieve centralized performance. This analysis involves studying the SED properties of the steady-state function f_r.

Theorem 3.2.7

Let $X = x(I + \beta\,\mathcal{W}_0(\lambda))$, with $(1+\beta)\lambda < 1$ and $X = X^\top$ symmetric. The action of function f_r yields

$$X^+ = f_r(X) = x^+ I + \sum_{k\geq 1} g_k(\beta,\lambda)\,\lambda^k\,U^k\ , \tag{3.25}$$

where

$$\begin{aligned} x^+ &= 2 + q - \tfrac{1}{x}\,\kappa_i(\beta,\lambda) \\ g_k(\beta,\lambda) &= \left(\beta_q + |1 - \tfrac{1}{x}\,\kappa_i(\beta,\lambda)|\right)\,\tfrac{1}{k!}\left(\tfrac{2\lambda_f}{\lambda}\right)^k \\ &\quad + \tfrac{1}{x}\,\kappa_w(\beta,\lambda)\,(1+\beta)^k\,\gamma_k\!\left(\tfrac{2\lambda_f}{(1+\beta)\lambda}\right)\ . \end{aligned} \tag{3.26}$$

□

Proof. See Appendix A.7. □

The g_k coefficients from Theorem 3.2.7 are complex and do not directly yield a compact weak-SED form for iterative f_r application. The behavior of g_k (Fig. 3.4, blue line) typically shows a peak then a valley and a ramp (decreasing then increasing), differing from the constant behavior of a standard weak-SED form (Fig. 3.4, red line).

A constant behavior can be achieved by shifting g_k weights: for $k \geq d$, U^k matrices are active everywhere, allowing free rearrangement for a flat level (level "c" in Fig. 3.4); for $k < d$, strength can be shifted to U^ℓ with $\ell > k$ (larger matrix span), leading to level "b" in Fig. 3.4. This results in the following theorem:

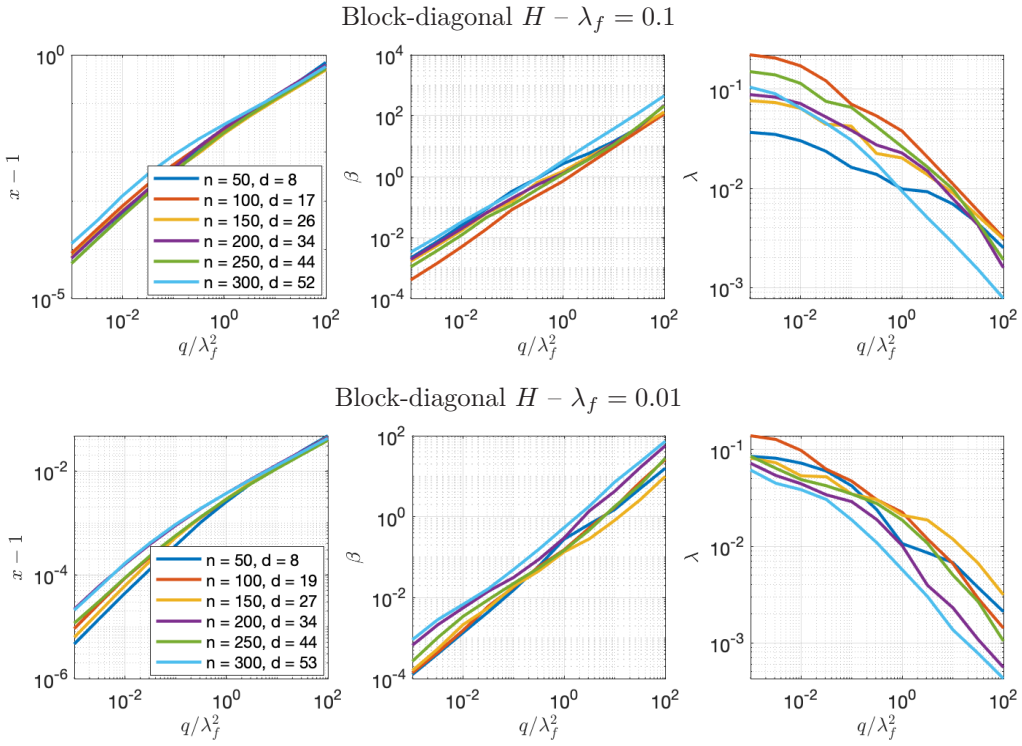


Figure 3.5: Estimated triplets (x, β, λ) obtained by fitting the covariance matrix $\bar{P}_r$ to the weak-SED form $x(I + \beta\,\mathcal{W}_0(\lambda))$, assuming a block-diagonal H. Results are shown as a function of q/λ_f^2, for $\lambda_f = 0.1$ (top) and $\lambda_f = 0.01$ (bottom).

Theorem 3.2.8 (*update*)

Consider a weak symmetric SED matrix $X = X^\top = x(I + \beta\,\mathcal{W}_0(\lambda))$ with parameters (x, β, λ) satisfying $\lambda(1+\beta) < 1$ and a maximum distance d. The action of function f_r yields a symmetric $X^+ = f_r(X) = x^+(I + \beta^+\,\mathcal{W}_0(\lambda^+))$ with

$$\begin{aligned} x^+ &= 2 + q - \tfrac{1}{x}\, k_i(\beta, \lambda) \\ \lambda^+ &= \lambda\alpha\ ,\ \alpha \leq 1 + \beta \\ \beta^+ &= \tfrac{1-\lambda^+}{x^+}\ \max(\xi_1\,\alpha^{-1}, \xi_d\,\alpha^{-d})\ , \end{aligned} \tag{3.27}$$

where

$$\begin{aligned}\xi_k &= \sum_{n=0}^{\infty} g_{k+n}\,\lambda^n \\ &= \left(\beta_q + |1 - \tfrac{1}{x}\,\kappa_i(\beta,\lambda)| + \tfrac{1}{x}\,\bar{\kappa}_w\right)\,\bar{\gamma}_k(2\lambda_f)\,\lambda^{-k} \\ &\qquad + \left(\tfrac{1}{x}\,\bar{\kappa}_w\right)\,\gamma_k\!\left(\tfrac{2\lambda_f}{(1+\beta)\lambda}\right)\,\lambda^{-1}\,(1+\beta)^{k-1} \\ \bar{\kappa}_w &= \kappa_w(\beta,\lambda)\,\tfrac{\lambda(1+\beta)}{1-\lambda(1+\beta)}\ ,\end{aligned} \tag{3.28}$$

and $\bar{\gamma}_1(2\lambda_f) = e^{2\lambda_f} - 1$, $\gamma_1(\cdot) = 1$ *at* $k = 1$. □

Proof. See Appendix A.8. □

Theorem 3.2.8 provides an upper bound useful for tracking purposes. For steady-state solutions $\bar{P}_r = f_r(\bar{P}_r)$, a similar approach applies when $(x^+, \beta^+\lambda^+) = (x, \beta, \lambda)$, considering the potential disagreement between x^+ and x.

Corollary 3.2.1 (*steady-state*)

Consider a symmetric weak-SED matrix $X = X^\top = x(I + \beta\,\mathcal{W}_0(\lambda))$ *with parameters* (x, β, λ) *satisfying* $\lambda(1+\beta) < 1$ *and maximum distance* d. *A sufficient condition for* X *to be mapped into an output matrix* $X^+ = f_r(X)$ *with the same weak-SED structure as* X *is to satisfy*

$$\beta \geq \tfrac{1-\lambda}{x}\,\max\left(\xi_1 + \tfrac{|x-x^+|}{\lambda}, \xi_d\right)\ . \tag{3.29}$$

□

To test the sufficient condition of Corollary 3.2.1, we define the excess value

$$\tau(\beta,\lambda) = \min_x \left(\tfrac{1-\lambda}{x\beta}\,\max\left(\xi_1 + \tfrac{|x-x^+|}{\lambda}, \xi_d\right) - 1\right)\ , \tag{3.30}$$

which indicates the reliability of the (β, λ) pair as a stable point. The condition is met when $\tau(\beta, \lambda) < 0$.

3.3 Simulation Results

3.3.1 Simulation scenario

To validate our findings, we conducted simulations with randomly generated system parameters, leading to the decaying properties illustrated in Fig. 3.5:

- Matrix L is generated by randomly activating links between a node and its neighbors up to a distance (in index term) of 7. Coefficients are randomly drawn to create an asymmetric, banded matrix L, as depicted in Fig. 3.1.
- Matrix H is either an identity matrix or an $n \times n$ block-diagonal matrix with random block sizes (average size 5). Its coefficients are uniformly random, and H is verified to be invertible and satisfy Assumption 3.1.1.
- For simplicity, error matrices are chosen as diagonal: $Q_c = q_c\,I$ and $R_c = r_c\,I$.
- The sampling spacing Δt is chosen to ensure a small $\lambda_f = \Delta t \|R_c^{-\frac{1}{2}} H L H^{-1} R_c^{\frac{1}{2}}\|$, typically in the range $[0.01, 0.1]$, because $2\lambda_f$ dictates the fastest possible decay of $\bar{P}_r$.

Fig. 3.5 illustrates the resulting decay characteristics of $\bar{P}_r$ for $\lambda_f = 0.1$ and $\lambda_f = 0.01$, plotted against q/λ_f^2 with a block-diagonal H and various network sizes ($n \in [50, 300]$). The estimated triplet (x, β, λ) values are derived from $\bar{P}_r$ decay plots (like Fig. 3.3), serving as useful references despite not guaranteeing an exact weak-SED form. In the

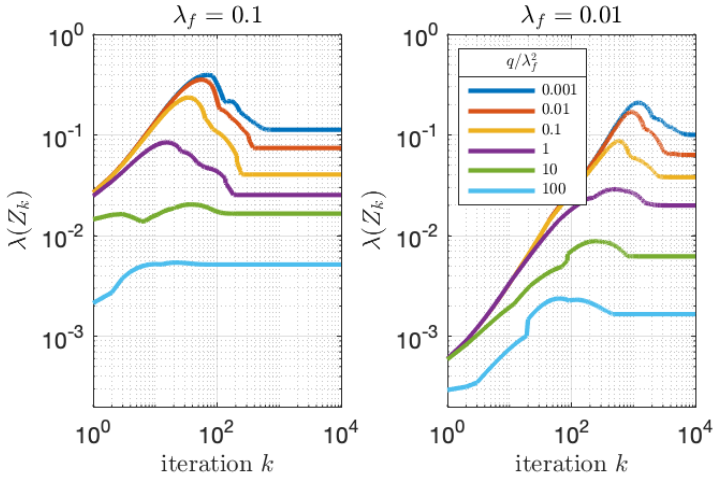


Figure 3.6: Estimated value of the weak-SED parameter λ for Z_k as the number of iterations k increases, for a network of $n = 300$ nodes with block-diagonal H. Results are shown as a function of q/λ_f^2, for $\lambda_f = 0.1$ (left) and $\lambda_f = 0.01$ (right).

block-diagonal H scenario, parameter dependence on network size n (e.g., λ values) varies with specific random realizations, but exhibits a consistent average behavior.

A general observation from Fig. 3.5 is that λ decreases as q increases. This aligns with existing literature, where higher q (proportional to the process noise covariance matrix norm and inversely proportional to the measurement covariance matrix norm, see (3.18)) is known to enhance Kalman filter robustness and improve observability, leading to faster convergence to steady-state $\bar{P}_r$ [219], [220], [221], [222]. This is consistent with a lower λ for increasing q, indicating a more compact $\bar{P}_r$ and stronger local dependence, which accelerates convergence. The weak-SED parameters of $\bar{P}_r$ are crucial for system behavior:

Theorem 3.3.1

A covariance matrix in the weak-SED form $\bar{P}_r = x(I + \beta\,\mathcal{W}(\lambda))$ implies

$$\begin{aligned} \bar{P} &= x_P\, I + x_P\, U + \alpha_P\, U \cdot \mathcal{W}_0(\lambda) \\ \bar{K}H &= x_K\, I + \alpha_K\, U \cdot \mathcal{W}_0(\lambda) + \beta_K\, U \cdot \mathcal{S}_0(2\lambda_f)\ , \end{aligned} \tag{3.31}$$

where $U \cdot \mathcal{W}_0(\lambda)$ denotes a shift-by-one in distance properties of the SED series, and constants are

$$\begin{aligned} x_P &= \tfrac{1}{2}(x-1)\,\|(A^\top A)^{-1}\| \\ \alpha_P &= x\beta\,\|A^{-1}\|^2 \\ x_K &= x - 1 - q \\ \alpha_K &= x\beta\, e^{2\lambda_f/\lambda}\,\|A^{-1}\|\,\|A\| \\ \beta_K &= |x - 1 - \beta_q|\,\|A^{-1}\|\,\|A\|\ . \end{aligned} \tag{3.32}$$

□

Proof. See Appendix A.9. □

Thus, the decay of both $\bar{P}$ and $\bar{K}H$ is related to λ for distances greater than 2, making λ a key factor in our investigation.

Fig. 3.6 illustrates the convergence properties of Z_k for an $n = 300$ node network, for various q/λ_f^2 values, with $\lambda_f = 0.1$ (left) and $\lambda_f = 0.01$ (right). The figure shows the λ value obtained by fitting Z_k in weak-SED form at each iteration. Convergence of x and β is similar but omitted for brevity. This behavior is consistent across network sizes in

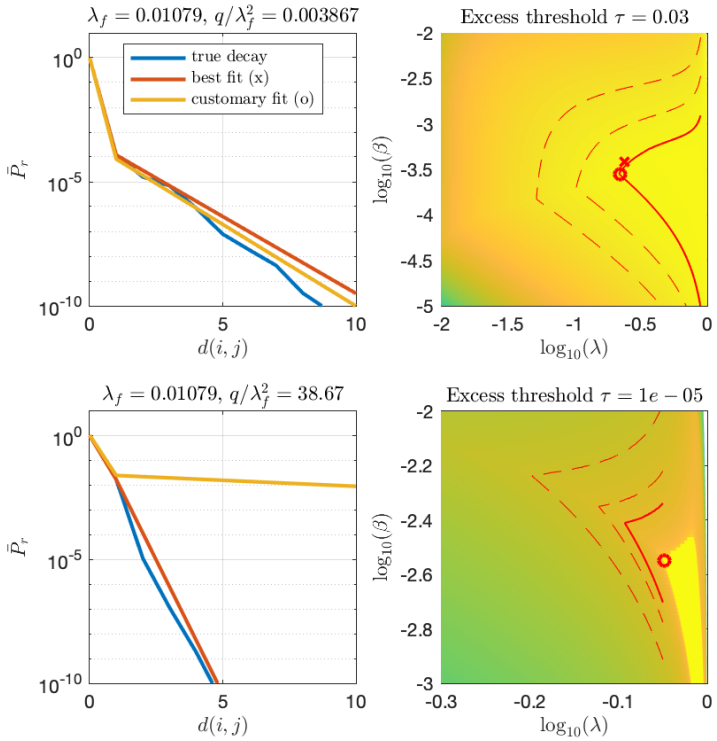


Figure 3.7: Real and estimated decay properties of $\bar{P}_r$ (left) and behavior of the excess threshold τ derived from Corollary 3.2.1 (right). Brighter colors (yellow) indicate lower values of τ, while darker colors (green) correspond to higher ones. Contour plots show the levels $\tau = 0.03$ (solid line) and $\tau = 0.05, 0.1$ (dashed lines).

the range $n \in [20, 300]$. We observe faster convergence when both λ_f and q/λ_f^2 increase, consistent with q's impact on system stability. In general, for appropriate q, convergence can take hundreds to thousands of iterations, which serves as our benchmark for future studies.

3.3.2 The search for stationary points

Fig. 3.7 shows aspects related to the steady-state description from Corollary 3.2.1 and the excess threshold τ from (3.30), for an $n = 300$ node network with maximum distance $d = 38$ and $\lambda_f \simeq 0.01$. The top plots in Fig. 3.7 use noise constants $q_c = 10^{-4} \|L\|^2$ (state) and $r_c = 10^{-3} \|H\|^2$ (observation), providing balanced noise levels with a slightly more pronounced observation noise. This yields $q/\lambda_f^2 \simeq 3 \cdot 10^{-3}$. The excess threshold (colored region on the top right of Fig. 3.7) shows lower τ values (yellow, in range $[0.02, 0.03]$) and higher τ values (green). The $\tau = 0.03$ contour (solid red line) indicates a peak (red "o" marker) close to the estimated fitting value (red "x" marker). Their fitting properties for $\bar{P}_r$ (top left of Fig. 3.7) closely match the true decay, though some overfitting is apparent since $\tau = 0.03 > 0$.

The bottom plots in Fig. 3.7 illustrate a scenario where $\tau < 0$. The bright yellow region on the bottom right represents $\tau < 0$, indicating stable points. Here, observation noise is minimized to $r_c = 10^{-7} \|H\|^2$, resulting in $q/\lambda_f^2 \simeq 39$. The reference point (red "o" marker) has $\lambda = 0.894$, very close to 1. The $\bar{P}_r$ fitting (bottom left of Fig. 3.7) closely matches the true decay at distance 1 (correctly estimated x and $\beta\lambda/(1-\lambda)$), but the λ decay estimate provides a loose bound.

Overall, the behavior in Fig. 3.7 is consistent across parameter choices: the SED formalism reliably identifies stable points only for higher q, with higher q leading to more valuable estimates, which always provide an upper bound to the true fit. This indicates

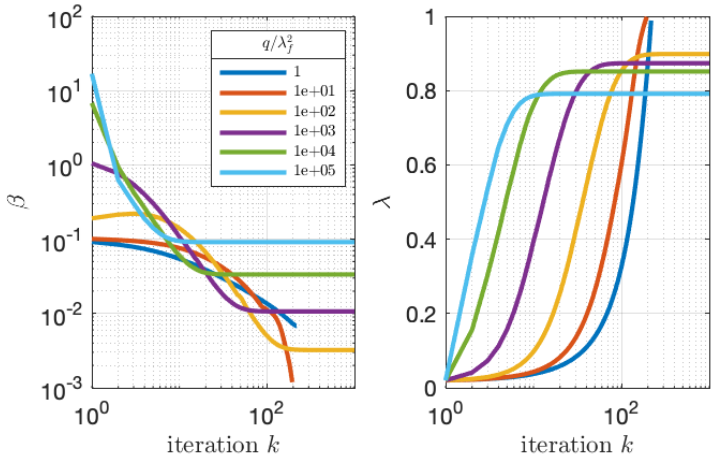


Figure 3.8: Convergence of the weak-SED estimates β (left) and λ (right) of Z_k as the number of iterations k increases. Parameters are tracked through the successive application of Theorem 3.2.8. Simulations are performed with $\lambda_f = 0.01$, maximum distance $d = 53$, and multiple values of q/λ_f^2.

some limitations of the current weak-SED formalism, but it still offers an upper bound on system performance in challenging scenarios and proves SED decay for numerous systems whose compact (λ_f, q, d) representation falls within the correct range. Further details follow.

3.3.3 Tracking properties

Fig. 3.8 and 3.9 show the tracking properties on Z_k using Theorem 3.2.8. We apply update (3.27) on $Z_k = f_r(Z_{k-1})$, starting from $Z_0 = \infty$ and

$$\begin{aligned} Z_1 &= Z_r + F_r F_r^\top \\ &= (2+q)I + (1+\beta_q)\,\mathcal{S}_0(2\lambda_f)\ , \end{aligned} \tag{3.33}$$

which is formalized in weak-SED form by retaining strong-SED form constants. The chosen α at each iteration is empirically set to minimize a target function $\lambda^+(1+\beta^+) - 0.05\,\beta/\beta^+$. The first term limits λ's increase, while the second encourages smaller β^+ values, consistent with the $1+\beta$ factor in inverse Theorems 3.2.5 and 3.2.6. We select the α that yields the minimum target within the finely sampled range $\alpha \in [0.9, 1+\beta]$, also considering $\alpha < 1$ for steady-state enforcement, consistent with Theorem 3.2.8.

Fig. 3.8 shows the resulting behavior as iterations progress, for a network with maximum distance $d = 53$ and $\lambda = 0.01$, and for a few q/λ_f^2 values. Convergence is observed only for sufficiently high q. This is primarily because at low q, the expected $\lambda(1+\beta)$ approaches 1, making Theorem 3.2.8's update less stable due to key parameters being inversely proportional to $1 - \lambda(1+\beta)$. Fig. 3.8 demonstrates the power of Theorem 3.2.8, providing meaningful output even after 10^3 successive applications.

Fig. 3.9 illustrates the steady-state estimates after 5000 iterations of Theorem 3.2.8. To ensure convergence, only high q values were selected. While the q range in Fig. 3.9 may be practically unfeasible, λ decreases with q, and β increases, consistent with Fig. 3.5. For very high q (corresponding to infinitely small observation noise), λ approaches limit values that decrease with λ_f. This suggests that finer time sampling improves tracking performance under consistent noise conditions (same q/λ_f^2). However, these limit values are very close, indicating that for practical purposes, selecting Δt such that $\lambda_f \in [0.01, 0.1]$ is a reasonable, conservative choice that also manages computational resources. Moreover, these limit values perfectly capture the higher levels in Fig. 3.5, which, though not shown for readability, saturate at smaller q/λ_f^2. This effect is mainly due to the SED formalism's specificity and the inverse theorem's improved efficiency at higher q.

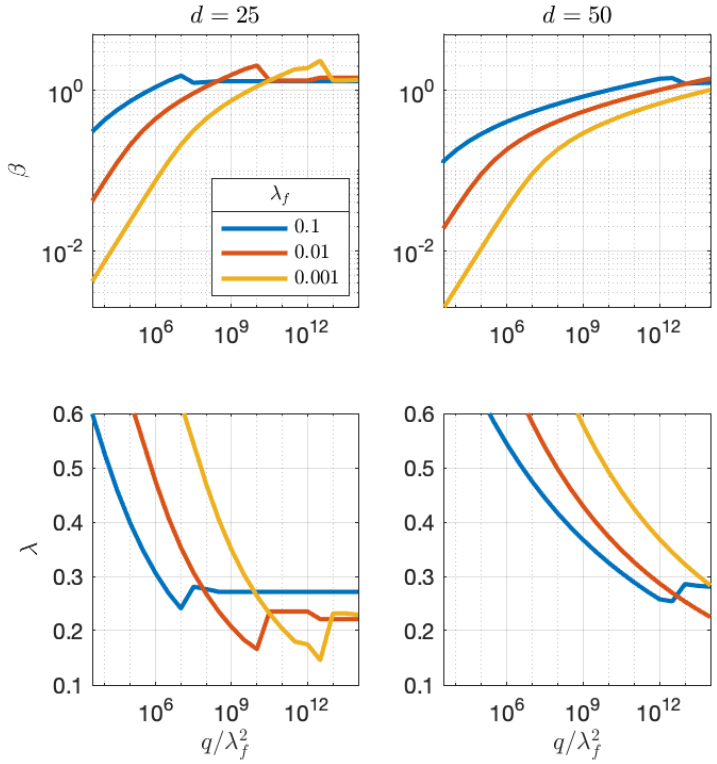


Figure 3.9: Steady-state behavior of the weak-SED parameters β (top) and λ (bottom) of $\bar{P}_r$, as a function of q/λ_f^2 for different values of λ_f. Simulations are performed on networks with maximum distance $d = 25$ (left) and $d = 50$ (right). Parameters are estimated by tracking the weak-SED properties of Z_k to the steady state through Theorem 3.2.8.

3.4 Implications on Kalman System Design

3.4.1 Distance-based pruning for spatially localized filters

The decaying structure of constituent matrices allows the Kalman filter to exploit this property by neglecting information exchange at sufficiently large distances where coefficients become negligible. This approach reduces communication and computation, enabling a fully localized Kalman process with minimal performance loss. Here, we link this performance loss to the main system parameters identified by our formalism.

We introduce a distance threshold $\kappa > 0$ and define the distance-based truncation operator

$$[\mathcal{T}_\kappa[M]]_{ij} = \begin{cases} M_{ij}, & \text{if } d(i,j) \leq \kappa \\ 0, & \text{otherwise.} \end{cases} \tag{3.34}$$

The truncated counterpart of matrix M is denoted $M^\kappa = \mathcal{T}_\kappa[M]$. In our scheme, this operator is applied directly to the state transition matrix F and the error covariance matrix P (specifically $P_{k|k}$ for prediction and $P_{k+1|k}$ for update) at each Kalman filter step. Consequently, the Kalman gain K_k is then calculated using these approximated matrices F^κ and $P^\kappa_{k+1|k}$, ensuring spatial localization throughout the filter computation. The key steps of this *pruned* Kalman filter algorithm are:

1. *Pruned Prediction Step:*

$$F^\kappa = \mathcal{T}_\kappa[F]$$
$$\hat{\boldsymbol{x}}^\kappa_{k+1|k} = F^\kappa \hat{\boldsymbol{x}}^\kappa_{k|k}$$

$$P^{\kappa}_{k+1|k} = \mathcal{T}_{\kappa}[F^{\kappa} P^{\kappa}_{k|k} (F^{\kappa})^T + Q]$$

2. *Pruned Update Step:*

$$\begin{aligned} S^{\kappa}_{k+1} &= H P^{\kappa}_{k+1|k} H^T + R \\ K^{\kappa}_{k+1} &= P^{\kappa}_{k+1|k} H^T (S^{\kappa}_{k+1})^{-1} \\ \hat{\boldsymbol{x}}^{\kappa}_{k+1|k+1} &= \hat{\boldsymbol{x}}^{\kappa}_{k+1|k} + K^{\kappa}_{k+1} (\boldsymbol{y}_{k+1} - H \hat{\boldsymbol{x}}^{\kappa}_{k+1|k}) \\ P^{\kappa}_{k+1|k+1} &= \mathcal{T}_{\kappa}[(I - K^{\kappa}_{k+1} H) P^{\kappa}_{k+1|k}] \end{aligned}$$

Here, truncation by $\mathcal{T}_{\kappa}$ means that both state estimate propagation and Kalman gain computation inherently disregard interactions beyond a distance $d(i,j) > \kappa$.

These approximations lead to a suboptimal filter, for which we can analyze the propagation of the estimation error bias. We define the *a posteriori* estimation error as $\boldsymbol{e}_k = \hat{\boldsymbol{x}}^{\kappa}_k - \boldsymbol{x}_k$, the difference between the κ-distance approximated state and the true state. Analyzing the expected bias $\mathbb{E}[\boldsymbol{e}_k]$ under this scheme, where K^{κ}_k is recursively computed from truncated matrices, requires considering error propagation at each filter step.

Theorem 3.4.1

Assume a system with limited state, $\|\mathbb{E}[\boldsymbol{x}_k]\| \leq \epsilon$, *and a Kalman filter exhibiting* stability *of the truncated dynamics matrices*

$$\begin{aligned} \|(I - K^{\kappa}_k) F^{\kappa}\| &\leq C_1 < 1 \\ \|(I - K^{\kappa}_k)(F^{\kappa} - F)\| &\leq C_2 < 1 \,, \end{aligned} \tag{3.35}$$

both valid over all iterations k. Then, the estimation error bias satisfies, in the limit,

$$\lim_{k \to \infty} \|\mathbb{E}[\boldsymbol{e}_k]\| \leq \frac{C_2 \, \epsilon}{1 - C_1} \,. \tag{3.36}$$

□

Proof. See Appendix A.10. □

This analysis leads to two main conclusions. First, to maintain a bounded bias, the distance threshold κ must be chosen large enough to ensure stability of the truncated dynamics matrix ($C_1 < 1$). Second, a smaller κ leads to sparser computation but a larger C_2, and thus a greater bias.

3.4.2 Key Performance Indicators

To assess the impact of distance-based pruning and link it to the key parameter λ from the SED formalism, we introduce several key performance indicators (KPIs). For simplicity, we omit κ from their notation.

Normalized Decay Properties The decay characteristics of the steady-state error covariance matrix $\bar{P}_r$ (e.g., in Fig. 3.7, left graphs) are a crucial link between system performance and the SED formalism. These properties are formally defined in a normalized form as

$$\Gamma_P = \frac{\max_{i,j:d(i,j)=\kappa} |[\bar{P}_r]_{i,j}|}{\max_{i,j} |[\bar{P}_r]_{i,j}|} \,. \tag{3.37}$$

This metric quantifies the relative magnitude of elements at a specific distance κ compared to the largest element, indicating spatial decay.

Normalized Matrix Perturbations Normalized perturbations quantify the distance between an original matrix and its κ-truncated counterpart. This is defined for the state transition matrix F and the Kalman gain-measurement matrix product $\bar{K}H$ (representing the steady-state or average effect of the Kalman gain on measurements) as:

$$\Gamma_F = \frac{\|F^\kappa - F\|}{\|F\|} , \tag{3.38}$$

$$\Gamma_{KH} = \frac{\|\bar{K}^\kappa H - \bar{K}H\|}{\|\bar{K}H\|} . \tag{3.39}$$

Here, $\|\cdot\|$ denotes a suitable matrix norm (e.g., the Frobenius norm). This metric measures the fractional change due to truncation, with its value decreasing as κ increases (less information is discarded).

Normalized Mean Squared Error The Normalized Mean Squared Error (NMSE) quantifies state estimation accuracy by measuring the average squared deviation between the estimated state from the pruned filter, $\hat{\boldsymbol{x}}_k^\kappa$, and the true state, $\boldsymbol{x}_k$. It is defined as

$$\text{NMSE} = \frac{\frac{1}{K}\sum_{k=1}^{K} |\hat{\boldsymbol{x}}_k^\kappa - \boldsymbol{x}_k|^2}{\frac{1}{K}\sum_{k=1}^{K} |\hat{\boldsymbol{x}}_k - \boldsymbol{x}_k|^2} , \tag{3.40}$$

where K is the total number of iterations. This ratio compares the MSE of the approximated filter's estimation to that of the optimal filter.

Normalized Innovation Squared The Normalized Innovation Squared (NIS) assesses filter statistical consistency by evaluating the innovation sequence. The innovation, $\tilde{\boldsymbol{y}}_k^\kappa = \boldsymbol{y}_k - H_k\hat{\boldsymbol{x}}_{k|k-1}^\kappa$, is the discrepancy between the actual measurement and the predicted measurement from the *pruned* filter. NIS weights the innovation by its covariance matrix $S_k^\kappa = H_k P_{k|k-1}^\kappa H_k^T + R$, also computed by the *pruned* filter (see (1.12)-(1.13)). NIS is defined as

$$\text{NIS} = \frac{1}{K}\frac{\sum_{k=1}^{K} (\tilde{\boldsymbol{y}}_k^\kappa)^\top (S_k^\kappa)^{-1} \tilde{\boldsymbol{y}}_k^\kappa}{s} , \tag{3.41}$$

where s is the measurement vector dimension. For a well-calibrated filter, $(\tilde{\boldsymbol{y}}_k^\kappa)^\top (S_k^\kappa)^{-1}\tilde{\boldsymbol{y}}_k^\kappa$ should approximately follow a chi-square distribution with s degrees of freedom, yielding an NIS value close to 1 [223].

3.4.3 Numerical results

We analysed a spatially distributed system with $n = 100$ and system matrices defined as in Section 3.3.1. In particular, for the scenario simulation in this section, the maximum distance among states is $d = 21$. We evaluate the distance-based truncation approach for truncation thresholds $\kappa \in \{1, \ldots, 10\}$. We use a fixed $\lambda_f = 0.0339$ and three different magnitudes of process noise, yielding three values of the ratio q/λ_f^2. This ratio, defined in (3.19), balances state and observation noises and is a key parameter in our SED formalism.

Fig. 3.10 validates the theoretical predictions of the SED framework. In the high-q/λ_f^2 regime (yellow curve), the steady-state covariance matrix $\bar{P}_r$ (and thus $\bar{P}$ via Theorem 3.3.1) exhibits the fastest spatial decay. Here, estimation heavily relies on measurements due to dominant process noise, and correlations between distant states vanish quickly. Consequently, even aggressive filter truncation ($\kappa = 1$) provides near-optimal performance, as shown by the immediate convergence of NMSE and NIS to 1, and the sharp decay of Γ_P and Γ_{KH}.

As q/λ_f^2 decreases (red line), process noise is less dominant, and system dynamics play a larger role. This corresponds to increased λ (Fig. 3.5), leading to slower spatial decay in $\bar{P}_r$ entries. This necessitates increasing the truncation threshold to around $\kappa = 4$ to maintain estimation performance. NMSE and NIS converge more gradually, and Γ_P and Γ_{KH} decay slower, reflecting wider spread of significant correlations.

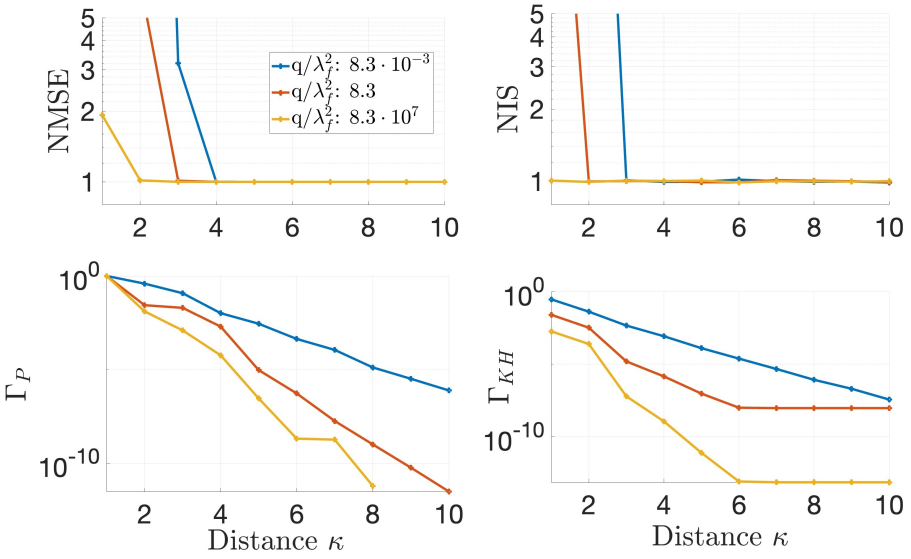


Figure 3.10: KPI comparison for distance-based truncation under varying process noise magnitude (q/λ_f^2). The number of state variables is fixed at $n = 100$.

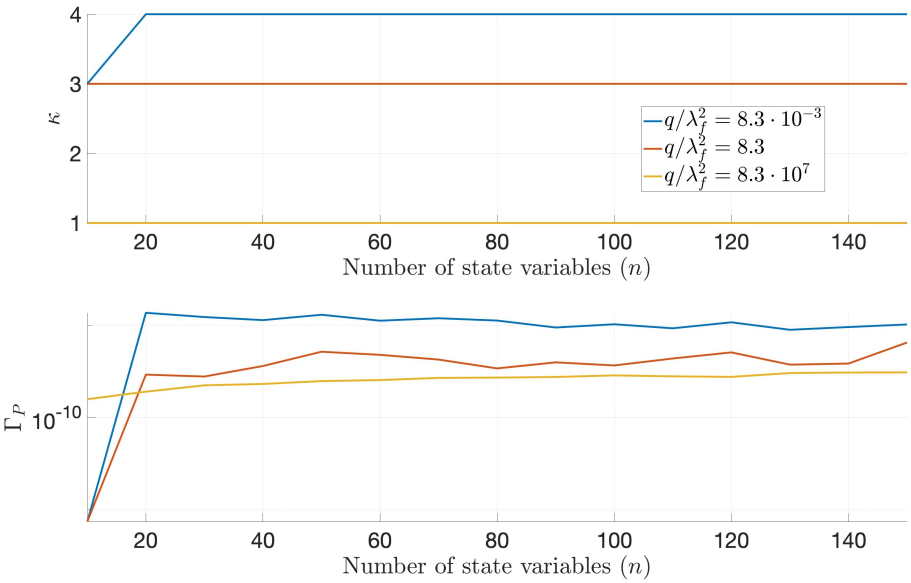


Figure 3.11: Impact of the number of nodes on the minimum required κ and on the normalized covariance perturbation Γ_P for near-optimal performance. Simulations are conducted with fixed spatial decay λ_f and process noise q.

In the low q/λ_f^2 regime (blue line), the filter relies primarily on model predictions, with minimal process noise influence. This scenario corresponds to a higher λ value, implying that covariance matrix correlations persist over larger spatial distances, with significantly slower decay. Here, a larger truncation radius is needed to preserve accuracy. NMSE and NIS converge only for higher κ values, and both Γ_P and Γ_{KH} exhibit slow decay, consistent with the extended range of spatial correlations and the expected less compact nature of $\bar{P}_r$ for smaller q/λ_f^2.

Fig. 3.11 investigates the impact of increasing system size n while keeping λ_f and process noise q fixed. We aim to determine the minimum κ for near-optimal performance,

defined as

$$\frac{1}{K}\sum_{k=1}^{K}\frac{|\hat{\boldsymbol{x}}_k^{\kappa}-\boldsymbol{x}_k|^2}{|\hat{\boldsymbol{x}}_k-\boldsymbol{x}_k|^2}<1+10^{-3}\ .$$

The top plot shows that the minimum required κ for a bounded estimation error bias (Theorem 3.4.1) initially increases with n. Crucially, for $n \geq 20$, κ plateaus and becomes independent of n. This indicates that the relevant interaction radius for accurate estimation, directly linked to decay parameter λ, is a *local property* of the system, not a function of its global dimension. The bottom plot shows that the normalized covariance perturbation Γ_P due to truncation remains essentially constant as n increases, consistent with SED theory and Theorem 3.3.1's implications for $\bar{P}$ decay. This reaffirms that the amount of neglected information, and thus pruning efficacy, depends only on local decay characteristics (captured by q and λ_f), not on the global system dimension.

3.5 Discussion

This chapter analyzes the sparsity of the Kalman filter for large-scale systems. A novel distance metric for state variables was introduced, revealing the exponential decay of key Kalman filter matrices with respect to this metric. This fundamental property enabled the approximation of system matrices through truncation, and we rigorously quantified the induced estimation bias. Our analysis confirms that this approximation effectively maintains high estimation accuracy while significantly reducing computational complexity, thereby establishing the scalability of our approach for large-scale networks. This work provides a principled method for efficient and accurate state estimation in extensive, interconnected environments. The inherent spatial decay characteristic of Kalman filter matrices, as demonstrated herein, opens exciting avenues for future research. We intend to leverage this property for clustering state variables, enabling the identification of naturally formed, tightly coupled subsystems. Furthermore, this localization will be instrumental in the design of truly distributed Kalman filters, where computations can be performed locally within interconnected regions, significantly enhancing scalability and real-time applicability in vast sensor networks and other complex, interconnected systems.

Chapter 4

Multirate Interlaced Kalman Filter and Attack Detection

Chapter 3 established that the matrices governing the Kalman filter—most notably the state transition, covariance, and gain matrices—exhibit *spatial sparsity* and *exponential decay* with distance in large-scale dynamical systems. This structural property enables scalable implementations of estimation algorithms, as computations and information exchanges can be localized with limited loss of accuracy. Building upon these foundations, this chapter addresses a complementary and pervasive aspect of distributed estimation: *asynchrony*.

In realistic networked cyber–physical systems, sensors, controllers, and estimators rarely operate in perfect synchrony. Differences in sampling rates, communication delays, and intermittent transmissions give rise to *multirate behavior*, where local estimators receive neighbor information at irregular intervals. If this temporal misalignment is ignored, the statistical coherence of the estimates degrades, leading to inconsistent covariance propagation and unreliable residuals—an issue that severely impacts model-based attack and attack detection.

To address these challenges, this chapter introduces the *Multirate Interlaced Kalman Filter* (MIKF), a distributed estimation scheme designed to maintain *covariance consistency* in asynchronous environments. The key idea is conceptually simple yet powerful: each local estimator treats incoming information from its neighbors as potentially *stale* and explicitly inflates its uncertainty before fusion. This time-dependent correction compensates for information aging, ensuring that local and global estimates remain statistically consistent even when subsystems operate at different rates or experience communication irregularities.

From a security perspective, such consistency is crucial. Residual-based detectors rely on well-characterized innovation statistics to guarantee correct detection thresholds. When updates are asynchronous, these properties may be distorted, leading to false alarms or undetected attacks. By explicitly accounting for the age of information, the proposed MIKF preserves the validity of residual statistics and extends model-based attack detection to multirate and asynchronous distributed settings.

The developments presented in this chapter extend the framework established in Chapter 3 along two main directions: (i) a structure-preserving decomposition of large-scale CPSs into weakly coupled subsystems, enabling distributed estimation under asynchronous operation while maintaining global consistency; and (ii) the design of residual-based statistical tests that preserve nominal false-alarm properties under delayed, heterogeneous, or missing updates, ensuring reliable detection of false data injection attacks.

Organization of the chapter.

- **Section 4.1.1** formalizes the distributed cyber–physical system (CPS) model, establishing the graph-based representation of subsystem interconnections and the decomposition inherited from the sparsity framework. It defines the continuous–time dynamics of each subsystem, together with the noise structure and coupling relationships, laying the foundation for localized estimation and control in large-scale systems.
- **Section 4.1.2** introduces the attacker model and the formulation of false data injection attacks. The section specifies the assumptions on attacker knowledge, capabilities, and scope of action, highlighting the implications of local and omniscient adversaries on the estimation process and motivating the need for resilient filtering mechanisms.

- **Section 4.2.1** presents the proposed *Multirate Interlaced Kalman Filter* . The section details the structure of local filters, the asynchronous prediction and correction steps, and the introduction of covariance compensation to counteract the effects of stale information. A formal stability analysis demonstrates that the proposed approach preserves mean-square boundedness and local exponential stability under asynchronous sampling and bounded disturbances.
- **Section 4.2.3** discusses the notions of *consistency* and *information content* as complementary indicators of estimator performance. It distinguishes between statistical and covariance consistency, defines quantitative measures based on innovation statistics, and analyzes how the proposed compensation mechanism maintains the desired balance between reliability and informativeness in distributed estimation.
- **Section 4.3** develops the residual-based attack detection framework tailored to the MIKF. Leveraging the statistical characterization of the innovations, it formulates hypothesis tests with analytically derived thresholds, ensuring a prescribed false-alarm probability. This section establishes how the proposed compensation allows these tests to remain valid even under asynchronous operation, enhancing the robustness of cyber–attack detection.
- **Section 4.4** provides numerical validation of the theoretical results through simulations on a networked mass–transport system. It illustrates the impact of asynchronous updates, the effectiveness of covariance compensation, and the behavior of the detection statistics under both nominal and attack conditions. The results confirm that the proposed MIKF preserves consistency and enables reliable attack detection despite heterogeneous sampling times.
- **Section 4.5** concludes the chapter by summarizing the main findings and emphasizing the practical implications of the proposed framework. It highlights the trade-off between consistency and information content and outlines potential extensions, including adaptive compensation tuning, event-triggered communication, and application to time-varying network topologies.

4.1 Problem Statement

4.1.1 System Model

The distributed system analyzed in this chapter builds directly upon the continuous–time dynamical model introduced in Section 3.1 and the sparsity framework developed in Chapter 3. We consider a large–scale cyber–physical system (CPS) whose overall behavior is described by the continuous–time linear model

$$\dot{\boldsymbol{x}}_t = L\boldsymbol{x}_t + \tilde{\boldsymbol{w}}_t, \qquad \boldsymbol{y}_t = H\boldsymbol{x}_t + \boldsymbol{v}_t, \tag{4.1}$$

where the notation follows that established in Chapter 3.1. Here, L and H denote the state and observation matrices, while $\tilde{\boldsymbol{w}}_t$ and $\boldsymbol{v}_t$ represent zero–mean white Gaussian noise processes with covariances Q_c and R_c, respectively.

As shown in Chapter 3, the entries of the matrices involved in (4.1)—and of the corresponding steady–state covariance and Kalman gain—exhibit *spatial exponential decay* with respect to the graph distance $d(i, j)$ defined on the system's interaction graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$. This decay property implies that distant states influence each other only weakly, thereby justifying a decomposition of the global system into smaller, locally coupled subsystems.

Let the set of states $\boldsymbol{x}_t$ be partitioned into N subsystems, each associated with a subset of state variables. The dynamics of subsystem i can then be expressed as the local

restriction of (4.1):

$$\begin{aligned} \dot{\boldsymbol{x}}_t^{(i)} &= L^{(i,i)}\boldsymbol{x}_t^{(i)} + \sum_{j\in\mathcal{N}_i^{\text{in}}} L^{(i,j)}\boldsymbol{x}_t^{(j)} + B^{(i)}\boldsymbol{u}_t^{(i)} + \tilde{\boldsymbol{w}}_t^{(i)}, \\ \boldsymbol{y}_t^{(i)} &= H^{(i)}\boldsymbol{x}_t^{(i)} + \boldsymbol{v}_t^{(i)}, \end{aligned} \tag{4.2}$$

where $\boldsymbol{x}_t^{(i)} \in \mathbb{R}^{n_i}$ is the local state vector of subsystem i, $\boldsymbol{u}_t^{(i)}$ and $\boldsymbol{y}_t^{(i)}$ are the corresponding local input and output signals, $L^{(i,j)}$ captures the coupling between subsystems i and j, and $\mathcal{N}_i^{\text{in}}$ denotes the set of neighbors of subsystem i in $\mathcal{G}$.

The noise processes $\tilde{\boldsymbol{w}}_t^{(i)}$ and $\boldsymbol{v}_t^{(i)}$ are assumed to be zero–mean, Gaussian, and mutually uncorrelated across subsystems, with local covariance matrices $Q_c^{(i)}$ and $R_c^{(i)}$.

This structure yields a *distributed continuous–time model* that preserves the spatial sparsity and locality properties identified in Chapter 3. Each subsystem interacts only with its direct neighbors, enabling scalable estimation and control architectures that approximate the behavior of a centralized Kalman filter within controlled error bounds.

To implement local estimation and control, each subsystem is assigned a computational unit called a *Control Agent*.

Definition 4.1.1 (*Control Agent*)

A Control Agent *is the local computational entity responsible for:*

1. *estimating the state of its subsystem through a continuous–time observer or filter;*
2. *generating control inputs based on the estimated state; and*
3. *evaluating residuals for attack detection.*

Each agent operates autonomously using its local model and measurements, while exchanging limited information with neighboring agents—typically the estimated state $\hat{\boldsymbol{x}}^{(i)}(t)$ and covariance $P^{(i,i)}(t)$—to account for inter–subsystem couplings. □

In practice, each agent:

- maintains a local continuous–time state estimate $\hat{\boldsymbol{x}}^{(i)}(t)$ and covariance $P^{(i,i)}(t)$;
- exchanges selected information with neighboring agents $j \in \mathcal{N}_i^{\text{in}}$ to maintain statistical consistency across the network; and
- computes residuals for fault or attack detection at runtime.

While Chapter 3 assumed synchronized and centralized estimation, practical CPSs frequently operate under *asynchronous* or *multirate* conditions, where local observers and communication channels evolve at different time scales. The central challenge addressed in this chapter is thus to extend the estimation process defined by (4.2) to such heterogeneous timing, while ensuring covariance consistency and reliable residual statistics. This motivates the development of the *Multirate Interlaced Kalman Filter* (MIKF), introduced in Section 4.2.1.

4.1.2 Attacker Model

In this chapter, a False Data Injection attack is considered. Let us make the following assumptions.

- We assume the communication channel among the control agents is secure, and thus, the information exchange is trusted.
- The attacker has perfect knowledge of $(F^{(i,i)}, B^{(i,i)}, H^{(i,i)})$, whereas he does not know the interconnection with neighboring systems.

- An attacker can only compromise a single subsystem i by injecting false sensor readings into a subset of sensors ($S_{bad}^{(i)}$).

Assuming an omniscient local attacker makes sense in practical settings, especially when model information can be acquired through intelligence-gathering methods. Additionally, we focus on the worst-case scenario where an omniscient local attacker is the most challenging to detect. This implies that at a given instant t, an attacker may be located near a subsystem.

The attacker model can be described as

$$\mathbf{y}_t^{(i)} = H^{(i,i)}\mathbf{x}_t^{(i)} + \mathbf{v}_t^{(i)} + \Gamma^{(i)} y_k^{(i),a} \tag{4.3}$$

where $\Gamma^{(i)} = diag(\alpha_1^{(i)}, ...\alpha_{m_i}^{(i)})$ is the sensor selection matrix; $\alpha_n^{(i)} = 1$ if and only if $n \in S_{bad}^{(i)}$; otherwise, $\alpha_n^{(i)} = 0$. $y_k^{(i),a} \in \mathbb{R}^{n_{y_i}}$ is the malicious input of the attacker. The main difficulty in detecting such cyberattacks is that the attacker can design $y_k^{(i),a}$ so that the effect is covert and is not distinguishable from the nominal behavior.

4.2 Observer Design

4.2.1 Multirate Interlaced Kalman Filter

The Interlaced Kalman Filter (IKF) enables distributed estimation of dynamical systems like Eq.(4.1) when discretized:

$$\begin{aligned} \boldsymbol{x}_{k+T_c^{(i)}}^{(i)} &= F^{(i,i)}\boldsymbol{x}_k^{(i)} + \sum_{j \in \mathcal{N}_i^{in}} F^{(i,j)}\boldsymbol{x}_k^{(j)} + B^{(i,i)}\boldsymbol{u}_t^{(i)} + \boldsymbol{w}_k^{(i)} \\ \boldsymbol{y}_k^{(i)} &= H^{(i,i)}\boldsymbol{x}_k^{(i)} + \boldsymbol{v}_k^{(i)} \end{aligned} \tag{4.4}$$

Here, $T_c^{(i)}$ denotes the sampling time of the i-th control agent, and the matrices $F^{(i,i)}$, $F^{(i,j)}$, $B^{(i,i)}$, and $H^{(i,i)}$ are derived from $F^{(i,i)}$, $F^{(i,j)}$, $B^{(i,i)}$ and $H^{(i,i)}$, once the sampling time is established. Each system, identified by the superscript, has a single control agent.

Let us make the following assumptions.

- Only the local dynamics matrices $F^{(i,i)}$, $B^{(i,i)}$, $H^{(i,i)}$, and the interconnection matrices $F^{(i,j)}$, $\forall j \in \mathcal{N}_i$, are available to the i-th control agent. This is a standard assumption for implementing IKF since information exchange between agents is constrained, necessitating distributed processing.
- The directed cyber-layer graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ is fixed and connected. A fixed and connected network topology ensures consistent communication among subsystems over time. Stability in the network structure allows for reliable data exchange, a crucial aspect in distributed systems.
- Each system i is locally observable, that is, the observability matrix $\left[H^{(i,i)}; H^{(i,i)}F^{(i,i)}; \ldots, H^{(i,i)}F^{(i,i)(n_i-1)}\right]^T$ has rank n_i, $\forall i \in \mathcal{V}$, $\forall k \geq 0$. For effective filtering, each system must be locally observable. Observability ensures that the state can be accurately estimated using the available measurements and dynamics. Without observability, the filtering algorithm may fail to converge to meaningful solutions.
- For any $\bar{k} \in \left[k; k + T_c^{(i)}\right]$, there exists a positive semidefinite, symmetrical $D_{\bar{k}}^{(i)}$, such that

$$\mathbb{E}\left\{\left(x_{\bar{k}}^{(i)} - x_k^{(i)}\right)\left(x_{\bar{k}}^{(i)} - x_k^{(i)}\right)^T\right\} \leq D_{\bar{k}}^{(i)}, \quad \forall i \in \mathcal{V}. \tag{4.5}$$

  where $T_c^{(i)}$ is the sampling time of the system i. Limited dynamic behavior implies that the state of each subsystem does not exhibit unbounded variation over time.

This assumption is essential to ensure consistency in distributed state estimation. By bounding the dynamics of the state, the filtering algorithm can maintain stable and reliable estimates. In other words, the existence of $D^{(i,j)}_{m_{(i,j)}}$ guarantees that over a limited time span, the variation of the state variables remains bounded.

In the IKF framework, each control agent develops its own filter to estimate the subsystem's state vector and exchanges information with neighboring agents to understand the system's current state. The information received from neighbors is treated as constant until the next update.

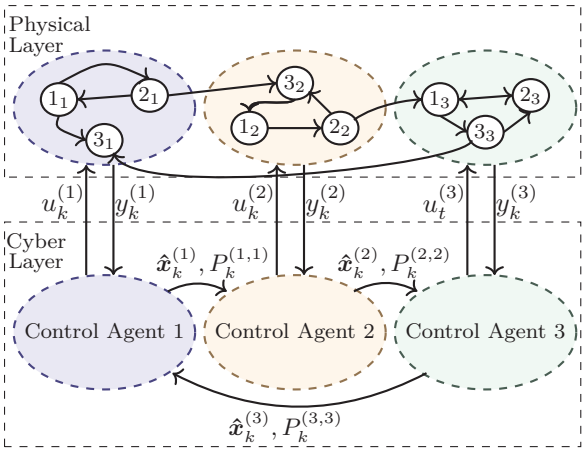


Figure 4.1: Example of physical and cyber interconnections. In the cyber layer, information exchange supports the implementation of the IKF. $\hat{\boldsymbol{x}}_k^{(i)}$ and $P_k^{(i,i)}$ denote, respectively, the state estimate and the covariance matrix of the i-th subsystem at time k.

Fig. 4.1 illustrates a physical system whose state can be estimated using IKF, with different colors and indices representing each subsystem. The system is partitioned into three subsystems, each with its state variables and control agent. The evolution of these variables depends on their history and interactions with variables in neighboring systems. Consider a large electrical grid that is divided into multiple subsystems. Each subsystem is managed by a different control agent. These control agents have knowledge only of their respective subsystem's internal state and the influence exerted on it by other subsystems.

To implement IKF, all stations must be perfectly synchronized. We consider a scenario with a looser assumption: each system operates at a predetermined sampled time denoted by $T_c^{(i)}$, when the prediction and correction cycles are executed. This assumption reflects the needs of many practical scenarios in which different control agents cannot operate simultaneously.

Synchronization is a cornerstone in distributed state estimate and control in various aspects such as security, attack detection, system stability, performance, and reliability.

In terms of security, a well-synchronized system ensures harmony among subsystems, maintaining a consistent state across the entire system. However, if synchronization is compromised, it could lead to inconsistencies in the system state, allowing attackers to exploit these vulnerabilities and potentially lead to security breaches.

When it comes to attack detection, predictability and consistency are key. Any deviation from the expected behaviour can be easily identified as an attack in a synchronised system. However, a lack of synchronization could lead to unpredictable system behaviour, complicating distinguishing between normal operation and anomalies and potentially increasing the rate of false positives or negatives.

System stability can also be affected by synchronization issues. For instance, in a control system, unsynchronized control inputs could lead to unstable system behaviour.

From a performance perspective, synchronization issues can lead to inefficiencies and reduced performance, as subsystems might struggle to effectively share resources or work together towards common goals.

Finally, regarding reliability, a synchronized system can continue to operate in a coordinated manner even if one subsystem fails. However, without proper synchronization, a single subsystem failure could have a cascading effect, potentially causing the entire system to fail.

For this, it is important to manage the lack of synchronization to ensure the effective operation of the system.

Remark 4.2.1

Due to the difference in operating times, the filters will run prediction and correction cycles at different time instants.In this chapter, for simplicity of notation, we denote the instant at which the filter i operates as k, but it is crucial to keep this distinction in mind nonetheless. □

When two control agents fall out of synchronization, their information about each other's system state becomes outdated. Synchronization in this scenario occurs at the least common multiple of their operating times. The notation $m_{i,j}$ denotes the last synchronized time instant between control agents i and j. At this point, the control agent i has access to the latest available estimation $\hat{\boldsymbol{x}}^{(j)}_{m_{i,j}}$ and $P^{(j,j)}_{m_{i,j}}$ of the system j. However, with time passing since the last update, $\hat{\boldsymbol{x}}^{(j)}_{m_{i,j}}$ may no longer accurately represent the current state of the system j, and $P^{(j,j)}_{m_{i,j}}$ might not sufficiently capture the uncertainty in the estimate.

To ensure covariance consistency, we need a measure of the mismatch between the available estimate and the actual state of the neighbors, or at least an approximation thereof. Thus, we propose a formalized approach to increase the uncertainty of the neighboring filter's estimated state, leveraging Eq. 4.5. Specifically, we exploit the fact that the covariance of the evolution of the neighboring system is upper bounded by $D^{(i,j)}_{m_{i,j}}$, and thus $\Pi^{(i,j)}_k = P^{(j,j)}_{m_{i,j}} + D^{(i,j)}_{m_{i,j}}$ increases uncertainty based on the available information. With these considerations, we present the IKF equations with compensation as follows.

Information Exchange:

$$\begin{aligned} \hat{\boldsymbol{x}}^{(i)}_{k-T_c^{(i)}} &= \hat{\boldsymbol{x}}^{(j)}_{m_{i,j}} \\ \Pi^{(i,j)}_{k-T_c^{(i)}} &= P^{(j,j)}_{m_{i,j}} + D^{(ij)}_{m_{i,j}} \end{aligned} \tag{4.6}$$

Prediction Step:

$$\begin{aligned} \hat{\boldsymbol{x}}^{(i)}_{k|k-T_c^{(i)}} &= F^{(i,i)}\hat{\boldsymbol{x}}^{(i)}_{k-T_c^{(i)}} + B^{(i,i)}\boldsymbol{u}^{(i)}_{k-T_c^{(i)}} + \\ &\quad + \sum_{j\in\mathcal{N}_i^{in}} F^{(i,j)}\hat{\boldsymbol{x}}^{(j)}_{k-T_c^{(i)}} \\ \tilde{Q}^{(i)}_{k-T_c^{(i)}} &= F^{(i,i)}P^{(i,i)}_{k-T_c^{(i)}}F^{(i,i)T} + 4|\mathcal{N}_i^{in}|\cdot \\ &\quad \sum_{j\in\mathcal{N}_i^{in}} F^{(i,j)}\Pi^{(i,j)}_{k-T_c^{(i)}}F^{(i,j)T} + Q^{(i)} \\ P^{(i,i)}_{k|k-T_c^{(i)}} &= F^{(i,i)}P^{(i,i)}_{k-T_c^{(i)}}F^{(i,i)T} + \tilde{Q}^{(i)}_{k-T_c^{(i)}} \end{aligned} \tag{4.7}$$

Correction Step:

$$\begin{aligned} S_k^{(i)} &= H^{(i,i)}P^{(i,i)}_{k|k-T_c^{(i)}}H^{(i,i)T} + R^{(i)} \\ K_k^{(i)} &= P^{(i,i)}_{k|k-T_c^{(i)}}H^{(i,i)T}\left[S_k^{(i)}\right]^{-1} \\ \hat{\boldsymbol{x}}_k^{(i)} &= \hat{\boldsymbol{x}}^{(i)}_{k|k-T_c^{(i)}} + K_k^{(i)}\left(\boldsymbol{y}_k^{(i)} - H^{(i,i)}\hat{\boldsymbol{x}}^{(i)}_{k|k-T_c^{(i)}}\right) \\ P_k^{(i,i)} &= (I - K_k^{(i)}H^{(i,i)})P^{(i,i)}_{k|k-T_c^{(i)}} \end{aligned} \tag{4.8}$$

where $\boldsymbol{x}^{(i)}_{k|k-T_c^{(i)}}$ and $P^{(i,i)}_{k|k-T_c^{(i)}}$ are, respectively, the a-priori state estimate and covariance matrix at time k, $K_k^{(i)}$ is the Kalman gain, $\hat{\boldsymbol{x}}_k^{(i)}$ and $P_k^{(i,i)}$ are, respectively,

the a-posteriori state estimate and a-posteriori covariance at time k. The reason $\tilde{Q}^{(i)}_{k-T_c^{(i)}}$ is updated in this way is described in the proof of Theorem 4.2.2.

4.2.2 Stability Analysis under Asynchronous Operation

We now analyze the stability properties proposed when subsystems operate at different sampling times, i.e., when each local filter i executes prediction and correction steps at discrete instants k_i, separated by its own control period $T_c^{(i)}$. In this asynchronous setting, the information received from neighboring filters may be outdated due to the lack of synchronization. To ensure consistency and boundedness of the estimation error, we analyze how the proposed covariance compensation term (see Eq. 4.6) affects the filter's convergence.

Let $\boldsymbol{e}_k^{(i)} = \boldsymbol{x}_k^{(i)} - \hat{\boldsymbol{x}}_k^{(i)}$ and $\boldsymbol{e}^{(i)}_{k|k-T_c^{(i)}} = \boldsymbol{x}_k^{(i)} - \hat{\boldsymbol{x}}^{(i)}_{k|k-T_c^{(i)}}$ denote the a priori and a posteriori estimation error for subsystem i at time k. We assume that the bounds and conditions listed below hold for all $k \geq 0$.

Assumption 4.2.1

We assume the following upper and lower bounds:

$$\underline{p}^{(i)} I \leq P_{k+1}^{(i)} \leq \overline{p}^{(i)} I \tag{4.9}$$

$$\underline{q}^{(i)} I \leq P_{k+1|k}^{(i)} \leq \overline{q}^{(i)} I \tag{4.10}$$

$$r_1 I \leq R^{(i)}, \qquad q_1 I \leq Q^{(i)} \tag{4.11}$$

$$\underline{p}^{(j)} I \leq P_{k+1}^{(j)} \leq \overline{p}^{(j)} I \tag{4.12}$$

$$\underline{h}^{(i)} I \leq ||H_{k+1}^{(ii)}|| \leq \overline{h}^{(i)} I \tag{4.13}$$

$$||F_k^{(ii)}|| \leq \overline{f}^{(i)} \tag{4.14}$$

$$\underline{f}^{(j)} \leq ||F_k^{(ij)}|| \tag{4.15}$$

$$\underline{h}^{(j)} I \leq ||H_{k+1}^{(ij)}|| \tag{4.16}$$

$$||w_k^{(i)}|| \leq \overline{w}^{(i)}, \qquad ||v_k^{(i)}|| \leq \overline{v}^{(i)} \tag{4.17}$$

□

First, we show that under the holding conditions, for each subsystem the local Kalman Gain is bounded bu a measurable quantity.

Lemma 4.2.1

If the conditions given in Assumption 4.2.1 holds then an upper bound of the norm of the Kalman gain matrix is given by

$$||K^{(i)}_{k+T_c^{(i)}}|| \leq \frac{\overline{q}^{(i)}\overline{h}^{(i)}}{\underline{q}^{(i)}\underline{h}^{(i)^2} + r_1} = \tilde{\kappa} \tag{4.18}$$

□

Before analyzing the full stability of the MIKF, it is necessary to establish how the compensated state transition matrix $\tilde{F}^{(i,i)\,T} = \left[I - K^{(i)}_{k+T_c^{(i)}} H^{(i,i)}\right]^T F_k^{(i,i)\,T}$ affects the propagation of the estimation error. The following lemma provides an upper bound on the evolution of the weighted error energy, ensuring that the compensated dynamics remain contractive in the mean-square sense.

Lemma 4.2.2

Let Assumption 4.2.1 hold. Consider subsystem i at its correction instants k and let

$$\tilde{F}^{(i,i)} := \left(I - K^{(i)}_{k+T_c^{(i)}} H^{(i,i)}\right) F_k^{(i,i)},$$

where $F_k^{(i,i)}$ and $H^{(i,i)}$ are the local state/output blocks, and $K_k^{(i)}$ is the Kalman gain of subsystem i. Assume $F_k^{(i,i)}$ is nonsingular for all $k \geq 0$. Then there exists a constant $\gamma^{(i)} \in (0,1)$ such that

$$\tilde{F}^{(i,i)\,T}\, P^{(i)\,-1}_{k+T_c^{(i)}}\, \tilde{F}^{(i,i)} \;\leq\; \big(1-\gamma^{(i)}\big)\, P_k^{(i)\,-1}. \tag{4.19}$$

□

The following theorem establishes the local stability properties of the MIKF under bounded disturbances and asynchronous sampling conditions. If all subsystems are proven to be stable, then the overall interconnected estimation process inherits global stability. The following theorem formalizes this result for the MIKF.

Theorem 4.2.1 (*Stability of the MIKF (linear case)*)

let Assumption 4.2.1 hold and Let $\tilde{\kappa}$ be the uniform upper bound on $\|K_k^{(i)}\|$ from Lemma 4.2.1, and let $\gamma^{(i)} \in (0,1)$ be the contraction constant given by Lemma 4.2.2. Then the following statements hold.

1. *If $\overline{\boldsymbol{w}}^{(i)} = \overline{\boldsymbol{v}}^{(i)} = 0$, there exists $\epsilon^{(i)} > 0$ such that for all initial errors with $\|\boldsymbol{e}_0^{(i)}\| \leq \epsilon^{(i)}$ the origin is (locally) exponentially stable:*

$$\|\boldsymbol{e}^{(i)}_{k+T_c^{(i)}}\| \leq c\,\beta^{\,k+1}\|\boldsymbol{e}_0^{(i)}\|, \quad \text{for some } c>0,\ \beta \in (0,1).$$

2. *If the noises satisfy $\|\boldsymbol{w}_k^{(i)}\| \leq \overline{\boldsymbol{w}}^{(i)}$ and $\|\boldsymbol{v}_k^{(i)}\| \leq \overline{\boldsymbol{v}}^{(i)}$ for all k, then there exists $\epsilon^{(i)} > 0$ such that for all $\|\boldsymbol{e}_0^{(i)}\| \leq \epsilon^{(i)}$ the error is uniformly bounded, with the explicit bound*

$$\big\|\boldsymbol{e}^{(i)}_{k+T_c^{(i)}}\big\|^2 \;\leq\; \frac{\overline{p}^{(i)}}{\underline{p}^{(i)}}\big(1+\xi^{(i)}\big)^{k+1}\,\big\|\boldsymbol{e}_0^{(i)}\big\|^2 \;+\; \frac{\overline{p}^{(i)}}{\big|\xi^{(i)}\big|}\,\rho^{(i)}\big(\tilde{\kappa}\big), \tag{4.20}$$

where

$$\xi^{(i)} := \frac{\gamma^{(i)}(1-\psi^{(i)})}{\psi^{(i)}} \in (-1,0) \quad \text{for any } \psi^{(i)} > 1,$$

and

$$\rho^{(i)}\big(\tilde{\kappa}\big) := \frac{1}{\underline{p}^{(i)}}\left(\frac{\overline{p}^{(i)}}{\underline{q}^{(i)}}\,\overline{\boldsymbol{w}}^{(i)} \;+\; \tilde{\kappa}\,\overline{\boldsymbol{v}}^{(i)}\right)^2. \tag{4.21}$$

□

Theorem 4.2.1 demonstrates that the proposed IKF with asynchronous operation and covariance compensation preserves local exponential stability. The additional uncertainty introduced by the term $D_{m_{i,j}}^{(i,j)}$ compensates for outdated information without destabilizing the estimation process. As shown in [224], increasing $P_{m_{i,j}}^{(j,j)}$ effectively decreases the Kalman gain $K_k^{(i)}$, resulting in a slower convergence rate but improved consistency and robustness. In steady state, the residual error variance depends on $\rho^{(i)}(\tilde{\kappa}^{(i)})$ and $\xi^{(i)}(\tilde{\kappa}^{(i)})$, quantifying the trade-off between convergence speed and robustness to asynchrony.

4.2.3 Consistency and Information Content

Consistency and information content are crucial to evaluating the performance of the estimator. Consistency gauges how well the estimated distribution matches the true parameters, while information content reflects the estimate's precision.

Balancing consistency and information content is essential as emphasizing one can compromise the other. This balance is critical for attack detection using filters such

as the Kalman filter. It directly impacts the test's ability to detect anomalies while accurately minimizing false positives or negatives. Overemphasizing consistency could lead to a conservative test that misses anomalies. In contrast, prioritizing the content of the information might result in a sensitive test prone to false positives. False positives in a Kalman filter are about individual incorrect acceptances of erroneous measurements or state estimates, while statistical consistency pertains to the overall match between predicted uncertainties and actual errors, ensuring the reliability and accuracy of the filter's performance over time.

Thus, achieving an optimal balance ensures effective attack detection, accurately identifying deviations while minimizing false alarms.

4.2.3.1 Consistency

In the literature, consistency can be defined in various ways depending on how the uncertainty of the estimate is represented. Various formats, such as probability distributions, covariance matrices, or confidence regions provide unique insights into this notion. To streamline our discussion, let's introduce two key definitions: statistical consistency and covariance consistency.

- *Statistical Consistency.* A filter whose a-posteriori estimation error vector and estimation error covariance matrix at time k are $\boldsymbol{e}_k$ and P_k is said to be statistically consistent if
$$\mathbb{E}\{\boldsymbol{e}_k\} = 0, \quad \mathbb{E}\left\{\boldsymbol{e}_k \boldsymbol{e}_k^T\right\} = P_k, \quad \forall k > \bar{k}. \tag{4.22}$$
This means that for each time greater than $\bar{k}$ the probability distribution of the estimation error aligns exactly with the true distribution of the parameter. This definition is very strong, and can only be achieved if the dynamical system is linear and if its parameters are known exactly. A possibility to evaluate consistency without ground-through knowledge is by using the *normalized innovation error squared* (NIS) is a metric, defined as:
$$\boldsymbol{\epsilon}_k^{(i)} = \boldsymbol{\nu}_k S_k^{-1} \boldsymbol{\nu}_k^T \tag{4.23}$$
where $\boldsymbol{\nu}_k = \boldsymbol{y}_k^{(i)} - H^{(i,i)} \hat{\boldsymbol{x}}^{(i)}_{k|k-T_c^{(i)}}$. Eq. 4.23 is a χ^2-distributed random variable with q_i degrees of freedom, where q_i is the length of the output vector [225].
- *Covariance Consistency.* A filter whose a-posteriori estimation error vector and estimation error covariance matrix at time k are $\boldsymbol{e}_k$ and P_k, is said to be statistically consistent if
$$\mathbb{E}\left\{\boldsymbol{e}_k \boldsymbol{e}_k^T\right\} \leq P_k, \quad \forall k > \bar{k}. \tag{4.24}$$
In many practical applications, the filter used is suboptimal due to model mismatches, where the actual system dynamics deviate from the assumed model. In such cases, the most one can aspire to is to ensure covariance consistency.

4.2.3.2 Information Content

Information content pertains to a predefined level associated with estimating a parameter, inversely related to the covariance matrix of the estimate. The balance of consistency and information content is crucial. Ideally, statistical consistency is optimal, converging the estimate to the true distribution with more data. If only covariance consistency is achievable, ensuring proximity to the covariance consistency equation optimizes the balance between reliability and precision in estimation.

4.2.4 Consistency of Multirate Interlaced Kalman Filter

The Multirate Interlaced Kalman Filter can ensure consistency in estimates under certain assumptions.

Theorem 4.2.2

Considering Multirate Interlaced Kalman Filter equation, under the previously stated assumptions, the pair $\left(\hat{\boldsymbol{x}}^{(i)}_{k+T_c^{(i)}}, P^{(i,i)}_{k+T_c^{(i)}}\right)$ *of the algorithm in* (4.6)-*4.8 is consistent, i.e.,*

$$\mathbb{E}\left\{\left(\hat{\boldsymbol{x}}^{(i)}_{k+T_c^{(i)}} - \boldsymbol{x}^{(i)}_{k+T_c^{(i)}}\right)\left(\hat{\boldsymbol{x}}^{(i)}_{k+T_c^{(i)}} - \boldsymbol{x}^{(i)}_{k+T_c^{(i)}}\right)^T\right\} \leq P^{(i,i)}_{k+T_c^{(i)}}, \tag{4.25}$$

$\forall i \in \mathcal{V}, k \geq 0.$ □

See proof in Appendix C.

As highlighted by Th. 4.2.2, a key factor in achieving consistency is the compensation parameter $D^{(i,j)}_{m_{ij}}$, which requires balancing the lack of knowledge about the neighbor dynamic model with the need for consistency.

Before employing the filter for attack detection, a calibration phase is essential to estimate the steady-state matrix $D^{(i,j)}_{m_{ij}}$. The challenge is that according due to the distributed nature of the system and the estimator, the i-th filter does not know the dynamic model of the system j, so, at most, one can aspire to approximate the matrix $D^{(i,j)}_{m_{ij}}$. Note that:

$$\mathbb{E}\left\{\left(\boldsymbol{x}^{(j)}_{m_{ij}} - \boldsymbol{x}^{(j)}_{k}\right)\left(\boldsymbol{x}^{(j)}_{m_{ij}} - \boldsymbol{x}^{(j)}_{k}\right)^T\right\} = \left(I - \hat{F}^{(j)}_{k,m_{ij}}\right)\left[\hat{\boldsymbol{x}}^{(j)}_{m_{ij}}\hat{\boldsymbol{x}}^{(j)T}_{m_{ij}} + P^{(j,j)}_{m_{ij}}\right]\left(I - \hat{F}^{(j)}_{k,m_{ij}}\right)^T$$

where $k > m_{ij}$, and $\hat{F}^{(j)}_{k,m_{ij}}$ is an unknown matrix such that $\boldsymbol{x}^{(j)}_{k} = \hat{F}^{(j)}_{k,m_{ij}}\boldsymbol{x}^{(j)}_{m_{ij}}$, quantifying the evolution of the state variable between two different time instants. Since $\hat{F}^{(j)}_{k,m_{ij}}$ is not directly observable, its exact value is unknown. However, if the system dynamics are sufficiently smooth—meaning that the evolution of the state does not vary abruptly over short time intervals—it is reasonable to assume that the transition behaviour between intervals of equal duration remains similar. Under this assumption, $\hat{F}^{(j)}_{k,m_{ij}}$ can be *estimated* by analysing past iterations of the same subsystem, that is, by evaluating how the state has evolved between previous pairs of time instants separated by comparable time steps. This approach allows one to approximate the local rate of change of the state, capturing the typical dynamics of subsystem j even when direct synchronization with its neighbours is not available.

Considering the j filter, $\boldsymbol{x}^{(j)}$ is a vector of n_j components. Then for $l = 1, \ldots, n_j$, we find a value $\alpha^{j_l}_{m_{ij}}$ such that:

$$\boldsymbol{x}^{(j_l)}_{m_{ij}} = \alpha^{j_l}_{m_{ij}}\,\boldsymbol{x}^{(j_l)}_{m_{ij}-\beta_{ij}},\ \bar{\alpha}^{(j)}_{m_{ij}} = \texttt{diag}\left\{|\alpha^{(j_1)}_{m_{ij}}|, \ldots, |\alpha^{(j_{n_j})}_{m_{ij}}|\right\},$$

where $\beta_{ij} = lcm\left(T_c^{(i)}, T_c^{(j)}\right)$ is the time between two successive synchronizations between filter i and filter j, and with $|\cdot|$ we denote the absolute value. Taking into account the history, it can be computed:

$$\alpha^{(j)}_{m_{ij}} = \max\{\bar{\alpha}^{(j)}_{m_{ij}}, \alpha^{j}_{m^{(j)}_{ij}-\beta_{ij}}\},$$

Then

$$\mathbb{E}\left\{\left(\boldsymbol{x}^{(j)}_{m_{ij}} - \boldsymbol{x}^{(j)}_{k}\right)\left(\boldsymbol{x}^{(j)}_{m_{ij}} - \boldsymbol{x}^{(j)}_{k}\right)^T\right\} \leq \left(I - \alpha^{(j)}_{m_{ij}}\right)\left[\hat{\boldsymbol{x}}^{(j)}_{m_{ij}}\hat{\boldsymbol{x}}^{(j)T}_{m_{ij}} + P^{(j,j)}_{m_{ij}}\right]\left(I - \alpha^{(j)}_{m_{ij}}\right)^T = D^{(i,j)}_{m_{ij}}$$

holds. Assuming that the uncertainty of the estimates provided by the neighbors grows linearly, given a time instant k such that $k > m_{ij}$, we can define $\boldsymbol{\Delta}^{(i,j)}_{k} = \boldsymbol{x}^{(j)}_{m_{ij}} - \boldsymbol{x}^{(j)}_{k}$ and state that:

$$\mathbb{E}\left\{\left(\boldsymbol{\Delta}^{(i,j)}_{k}\right)\left(\boldsymbol{\Delta}^{(i,j)}_{k}\right)^T\right\} \leq D^{(i,j)}_{m_{ij}}. \tag{4.26}$$

In particular, according to the theorem to guarantee consistency, a choice for $\Pi^{(i,j)}_{k-T_c^{(i)}}$ is $\Pi^{(i,j)}_k = P^{(i,j)}_{m_{ij}} + D^{(i,j)}_{m_{ij}}$.

Remark 4.2.2

Calculating $\alpha^j_{m_{ij}}$ does not guarantee that $D^{(i,j)}_{m_{ij}}$ is an upper bound for the mismatch between the i-th filter's available value and the true state. However, $\alpha^{(j)}_{m_{ij}}$ serves as a reasonably accurate approximation during stable system conditions and initial learning stages. This step is needed to calculate a parameter that maximizes the filter information content. For consistency alone, without concern for information content, opt for a large $D^{(i,j)}_{m_{ij}}$. □

Once the parameters are calculated, Monte Carlo simulations are performed to provide insight into filter consistency and information content. In particular, for the i-th filter, we have:

$$\bar{\epsilon}^{(i)}_k = \frac{1}{V}\sum_{v=1}^{V} \boldsymbol{\epsilon}^{(i),[v]}_k \tag{4.27}$$

where $[v]$ is the index of the run, and $\epsilon^{(i),[v]}_k$ is the NIS metric defined in Eq. 4.23. Given the ordered temporal vector $\bar{\epsilon}^{(i)}$, where $\bar{\epsilon}^{(i)}_k$ represents the k-th element of $\bar{\epsilon}^{(i)}$, and considering a sufficiently large V (larger than 30), invoking the central limit theorem [226], we obtain the approximations:

$$\tilde{\epsilon}^{(i)} = \frac{1}{K}\sum_{k=1}^{K} \bar{\epsilon}^{(i)}_k \approx q_i,\ \ \tilde{S}^{(i)} = \frac{1}{K-1}\sum_{k=1}^{K}\left(\bar{\epsilon}^{(i)}_k - q_i\right)^2 \approx \frac{2q_i}{V}.$$

If $\tilde{\epsilon}^{(i)}$ and $\tilde{S}^{(i)}$ are too far from the theoretical values, adjustment of parameters may be necessary to strike a balance between consistency and information content. To evaluate the consistency of the filter, please refers to [225].

4.3 Attack Detection Method

The attack detection process hinges on evaluating the dissimilarity between the predicted measurement, derived from the preceding estimation and the system's dynamics, and the current actual measurement. Significant disparities between prediction and actual measurement show potential anomalies. This difference is defined as innovation. This disparity should be zero in an ideal scenario, indicating a perfect match between the predicted and the real measurements. However, natural fluctuations occur around this ideal in real-world scenarios due to dynamic and measurement errors. In particular, considering a $(1-a)$ confidence level, a false-positive rate roughly equal to a is physiological, so we do a batch analysis of innovation.

To do this, exploiting the central limit theorem [226], we built the following metric:

$$\hat{\boldsymbol{\epsilon}}^{(i)}_k = \frac{1}{K_b}\sum_{s=k-K_b}^{k} \epsilon^{(i)}_s \tag{4.28}$$

where K_b is the batch size and k is the current time instant. The value of K_b must be large enough (at least equal to 30). Then, if the filter is well calibrated, the mean and variance of Eq. 4.28 is known, so we perform the following mean Gaussian test [227].

$$q_i - z_{\alpha/2}\sqrt{\frac{2q_i}{K_b}} \leq \hat{\boldsymbol{\epsilon}}^{(i)}_k \leq q_i + z_{\alpha/2}\sqrt{\frac{2q_i}{K_b}} \tag{4.29}$$

Our proposed detection method triggers an alert when the chain of inequalities in Eq. 4.29 is unmet. A key advantage of this approach lies in deriving thresholds based

on the statistical properties of the observed quantity rather than through simulation or empirical methods. Consequently, we establish a known false alarm probability by setting a threshold and vice versa. For instance, with a threshold set at 95%, it is understood that 5% of samples could normally exceed the threshold, but accepting this level of false alarms allows us to detect true anomalies.

4.4 Simulations

Mass transport networks like water networks are common examples of cyber-physical systems. They involve modeling differential-algebraic equations to represent the flow dynamics. The key variables of interest are typically the water levels in each tank throughout the network. For this reason, in our simulation, we consider a linear model for 9 interconnected tanks whose dynamics is described in [228].

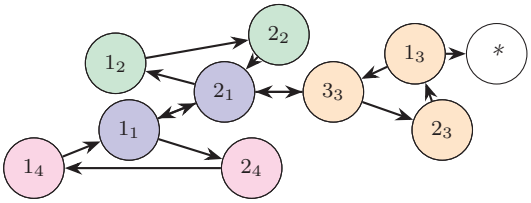


Figure 4.2: Physical connection graph. Colors denote different subsystems, and subscripts identify the variables within each control agent. The asterisk node represents an external portion of the environment that is connected to the system but not explicitly estimated.

The simulation's physical connection graph, as depicted in Fig. 4.2, represents tanks as nodes. The control agents are identified by an index ($i = 1, 2, 3, 4$, and $|\mathcal{V}| = 4$). A white node with an asterisk denotes the *outside world*, representing a portion of the system not explicitly modeled but still connected in some manner. The edges in the graph represent pipes connecting the tanks. Given the pipe connections among the tanks shown in Fig. 4.2, the state transition matrix can be computed as follows:

$$F = p\ (L - G) = \begin{bmatrix} F^{(1,1)} & \dots & F^{(1,N)} \\ \vdots & \ddots & \vdots \\ F^{(N,1)} & \dots & F^{(N,N)} \end{bmatrix}$$

In our simulation, we set $p = 0.4$ to represent the proportion of water exchanged among the tanks. The Laplacian matrix $L \in \mathbb{R}^{n \times n}$ corresponds to the physical layer graph excluding connections to the outside world. Additionally, G is a diagonal matrix where $g_{a,a}$ equals one if an edge connects v_a to the outside world, and zero otherwise. F represents the overall system transition matrix; with appropriate sorting of rows and columns, it is possible to obtain the partition into subsystems.

The output selection matrix and input matrix for each subsystem are, respectively:

$$H^{(i,i)} = \left\{ \begin{bmatrix} 0.55 & 0.44 \\ 0.71 & 0.03 \end{bmatrix}, \begin{bmatrix} 0.29 & 0.94 \\ 0.70 & 0.98 \end{bmatrix}, \begin{bmatrix} 0.40 & 0.44 & 0.70 \\ 0.47 & 0.20 & 0.69 \\ 0.27 & 0.20 & 0.38 \end{bmatrix}, \begin{bmatrix} 0.70 & 0.10 \\ 0.56 & 0.13 \end{bmatrix} \right\}$$

$$B^{(i,i)} = \left\{ \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 \\ 0 \end{bmatrix} \right\}$$

The control agents sample at times $T_c = [0.1, 0.2, 0.3, 0.5]$ seconds within a time window of 300 seconds. The model and measurement noise were set to $Q^{(i)} = \mathbb{I}_{n_i}$ and

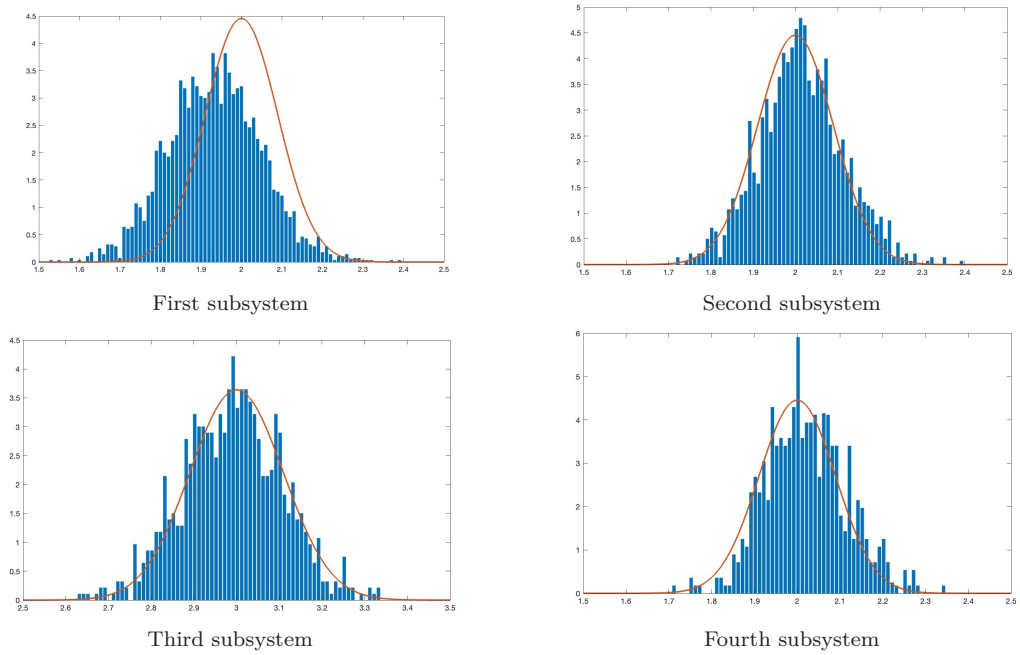


Figure 4.3: Frequency histogram of the realizations of Eq. 4.27 (blue) obtained from Monte Carlo simulations, compared with the theoretical χ^2 distribution (red).

$R^{(i)} = 0.5\,\mathbb{I}_{q_i}$ for each subsystem i, where $\mathbb{I}_a$ denotes the $a \times a$ identity matrix. The control inputs (one per control agent) are $u_k^{(i)} = 0.2\,i\,\sin(k)$.

We conducted 500 Monte Carlo simulations with a 300-second time window using the compensation parameter obtained during the calibration phase. For each system, we calculate the histogram of the normalized innovation squared (NIS) frequencies, as shown in Fig. 4.3. In these simulations, we computed the instantaneous averages and then plotted the histogram of this average across the entire simulation time. Notably, the observed sample distribution in Fig. 4.3 (first panel) for the first subsystem lies to the left of the theoretical distribution, indicating that compensation parameters larger than necessary were computed. Since our algorithm provides an upper bound on the compensation, this effect is most visible in the first subsystem, which is the most connected.

Reducing the compensation parameters computed for the first system, we re-ran the Monte Carlo simulations with the other parameters unchanged. This adjustment led to the sample distribution shown in Fig. 4.4, where the innovation distribution follows the theoretical χ^2 very closely.

Next, using the statistical batch test (window size $K_b = 100$ samples), we compared the compensated Multirate IKF with the traditional IKF (Fig. 4.5). Because filters run at different rates, test start times differ across subsystems. With our compensation in place, the test statistic remains within the admissible band with violations below 5%, matching the nominal false alarm rate. Without compensation (red), especially for faster filters, out-of-band excursions become frequent due to stale neighbor information, which compromises covariance consistency.

We then conducted a false data injection attack on the first sensor of the first subsystem (the most compensated one) in the window $t \in [100, 200]$ s by adding Gaussian noise $\mathcal{N}(0, 0.5)$. As shown in Fig. 4.6, with compensation the batch test fails significantly more often during the attack, correctly signaling an attack. This illustrates that the proposed compensation preserves nominal false alarm rates while enabling effective detection when the measurements are corrupted.

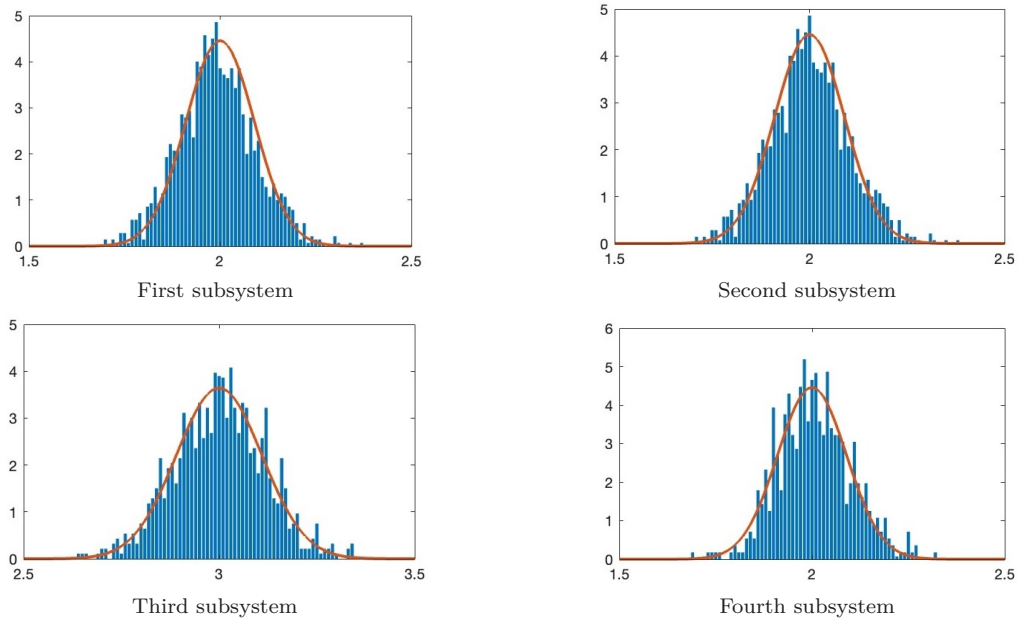


Figure 4.4: Frequency histogram of Eq. 4.27 (blue, $V = 500$) compared with the theoretical χ^2 distribution (red), under reduced compensation on subsystem 1. The proposed compensation represents an upper bound; reducing it can improve test informativeness while maintaining statistical consistency.

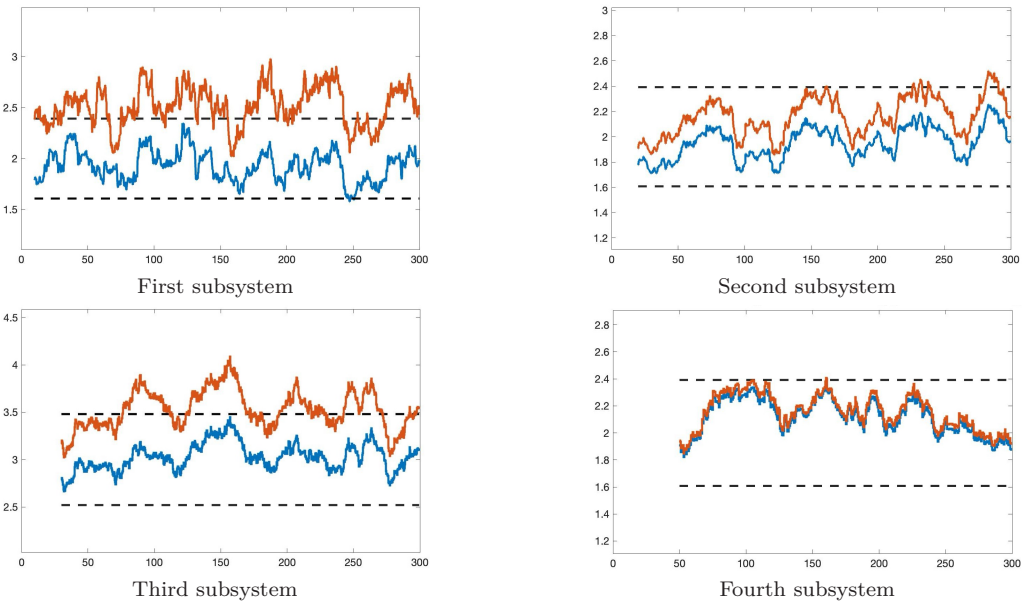


Figure 4.5: Comparison between the detection statistics $\hat{\epsilon}_k^{(i)}$ and the test thresholds (black). The proposed method after calibration (blue) is compared with the traditional IKF approach (red). For a correctly calibrated 95% confidence level, the statistic should remain within the thresholds, with an expected failure rate of approximately 5%.

4.5 Discussion

This chapter addressed attack detection in CPS under realistic communication and sampling constraints. We introduced a MIKF with covariance compensation that preserves *covariance consistency* while retaining as much *information content* as possible. The key insight is to treat neighbor estimates as time–aged signals and explicitly inflate their

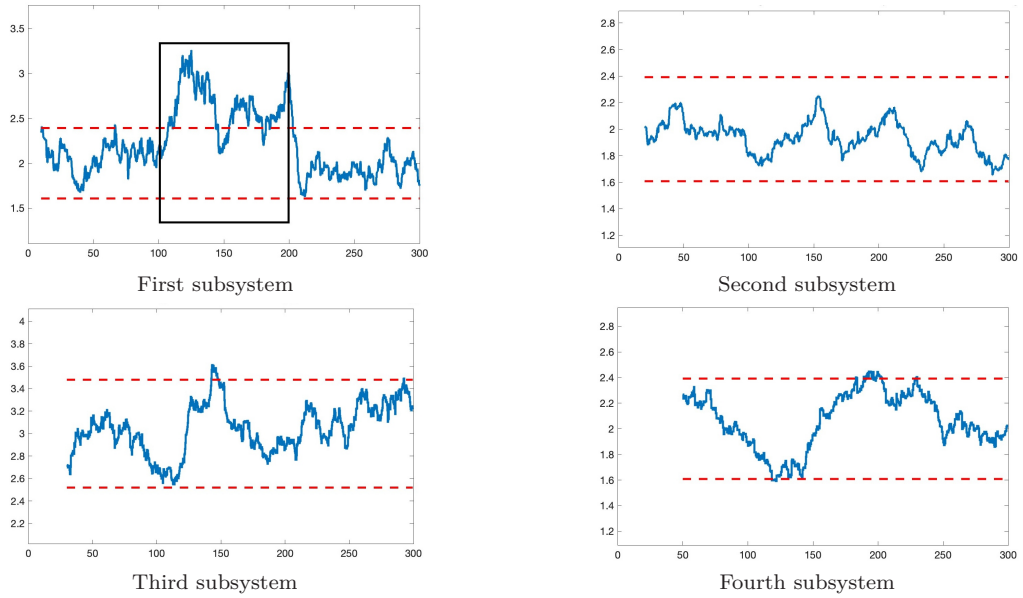


Figure 4.6: Batch detection statistic $\hat{\epsilon}_k^{(i)}$ (blue) under a fault detection and isolation (FDI) attack on the first sensor of subsystem 1, active between $t = 100$ and $t = 200$ s. The red lines indicate the Gaussian test thresholds used for anomaly detection.

uncertainty before using them in the local prediction step. This prevents overconfident covariances that would otherwise corrupt the statistics of the innovations and undermine residual-based detection.

Methodologically, we modified the a priori covariance computation $P_{k|k-1}^{(i,i)}$ to account for asynchronous updates, leveraging past synchronization instants to quantify how neighbor states may have evolved since the last exchange. This yields a principled compensation term that is simple to implement and compatible with heterogeneous sampling times—particularly valuable for fast subsystems that cannot always incorporate fresh neighbor information at every iteration.

On the detection side, we adopted a batch test on the normalized innovations, exploiting the central limit theorem to derive thresholds with a prescribed false alarm rate. The experiments showed two complementary effects. First, when asynchrony is ignored (traditional IKF), covariance consistency is frequently violated, inflating the false alarm rate and weakening detection. Second, with compensation enabled and after a light calibration of the compensation parameters, the empirical innovation statistics match their nominal distribution closely in nominal conditions, while deviations become conspicuous under false data injection, enabling reliable detection.

Two practical trade-offs emerge. Increasing compensation promotes consistency but may reduce informativeness (slower convergence, larger steady-state variances). Conversely, aggressive tuning of compensation can yield sharper estimates but risks overconfidence. In practice, a short calibration phase that minimizes the discrepancy between empirical and theoretical innovation statistics is effective to balance these objectives.

Several extensions are natural. First, one could automate the calibration of compensation terms via online adaptation (e.g., stochastic approximation on innovation moments). Second, robustness to model mismatch could be further improved by incorporating set-based or risk-sensitive updates. Third, communication-aware policies (event-triggered sharing of $(\hat{\boldsymbol{x}}, P)$) could reduce bandwidth while preserving consistency guarantees. Finally, extending the analysis to time-varying interconnection graphs and correlated process/measurement noises would broaden applicability.

Overall, the proposed MIKF with compensation restores the statistical validity of residuals in multirate, asynchronous settings and, as a result, significantly strengthens

model-based attack detection in distributed CPS.

Chapter 5

Sliding Mode Observer for Distributed Attack Reconstruction

An essential aspect of resilient estimation in networked control systems is the ability to *reconstruct the true physical measurements* when sensor data are corrupted, either by disturbances or deliberate attacks. Even in the presence of reliable communication and synchronization, local measurements may not faithfully reflect the actual system state due to false data injection, replay, or denial-of-service actions that compromise the sensing layer. When such corruptions propagate through distributed estimators, they distort the information exchanged among agents, leading to inconsistent state estimates and potentially unsafe control actions. Hence, attack-resilient observers must not only detect the presence of adversarial manipulation but also *recover the genuine measurement signal* required for correct estimation and control.

While the distributed filtering framework established in Chapter 3 enabled scalable state estimation under structural sparsity and asynchronous sampling, it did not explicitly address the case where measurements themselves become unreliable. Standard Kalman-based approaches are inherently statistical and rely on nominal noise models, which are easily violated under adversarial conditions. To ensure estimation reliability in such contexts, robustness must be introduced at the level of the observer dynamics, allowing each subsystem to reject and reconstruct corrupted information in real time.

This chapter introduces a *distributed Sliding Mode Observer (SMO)* architecture specifically designed for *attack reconstruction* in interconnected cyber–physical systems. The proposed approach integrates robust sliding-mode techniques with distributed estimation principles, allowing each agent to identify and compensate malicious measurement corruptions using only local information and limited neighbor communication. The main objective is twofold: to maintain accurate state estimation despite sensor attacks and to reconstruct the true underlying measurements that are essential for safe control and anomaly detection.

The core idea lies in a *two-layer modular structure*. The first layer enforces output robustness by driving the measurement residual to a sliding manifold, effectively rejecting unknown inputs and isolating the equivalent disturbance acting on the output channel. The second layer then exploits this equivalent signal to estimate both the system state and the attack component through a distributed extended observer. This decomposition enables a clean separation between robustness enforcement and signal reconstruction, ensuring that each subsystem contributes autonomously to the global estimation process without centralized coordination.

From a theoretical perspective, the chapter establishes explicit guarantees for finite-time convergence of the sliding variable, boundedness of the estimation errors under disturbances of the distributed estimation network. In addition, a detectability analysis demonstrates that the augmented system, which jointly models physical states and attack variables, remains structurally observable through local outputs and neighbor exchanges. The resulting framework achieves distributed attack reconstruction with quantifiable error bounds and no false isolation of benign nodes.

Overall, the distributed SMO framework presented here extends classical sliding-mode estimation principles to large-scale interconnected systems operating under adversarial sensing conditions. It provides both the analytical foundation and practical mechanisms for reconstructing genuine measurements, thereby enhancing the integrity and resilience of distributed estimation architectures.

Organization of the Chapter. The remainder of this chapter is structured as follows:

- **Section 5.1** introduces the distributed system model and formulates the attack reconstruction problem. It defines the structure of sensor corruptions, characterizes the main attack classes (FDI, replay, DoS), and outlines the detectability and boundedness assumptions.
- **Section 5.2** describes the overall modular observer architecture, highlighting the interaction between the sliding-mode regulation layer and the state–attack estimation layer within a distributed communication network.
- **Section 5.3** develops the Level 1 sliding-mode output regulator. It introduces the adaptive gain law, analyzes the output error dynamics, and establishes finite-time convergence of the sliding variable despite bounded disturbances.
- **Section 5.4** presents the Level 2 distributed state–attack observer, which leverages the equivalent injection generated by the first layer to reconstruct both the true state and the latent attack signal. Detectability and stability conditions are derived using an input-to-state stability and small-gain framework.
- **Section 5.5** integrates the results from the two layers, providing the main theoretical guarantees for distributed attack reconstruction and bounded estimation errors.
- **Section 5.6** illustrates the performance of the proposed framework through simulations on a multi-agent network subject to multiple concurrent attacks, highlighting reconstruction accuracy and scalability.
- **Section 5.7** concludes the chapter by summarizing the theoretical contributions, discussing their implications for resilient control, and outlining future research directions such as adaptive tuning, nonlinear extensions, and event-triggered communication.

5.1 System Model

5.1.1 Networked System Model

We consider a large-scale networked cyber–physical system composed of N interconnected linear subsystems (or agents), indexed by the finite set

$$\mathcal{V} := \{1, 2, \dots, N\}.$$

The physical and informational interconnections among agents are described by a directed graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $(j, i) \in \mathcal{E}$ indicates that the dynamics of node j influence those of node i. The in-neighborhood of node i is defined as

$$\mathcal{N}_i^{\text{in}} := \{\, j \in \mathcal{V} : (j, i) \in \mathcal{E} \,\}.$$

Each agent $i \in \mathcal{V}$ is modeled by the following dynamics:

$$\dot{\boldsymbol{x}}^{(i)}(t) = F^{(i,i)}\boldsymbol{x}^{(i)}(t) + \sum_{j \in \mathcal{N}_i^{\text{in}}} F^{(i,j)}\boldsymbol{x}^{(j)}(t) + B^{(i)}\boldsymbol{u}^{(i)}(t) + D^{(i)}\boldsymbol{w}^{(i)}(t), \tag{5.1a}$$

$$y^{(i)}(t) = H^i\boldsymbol{x}^{(i)}(t) + \boldsymbol{d}^{(i)}(t), \tag{5.1b}$$

where $\boldsymbol{x}^{(i)} \in \mathbb{R}^{n_i}$ is the local state, $\boldsymbol{u}^{(i)} \in \mathbb{R}^{m_i}$ the control input, $\boldsymbol{y}^{(i)} \in \mathbb{R}^{p_i}$ the measured output, $\boldsymbol{w}^{(i)}(t)$ an external disturbance, and $\boldsymbol{d}^{(i)}(t)$ an unknown sensor-side corruption.

The system matrices $F^{(i,i)}$, $B^{(i)}$, H^i, and $D^{(i)}$ are known and of appropriate dimensions, while $F^{(i,j)} \in \mathbb{R}^{n_i \times n_j}$ represents the coupling from node j to node i. For notational compactness, we define:

$$x^{(-i)} := \text{col}(x^{(j)})_{j \in \mathcal{N}_i^{\text{in}}}, \quad F^{(-i)} := [\, F^{(i,j)} \,]_{j \in \mathcal{N}_i^{\text{in}}},$$

and rewrite (5.1a) as:

$$\dot{x}^{(i)} = F^{(i,i)}\boldsymbol{x}^{(i)} + F^{(-i)}\boldsymbol{x}^{(-i)} + B^{(i)}\boldsymbol{u}^{(i)} + D^{(i)}\boldsymbol{w}^{(i)}.$$

We impose the following standard assumptions:

Assumption 5.1.1

For each $i \in \mathcal{V}$, the pair $(F^{(i,i)}, H^i)$ has no invariant zeros in the closed right half-plane. □

Assumption 5.1.2

The disturbance $\boldsymbol{w}^{(i)}(t)$ and measurement noise $\boldsymbol{v}^{(i)}(t)$ are bounded for all $t \geq 0$:

$$\|\boldsymbol{w}^{(i)}(t)\| \leq \bar{\boldsymbol{w}}^{(i)}, \quad \|v^{(i)}(t)\| \leq \bar{\boldsymbol{v}}^{(i)}, \quad \|\dot{\boldsymbol{v}}^{(i)}(t)\| \leq \bar{\dot{\boldsymbol{v}}}^{(i)},$$

for unknown but finite constants $\bar{\boldsymbol{w}}^{(i)}, \bar{\boldsymbol{v}}^{(i)}, \bar{\dot{\boldsymbol{v}}}^{(i)} > 0$. □

Assumption 5.1.1 ensures detectability, a critical prerequisite for any observer-based reconstruction method, particularly under bounded-input–bounded-output frameworks [66], [229]. Assumption 5.1.2 allows for a residual-based design and threshold tuning without requiring exact disturbance statistics, which is consistent with realistic CPS conditions.

To expose the measurable output dynamics explicitly, we introduce a coordinate transformation. Let $T^{(i)} \in \mathbb{R}^{n_i \times n_i}$ be a nonsingular matrix of the form:

$$T^{(i)} = \begin{bmatrix} S^{(i)} \\ H^i \end{bmatrix},$$

where the rows of $S^{(i)}$ are chosen to complete those of H^i, making $T^{(i)}$ invertible. Define the transformed state $\tilde{x}^{(i)} := T^{(i)}x^{(i)}$, partitioned as

$$\tilde{\boldsymbol{x}}^{(i)} = \begin{bmatrix} \boldsymbol{x}_1^{(i)} \\ \boldsymbol{y}_c^{(i)} \end{bmatrix},$$

where $\boldsymbol{y}_c^{(i)} := H^i\boldsymbol{x}^{(i)}$ is the noise-free component of the output. This yields the transformed dynamics:

$$\begin{bmatrix} \dot{\boldsymbol{x}}_1^{(i)} \\ \dot{\boldsymbol{y}}_c^{(i)} \end{bmatrix} = \mathcal{F}^{(i,i)} \begin{bmatrix} \boldsymbol{x}_1^{(i)} \\ \boldsymbol{y}_c^{(i)} \end{bmatrix} + \sum_{j \in \mathcal{N}_i^{\text{in}}} \mathcal{F}^{(i,j)} \begin{bmatrix} \boldsymbol{x}_1^{(j)} \\ \boldsymbol{y}^{(j)} \end{bmatrix} + \mathcal{B}^{(i)}\boldsymbol{u}^{(i)} + \mathcal{D}^{(i)}\boldsymbol{w}^{(i)}, \tag{5.2}$$

where $\mathcal{F}^{(i,i)} := T^{(i)}F^{(i,i)}(T^{(i)})^{-1}$, $\mathcal{F}^{(i,j)} := T^{(i)}F^{(i,j)}(T^{(j)})^{-1}$, and $\mathcal{B}^{(i)} = T^{(i)}B^{(i)}$, $\mathcal{D}^{(i)} = T^{(i)}D^{(i)}$. The output is then:

$$\boldsymbol{y}^{(i)} = \begin{bmatrix} 0 & I \end{bmatrix} \begin{bmatrix} \boldsymbol{x}_1^{(i)} \\ \boldsymbol{y}_c^{(i)} \end{bmatrix} + \boldsymbol{d}^{(i)}(t).$$

This structure enables the design of sliding-mode injection terms directly on measurable outputs, which is central to the methodology we develop.

5.1.2 Sensor Attack Model

We now describe the structure of the sensor-side corruption $\boldsymbol{d}^{(i)}(t)$ in (5.1b), decomposed as:

$$\boldsymbol{d}^{(i)}(t) = \boldsymbol{v}^{(i)}(t) + \boldsymbol{a}^{(i)}(t), \tag{5.3}$$

where $\boldsymbol{v}^{(i)}(t)$ denotes benign noise and $\boldsymbol{a}^{(i)}(t)$ represents an adversarial injection.

The malicious input is modeled as:

$$\boldsymbol{a}^{(i)}(t) = E^{(i)}\boldsymbol{\delta}^{(i)}(t), \tag{5.4}$$

where $\boldsymbol{\delta}^{(i)}(t) \in \mathbb{R}^{r_i^{\mathrm{a}}}$ is a latent attack signal and $E^{(i)} \in \mathbb{R}^{p_i \times r_i^{\mathrm{a}}}$ maps it into the measurement space. Typically, $E^{(i)}$ selects a known subset of vulnerable sensor channels. A canonical structure is:

$$E^{(i)} = \begin{bmatrix} 0 \\ I_{r_i^{\mathrm{a}}} \end{bmatrix},$$

where the nonzero rows correspond to compromised outputs. This reflects scenarios where only a subset of sensors is attackable due to physical exposure, communication vulnerabilities, or unbalanced security provisioning [66], [229], [230].

Attackers are typically resource-constrained and capable of targeting only a limited number of sensors at a time. Furthermore, the vulnerable sensor set can often be identified based on architectural knowledge, vulnerability assessments, or historical attacks [10], [231]. This supports the use of a known and fixed $E^{(i)}$ in observer design.

The latent input $\boldsymbol{\delta}^{(i)}(t)$ is assumed to be activated at some unknown time $T_a^{(i)}$, with bounded magnitude and rate:

$$\delta^{(i)}(t) = \boldsymbol{\psi}^{(i)}(t)\,\mathbb{1}(t - T_a^{(i)}), \tag{5.5}$$

where $\boldsymbol{\psi}^{(i)}(\cdot)$ is bounded, and $\mathbb{1}(\cdot)$ denotes the unit step function.

Assumption 5.1.3

The attack signal and its derivative satisfy:

$$\|\boldsymbol{\delta}^{(i)}(t)\| \le \bar{\delta}^{(i)}, \quad \|\dot{\boldsymbol{\delta}}^{(i)}(t)\| \le \dot{\bar{\boldsymbol{\delta}}}^{(i)},$$

for all $t \ge 0$, with unknown finite bounds. □

The general model (5.4) captures several types of sensor attacks that have been widely studied in the literature [10], [230], [231]. We provide formal expressions for three canonical attack classes:

- *False Data Injection :* The attacker injects an arbitrary but smooth or piecewise-constant signal into the vulnerable measurement channels. This corresponds to:

  $$\boldsymbol{\delta}^{(i)}(t) = \boldsymbol{\psi}^{(i)}(t), \quad \text{with } \boldsymbol{\psi}^{(i)}(t) \in \mathcal{C}^1 \text{ or piecewise constant.}$$

  In this case, the attack is persistent and can introduce a constant or slowly varying bias.

- *Replay Attacks:* A previously recorded sequence of valid outputs is replayed to mask the true behavior of the system. The injected signal replicates past measurements:

  $$\boldsymbol{\delta}^{(i)}(t) = (E^{(i)})^{\dagger}\left(\boldsymbol{y}^{(i)}(t - \tau_i) - H^i\boldsymbol{x}^{(i)}(t) - \boldsymbol{v}^{(i)}(t)\right), \quad \tau_i > 0,$$

  where $(E^{(i)})^{\dagger}$ denotes the Moore–Penrose pseudoinverse, assuming the attack channels are known. This simulates legitimate data and may evade standard anomaly detectors.

- *Denial-of-Service (DoS):* The attacker forces the sensor to output a constant, typically the last known value or a fixed spoofed reading. This yields:

  $$\boldsymbol{\delta}^{(i)}(t) = (E^{(i)})^{\dagger}\left(\boldsymbol{y}_{\mathrm{DoS}}^{(i)} - H^i\boldsymbol{x}^{(i)}(t) - \boldsymbol{v}^{(i)}(t)\right),$$

  where $y_{\mathrm{DoS}}^{(i)} \in \mathbb{R}^{p_i}$ is the constant output enforced during the attack. In this case, the output becomes uninformative or misleading.

Each attack type above corresponds to a specific structure of the latent signal $\boldsymbol{\delta}^{(i)}(t)$, which our observer framework is designed to estimate accurately despite disturbances and unknown activation times.

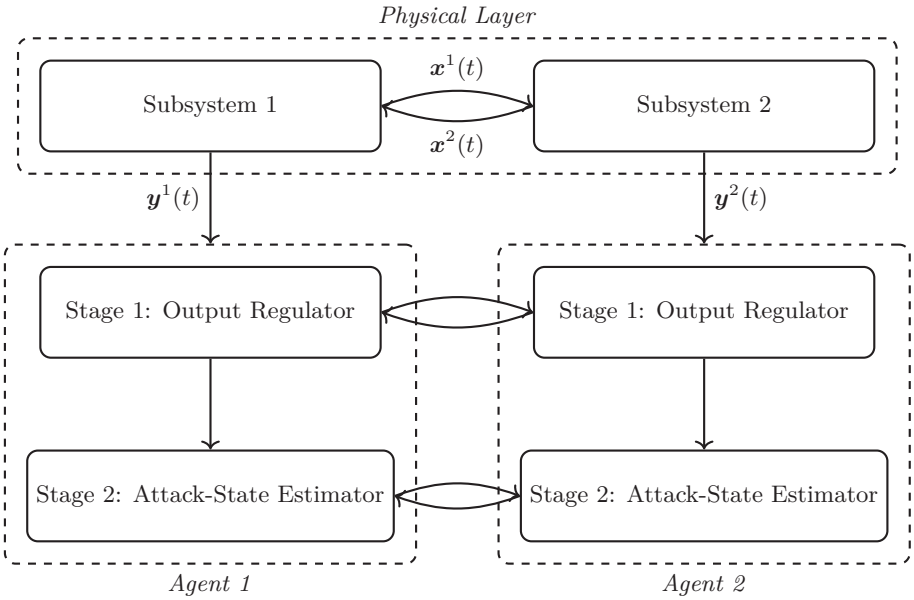


Figure 5.1: Modular observer architecture. Each agent processes local measurements through a two-stage pipeline composed of a sliding-mode output regulator and a state–attack observer. Arrows indicate the communication links between modules, while the specific exchanged signals are defined in the subsequent sections.

5.1.3 Problem formulation

The goal of this work is to design a distributed estimation architecture capable of reconstructing the unknown attack signal $\boldsymbol{\delta}^{(i)}(t)$ at each agent, with minimal assumptions on global model knowledge or centralized coordination.

Problem 5.1.1

Given the system dynamics (5.1)*, the attack model* (5.4)–(5.5)*, and Assumptions 5.1.1–5.1.3, design a fully distributed observer such that each agent* $i \in \mathcal{V}$*:*

i) *Accurately estimates the unknown signal* $\delta^{(i)}(t)$ *affecting its measurements;*

ii) *Guarantees that the estimation error* $\hat{\boldsymbol{\delta}}^{(i)}(t) - \boldsymbol{\delta}^{(i)}(t)$ *converges to a known bounded set in finite time;*

iii) *Operates using only its own measurements and information received from in-neighbors, without access to centralized or global information.*

□

This formulation captures the challenge of reconstructing sensor attack signals in a scalable, robust, and decentralized manner—key requirements for modern CPS protection in adversarial environments.

5.2 Modular Distributed Observer Architecture

This section introduces the proposed distributed observer framework for resilient state reconstruction and attack detection in interconnected cyber–physical systems. The design is built on a modular two-layer structure that combines disturbance rejection with the estimation of latent adversarial inputs. Each agent runs a local observer that communicates with its neighbors, enabling accurate and robust reconstruction in a fully distributed manner [232].

5.2.1 Functional Overview

The architecture follows a two-stage design philosophy, which separates the estimation task into two complementary layers. This decomposition allows each module to focus on a distinct objective—robust regulation and state–attack estimation—thereby simplifying the design and improving interpretability.

Stage 1 — Robust Output Regulation via Sliding Modes: The first layer implements a sliding-mode output regulator [233] that attenuates the effect of unknown disturbances and adversarial inputs. It drives the output estimation error to a small neighborhood of zero in finite time, and in doing so, produces an auxiliary signal—the so-called *equivalent injection*—which acts as a filtered representation of the total disturbance affecting the output channel.

Stage 2 — State and Attack Estimation: The second layer leverages the equivalent injection as a proxy measurement to infer both the plant state and the unknown attack signal. This is achieved via an augmented observer that models the attack input as an additional state variable. The observer is fully distributed: it processes only local measurements and exchanges information with neighboring agents [10].

This modular organization separates robustness enforcement (Stage 1) from state reconstruction (Stage 2), allowing each layer to be independently designed and systematically integrated. The result is a flexible estimation framework applicable to heterogeneous networks with limited sensing and communication capabilities [232].

5.2.2 Architecture Diagram and Information Flow

Figure 5.1 offers a schematic representation of the observer architecture in a simplified setting with two agents (Agent 1 and Agent 2) connected via a bidirectional communication link. Although minimal, this configuration captures the essential principles of the design, which naturally generalizes to larger networks with arbitrary topologies.

Each agent operates a two-layer estimation pipeline. The vertical flow between layers enables local integration of robustness and estimation, while horizontal communication between neighboring agents supports distributed cooperation.

The diagram highlights *communication capabilities* rather than fixed signal types. The actual content of the exchanged messages—be it estimates, innovation terms, or internal residuals—will be formally defined in the subsequent sections. This level of abstraction emphasizes the modularity and scalability of the framework.

5.2.3 Observer Decomposition

The overall observer structure is intentionally modular, with each layer solving a specific estimation subproblem and exposing a clear interface to the next. This design choice enhances reusability, facilitates tuning, and supports layered deployment.

In the first stage, the agent suppresses the influence of unknown disturbances by enforcing a sliding condition on the output error [233]. Once sliding is established, the equivalent injection provides a disturbance-consistent signal that aggregates the effect of modeling errors, external noise, and possible attacks.

The second stage uses this signal to reconstruct an extended state that includes both the true plant state and the attack component. This reconstruction relies on dynamic consistency relations derived from the quasi-sliding regime and operates with partial knowledge of the network state through neighbor communication [10], [232].

This layered decomposition enables a clean separation between robust error regulation and latent signal estimation. The next two sections develop each stage in detail, including controller design, observability analysis, and convergence guarantees under bounded perturbations.

5.3 Level 1 – Sliding-Mode Output Regulator

This section develops the Level 1 module, which enforces a sliding condition on the output error to generate a disturbance-consistent signal—referred to as the *equivalent injection*. This signal does not aim to reconstruct the full system state but rather provides an accurate estimate of the aggregated disturbance affecting the measured output. This estimate will serve as internal input for the Level 2 module, which handles joint state and attack reconstruction.

5.3.1 Output Regulator Design

Consider the agent dynamics expressed in the transformed coordinates (5.2). For each agent $i \in \mathcal{V}$, let $\hat{\boldsymbol{x}}^{(i)} \in \mathbb{R}^{n^{(i)}}$ denote the internal signal generated by the local correction dynamics in Level 1. Due to the specific block structure of the transformation, this signal can be naturally partitioned as

$$\hat{\boldsymbol{x}}^{(i)} := \mathrm{col}(\hat{\boldsymbol{x}}_1^{(i)}, \hat{\boldsymbol{y}}^{(i)}),$$

where $\hat{\boldsymbol{x}}_1^{(i)} \in \mathbb{R}^{n^{(i)}-p^{(i)}}$ represents the estimate of the internal state component not directly measured, while $\hat{\boldsymbol{y}}^{(i)} \in \mathbb{R}^{p^{(i)}}$ corresponds to a regulated approximation of the plant output.

To quantify the discrepancy between the measured and regulated outputs, we define the output error

$$\boldsymbol{e}_y^{(i)} := \boldsymbol{y}^{(i)} - \hat{\boldsymbol{y}}^{(i)},$$

which reflects the mismatch between the actual sensor reading $\boldsymbol{y}^{(i)}$ and the corresponding reconstructed signal $\hat{\boldsymbol{y}}^{(i)}$. This quantity serves as the driving signal for the regulation strategy introduced in this layer. Unlike standard state observers, the goal of this module is not to recover the full system state, but to asymptotically suppress the output error $\boldsymbol{e}_y^{(i)}$, thereby enabling the extraction of a disturbance-consistent injection signal, to be used in a later stage.

To achieve this objective, we design a dynamic regulator that asymptotically enforces a sliding condition on the output error. The regulation dynamics are defined as follows:

$$\dot{\hat{\boldsymbol{x}}}^{(i)} = \mathcal{F}^{(i,i)}\hat{\boldsymbol{x}}^{(i)} - \mathcal{L}^{(i)}\boldsymbol{e}_y^{(i)} - \sum_{j\in\mathcal{N}_i^{\mathrm{in}}} \mathcal{L}^{(i,j)}\boldsymbol{e}_y^{(j)} + \mathcal{F}^{(-i)}\hat{\boldsymbol{x}}^{(-i)} + \mathcal{B}^{(i)}\boldsymbol{u}^{(i)} + \mathcal{K}^{(i)}(t)\boldsymbol{\nu}^{(i)}, \tag{5.6}$$

where the vector $\hat{\boldsymbol{x}}^{(-i)} := \mathrm{col}(\hat{\boldsymbol{x}}^{(j)})_{j\in\mathcal{N}_i^{\mathrm{in}}}$ collects the correction signals received from agent i's neighbors. The correction matrices in (5.6) are defined as:

$$\mathcal{L}^{(i)} := \begin{bmatrix} -\mathcal{F}_{12}^{(i)} \\ -(\mathcal{F}_{22}^{(i)} - \mathcal{F}_{22}^{\mathrm{des},i}) \end{bmatrix}, \qquad \mathcal{L}^{(i,j)} := \begin{bmatrix} -\mathcal{F}_{12}^{(i,j)} \\ -\mathcal{F}_{22}^{(i,j)} \end{bmatrix},$$

$$\mathcal{K}^{(i)}(t) := \begin{bmatrix} 0 \\ K^{(i)}(t) \end{bmatrix}.$$

The last term in (5.6) introduces a corrective action that pushes the error signal $\boldsymbol{e}_y^{(i)}$ toward zero using a sliding-mode strategy. The direction of this injection is given by a smooth saturation function,

$$\boldsymbol{\nu}^{(i)} := \mathrm{sat}_{\phi^{(i)}}(\boldsymbol{e}_y^{(i)}) := \frac{\boldsymbol{e}_y^{(i)}}{\max\{\phi^{(i)}, \|\boldsymbol{e}_y^{(i)}\|\}},$$

where $\phi^{(i)} > 0$ defines the width of a boundary layer within which chattering is suppressed. This formulation preserves robustness to measurement noise while avoiding the discontinuities associated with standard sliding-mode control laws.

To enforce convergence of the output error within the sliding layer

$$\mathcal{S}_\phi := \left\{ \mathrm{col}(\boldsymbol{e}_y^1, \ldots, \boldsymbol{e}_y^N) : \|\boldsymbol{e}_y^{(i)}\| \le \phi^{(i)},\ \forall i \in \mathcal{V} \right\},$$

the scalar gain $K^{(i)}(t)$ is adapted online according to the following rule:

$$\dot{K}^{(i)}(t) = \begin{cases} \gamma_i, & \text{if } \|\boldsymbol{e}_y^{(i)}(t)\| > \phi^{(i)}, \\ -\lambda_i \left(K^{(i)}(t) - K_b^{(i)}(t) \right), & \text{if } \|\boldsymbol{e}_y^{(i)}(t)\| \le \phi^{(i)}, \end{cases} \tag{5.7}$$

where $\gamma_i > 0$ controls the growth rate outside the boundary layer, and $\lambda_i \in (0, 1)$ governs the decay rate once regulation is achieved. To retain robustness against time-varying and possibly discontinuous disturbances, the gain is not allowed to fall below a time-varying adaptive floor $K_b^{(i)}(t)$, which evolves as

$$K_b^{(i)}(t^+) = \begin{cases} \max\{K_b^{(i)}(t), K^{(i)}(t)\}, & \text{if } \|\boldsymbol{e}_y^{(i)}(t)\| = \phi^{(i)}, \\ K_b^{(i)}(t), & \text{otherwise.} \end{cases} \tag{5.8}$$

Adaptive gain tuning strategies of this form are often employed to ensure robustness against unknown and time-varying disturbances, avoiding conservative overdesign [234].

The overall mechanism allows the regulating gain to increase when needed and stabilize once sufficient corrective action is achieved, while memorizing the minimal effort required to preserve sliding. This modular correction law ensures that the system can respond to unexpected disturbances or attack events without relying on conservative bounds, while maintaining the output error within a predefined invariant region. The corrective signal $\mathcal{K}^{(i)}(t)\boldsymbol{\nu}^{(i)}$ represents the raw discontinuity injection that enforces the output sliding condition. To obtain a smooth and disturbance-consistent estimate of the total unknown input acting on the output channel, this signal is post-processed through a low-pass filter. The resulting filtered signal is referred to as the *equivalent injection*, and will serve as the key input to the subsequent estimation stage responsible for joint attack and state reconstruction. This concept originates from the theory of sliding-mode observers, where the equivalent injection reflects the effect of unknown inputs on the output channel [235], [236]

5.3.2 Error Dynamics

We now characterize the internal behavior of the proposed output regulation scheme by analyzing the dynamics of the estimation error. Define the total estimation error for agent i as

$$\tilde{\boldsymbol{x}}^{(i)} - \hat{\boldsymbol{x}}^{(i)} = \mathrm{col}(\boldsymbol{e}_1^{(i)}, \boldsymbol{e}_y^{(i)}),$$

where $\boldsymbol{e}_1^{(i)}$ denotes the unmeasured state estimation error and $\boldsymbol{e}_y^{(i)}$ is the output prediction error introduced earlier.

The dynamics of $(\boldsymbol{e}_1^{(i)}, \boldsymbol{e}_y^{(i)})$ can be derived by subtracting (5.6) from the transformed system equations (5.2). This leads to the following block lower-triangular structure:

$$\begin{aligned} \begin{bmatrix} \dot{\boldsymbol{e}}_1^{(i)} \\ \dot{\boldsymbol{e}}_y^{(i)} \end{bmatrix} &= \begin{bmatrix} \mathcal{F}_{11}^{(i)} & 0 \\ \mathcal{F}_{21}^{(i)} & \mathcal{F}_{22}^{\mathrm{des},i} \end{bmatrix} \begin{bmatrix} \boldsymbol{e}_1^{(i)} \\ \boldsymbol{e}_y^{(i)} \end{bmatrix} + \sum_{j \in \mathcal{N}_i^{\mathrm{in}}} \begin{bmatrix} \mathcal{F}_{11}^{(i,j)} & 0 \\ \mathcal{F}_{21}^{(i,j)} & 0 \end{bmatrix} \begin{bmatrix} \boldsymbol{e}_1^{(j)} \\ \boldsymbol{e}_y^{(j)} \end{bmatrix} \\ &\quad - \mathcal{K}^{(i)}(t)\boldsymbol{\nu}^{(i)} + \begin{bmatrix} 0 \\ I \end{bmatrix} \dot{\boldsymbol{d}}^{(i)} - \begin{bmatrix} \mathcal{F}_{12}^{(i)} \\ \mathcal{F}_{22}^{(i)} \end{bmatrix} \boldsymbol{d}^{(i)} \\ &\quad - \sum_{j \in \mathcal{N}_i^{\mathrm{in}}} \begin{bmatrix} \mathcal{F}_{12}^{(i,j)} \\ \mathcal{F}_{22}^{(i,j)} \end{bmatrix} \boldsymbol{d}^{(j)} + \begin{bmatrix} \mathcal{D}_1^{(i)} \\ \mathcal{D}^{(i)} \end{bmatrix} \boldsymbol{w}^{(i)}. \end{aligned} \tag{5.9}$$

The absence of terms involving $\boldsymbol{e}_y^{(i)}$ in the $\dot{\boldsymbol{e}}_1^{(i)}$ equation highlights the triangular decoupling. This hierarchical separation is common in sliding-mode observer design [233],

[236], enabling modular stability analysis. We therefore study the evolution of the internal error dynamics $\dot{\boldsymbol{e}}_1^{(i)}$, and show that—regardless of the behavior of the measurable components—this unmeasured part converges to a compact set whose size depends only on the magnitude of bounded exogenous inputs and the network topology.

Proposition 5.3.1

Let Assumptions 5.1.1–5.1.3 hold, and consider the error dynamics in (5.9). Suppose that for each agent $i \in \mathcal{V}$, there exist a positive definite matrix $P_1^{(i)} \succ 0$ and a scalar $\alpha_i > 0$ such that

$$(\mathcal{F}_{11}^{(i)})^\top P_1^{(i)} + P_1^{(i)} \mathcal{F}_{11}^{(i)} \preceq -I, \tag{5.10a}$$

$$\alpha_i > \|P_1^{(i)}\|, \tag{5.10b}$$

$$Q(\alpha)^\top + Q(\alpha) \succ 0, \tag{5.10c}$$

where $Q(\alpha) \in \mathbb{R}^{N\times N}$ is defined element-wise by

$$[Q(\alpha)]_{ij} = \begin{cases} 1, & \text{if } i = j, \\ -2\alpha_i \|\mathcal{F}_{11}^{(i,j)}\|, & \text{if } j \in \mathcal{N}_i^{\text{in}}, \\ 0, & \text{otherwise.} \end{cases}$$

Define the vector $\mathbf{r}(\alpha) \in \mathbb{R}^N_{>0}$ as

$$\mathbf{r}(\alpha) := \operatorname{col}(r_1, \dots, r_N),$$

where the i-th component is

$$r_i := 2\alpha_i \left(\|\mathcal{D}_1^{(i)}\| \bar{w}^{(i)} + \|\mathcal{F}_{12}^{(i)}\| \bar{d}^{(i)} + \sum_{j\in\mathcal{N}_i^{\text{in}}} \|\mathcal{F}_{12}^{(i,j)}\| \bar{d}^{(j)} \right).$$

Then the stacked error vector $e_1 = \operatorname{col}(\boldsymbol{e}_1^1, \dots, \boldsymbol{e}_1^N) \in \mathbb{R}^{n_1+\cdots+n_N}$ is ultimately bounded. Specifically, defining the vector of componentwise norms

$$\bar{\boldsymbol{e}} := \operatorname{col}\left(\|\boldsymbol{e}_1^1\|, \dots, \|\boldsymbol{e}_1^N\|\right) \in \mathbb{R}^N,$$

the trajectories of $\boldsymbol{e}_1(t)$ remain eventually confined within the compact set

$$\mathcal{X} := \left\{ \bar{\boldsymbol{e}} \in \mathbb{R}^N \;\middle|\; \bar{\boldsymbol{e}} \le \boldsymbol{\rho} \right\},$$

where $\boldsymbol{\rho} := 2\left(Q^\top(\alpha) + Q(\alpha)\right)^{-1} \mathbf{r}(\alpha)$. □

Proof. Define a composite Lyapunov function as

$$V := \sum_{i\in\mathcal{V}} (\boldsymbol{e}_1^{(i)})^\top P_1^{(i)} \boldsymbol{e}_1^{(i)}.$$

Differentiating V along the trajectories of (5.19) and exploiting (5.10a), yields

$$\dot{V} \le -\sum_i \|\boldsymbol{e}_1^{(i)}\|^2 + 2\sum_i \sum_{j\in\mathcal{N}_i^{\text{in}}} \alpha_i \|\mathcal{F}_{11}^{(i,j)}\| \|\boldsymbol{e}_1^{(i)}\| \|\boldsymbol{e}_1^{(j)}\| + 2\sum_i \alpha_i \|\tilde{d}^{(i)}\| \|\boldsymbol{e}_1^{(i)}\|,$$

where the lumped disturbance term is

$$\tilde{\boldsymbol{d}}^{(i)} := \mathcal{D}_1^{(i)} \boldsymbol{w}^{(i)} - \mathcal{F}_{12}^{(i)} \boldsymbol{d}^{(i)} - \sum_{j\in\mathcal{N}_i^{\text{in}}} \mathcal{F}_{12}^{(i,j)} \boldsymbol{d}^{(j)},$$

and satisfies the bound

$$\|\tilde{\boldsymbol{d}}^{(i)}\| \leq \frac{r_i}{2\alpha_i}.$$

Defining the vector of norms $\bar{\boldsymbol{e}} := \mathrm{col}(\|\boldsymbol{e}_1^1\|, \ldots, \|\boldsymbol{e}_1^N\|) \in \mathbb{R}^N$, and explicitly substituting the definitions of $Q(\alpha)$ and $\mathbf{r}(\alpha)$, the derivative of the Lyapunov function becomes

$$\dot{V} \leq -\tfrac{1}{2}\bar{\boldsymbol{e}}^\top \left(Q^\top(\alpha) + Q(\alpha)\right)\bar{\boldsymbol{e}} + \bar{\boldsymbol{e}}^\top \mathbf{r}(\alpha),$$

Since the matrix $Q^\top(\alpha) + Q(\alpha)$ is positive definite by assumption, the set

$$\left\{\bar{\boldsymbol{e}} \in \mathbb{R}^N \;\middle|\; \bar{\boldsymbol{e}} \leq \boldsymbol{\rho} := 2\left(Q^\top(\alpha) + Q(\alpha)\right)^{-1}\mathbf{r}(\alpha)\right\}$$

is forward invariant and attractive, establishing ultimate boundedness of the error. □

This result establishes that the unmeasured internal component $\boldsymbol{e}_1^{(i)}(t)$ remains bounded for every agent $i \in \mathcal{V}$, with an explicit bound that depends solely on known system parameters, network topology, and the magnitude of bounded exogenous inputs. Since all terms involved are either design choices or assumed to be uniformly bounded, the boundedness of $\boldsymbol{e}_1^{(i)}(t)$ is guaranteed uniformly across the entire network.

In view of this property, it is both technically sound and analytically convenient to treat $\boldsymbol{e}_1^{(i)}(t)$ as a bounded signal for all $i \in \mathcal{V}$ in the remainder of the analysis. This assumption simplifies the characterization of the dynamics governing the output estimation error , without compromising the generality or robustness of the proposed framework. We now proceed to show that, under the proposed adaptive gain law (5.7), the output error $\boldsymbol{e}_y^{(i)}(t)$ for all $i \in \mathcal{V}$ is guaranteed to reach the sliding manifold in finite time, despite the presence of bounded disturbances and modeling uncertainties.

Proposition 5.3.2

Let Assumptions 5.1.1–5.1.3 hold, and suppose the conditions of Proposition 5.3.1 are satisfied. Consider the error dynamics described in (5.9) with adaptive gain law (5.7) and fix a time interval $[t_1, t_2] \subset \mathbb{R}_{\geq 0}$. Define the aggregated disturbance signal

$$D_{net}^{(i)}(t) := \mathcal{F}_{21}^{(i)} e_1^{(i)}(t) + \sum_{j \in \mathcal{N}_i^{\text{in}}} \mathcal{F}_{21}^{(i,j)} e_1^{(j)}(t) + \dot{d}^{(i)}(t) + \mathcal{D}^{(i)} w^{(i)}(t) - \mathcal{F}_{22}^{(i)} d^{(i)}(t) - \sum_{j \in \mathcal{N}_i^{\text{in}}} \mathcal{F}_{22}^{(i,j)} d^{(j)}(t).$$

Suppose that $D_{net}^{(i)}(t)$ is bounded on $[t_1, t_2]$, and that its worst-case magnitude is attained at the initial instant t_1, that is,

$$\|D_{net}^{(i)}(t)\| \leq \bar{D}_{net}^{(i)} := \|D_{net}^{(i)}(t_1)\| < \infty, \qquad \forall\, t \in [t_1, t_2].$$

Then, for any $\varepsilon_i > 0$, the adaptive gain law (5.7) ensures that the output error enters the sliding layer in finite time. Specifically, if the time window is sufficiently long such that

$$t_2 - t_1 > T_{\min}^{(i)} := \frac{K_{crit}^{(i)} - K^{(i)}(t_1)}{\gamma_i} + \frac{\|e_y^{(i)}(t_K^{(i)})\| - \phi^{(i)}}{\varepsilon_i},$$

then there exists a time $t_s^{(i)} \in [t_1, t_2]$ such that

$$\|e_y^{(i)}(t)\| \leq \phi^{(i)}, \qquad \forall t \in [t_s^{(i)}, t_2],$$

where

$$K_{crit}^{(i)} := \lambda_{\max}^{(i)} \phi^{(i)} + \bar{D}_{net}^{(i)} + \varepsilon_i,$$

$$\lambda_{\max}^{(i)} := \lambda_{\max}\left(\tfrac{1}{2}\left(\mathcal{F}_{22}^{\text{des},i} + \mathcal{F}_{22}^{\text{des},i\top}\right)\right),$$

and

$$t_K^{(i)} := t_1 + \frac{K_{crit}^{(i)} - K^{(i)}(t_1)}{\gamma_i}.$$

□

Proof. Define the Lyapunov candidate function

$$V_i(t) := \tfrac{1}{2}\|e_y^{(i)}(t)\|^2,$$

which is positive definite and radially unbounded with respect to $e_y^{(i)}$. Differentiating $V_i(t)$ along the trajectories of the error dynamics (5.9) yields

$$\dot{V}_i(t) = (e_y^{(i)})^\top \dot{e}_y^{(i)}.$$

From the structure of the error dynamics, and using the bound $\|D_{\text{net}}^{(i)}(t)\| \le \bar{D}_{\text{net}}^{(i)}$, we obtain the estimate:

$$\dot{V}_i(t) \le \lambda_{\max}^{(i)}\|e_y^{(i)}(t)\|^2 + \bar{D}_{\text{net}}^{(i)}\|e_y^{(i)}(t)\| - K^{(i)}(t)\|e_y^{(i)}(t)\|,$$

where $\lambda_{\max}^{(i)} := \lambda_{\max}\left(\frac{1}{2}(\mathcal{F}_{22}^{\text{des},i} + \mathcal{F}_{22}^{\text{des},i\top})\right)$ denotes the maximum eigenvalue of the symmetric part of the local desired output dynamics matrix.

Now suppose that the adaptive gain $K^{(i)}(t)$ satisfies

$$K^{(i)}(t) \ge K_{\text{crit}}^{(i)} := \lambda_{\max}^{(i)}\phi^{(i)} + \bar{D}_{\text{net}}^{(i)} + \varepsilon_i,$$

for some small constant $\varepsilon_i > 0$. Then, whenever $\|e_y^{(i)}(t)\| > \phi^{(i)}$, it follows that

$$\dot{V}_i(t) \le -\varepsilon_i\|e_y^{(i)}(t)\| < 0.$$

Therefore, the energy $V_i(t)$ strictly decreases at least at rate $\varepsilon_i\|\boldsymbol{e}_y^{(i)}(t)\|$, which implies that the norm $\|e_y^{(i)}(t)\|$ decreases by at least ε_i per unit time as long as it remains outside the sliding layer.

The gain $K^{(i)}(t)$ increases at rate γ_i while $\|\boldsymbol{e}_y^{(i)}(t)\| > \phi^{(i)}$, and reaches the threshold $K_{\text{crit}}^{(i)}$ no later than time

$$t_K^{(i)} := t_1 + \frac{K_{\text{crit}}^{(i)} - K^{(i)}(t_1)}{\gamma_i}.$$

After this point, the descent rate is guaranteed. The time required to reduce the error from $\|\boldsymbol{e}_y^{(i)}(t_K^{(i)})\|$ to the boundary $\phi^{(i)}$ at rate ε_i is at most

$$\frac{\|\boldsymbol{e}_y^{(i)}(t_K^{(i)})\| - \phi^{(i)}}{\varepsilon_i}.$$

Summing both contributions yields the total convergence time bound:

$$t_s^{(i)} \le t_1 + \frac{K_{\text{crit}}^{(i)} - K^{(i)}(t_1)}{\gamma_i} + \frac{\|\boldsymbol{e}_y^{(i)}(t_K^{(i)})\| - \phi^{(i)}}{\varepsilon_i}.$$

In particular, if we consider a generic time interval $[t_1, t_2]$, and assume that the worst-case bound $\bar{D}_{\text{net}}^{(i)}$ holds throughout the interval, then convergence is guaranteed if the duration satisfies

$$t_2 - t_1 > T_{\min}^{(i)} := \frac{K_{\text{crit}}^{(i)} - K^{(i)}(t_1)}{\gamma_i} + \frac{\|\boldsymbol{e}_y^{(i)}(t_K^{(i)})\| - \phi^{(i)}}{\varepsilon_i}.$$

□

Proposition 5.3.2 ensures that the output error $\boldsymbol{e}_y^{(i)}(t)$ reaches the boundary layer $\|\boldsymbol{e}_y^{(i)}(t)\| \leq \phi^{(i)}$ in finite time, provided the adaptive gain $K^{(i)}(t)$ is allowed to increase sufficiently. Once this sliding condition is established, the corrective term $K^{(i)}(t)\boldsymbol{\nu}^{(i)}(t)$ effectively reflects the net influence of unknown disturbances acting on the output.

To extract a usable estimate of these disturbances, the signal $K^{(i)}(t)\boldsymbol{\nu}^{(i)}(t)$ is passed through a low-pass filter, yielding a smooth approximation known as the *equivalent injection*. This quantity corresponds to the classical concept introduced in [237] and later employed in observer design, such as in [233].

A key advantage of the proposed adaptive scheme is that it removes the need to predefine worst-case bounds on disturbances—an inherent limitation in conventional fixed-gain sliding-mode designs. This is especially relevant in cyber–physical systems subject to uncertain or adversarial environments, where disturbance profiles may be time-varying and unpredictable.

Moreover, even in the presence of transient deviations—such as impulsive attacks or abrupt measurement spikes—the adaptive gain guarantees re-entry into the sliding manifold, ensuring continued disturbance rejection and robust estimation under bounded uncertainties.

5.4 Level 2 – State–Attack Observer

The second layer of the estimation architecture is devoted to the reconstruction of both the true local state in the transformed coordinates $\tilde{\boldsymbol{x}}^{(i)}$ and the unknown attack signal $\delta^{(i)}$ affecting the measurement channel of agent $i \in \mathcal{V}$. While the previous stage introduced a correction mechanism based on output regulation, it is important to note that the Level 1 module does not act as a full state observer in the classical sense. Instead, its primary function is to drive the output estimation error $\boldsymbol{e}_y^{(i)}$ toward a predefined sliding manifold, thus enforcing robust behavior in the presence of bounded perturbations [233], [236].

The main object of interest produced by the Level 1 regulator is the injection signal $K^{(i)}(t)\boldsymbol{\nu}^{(i)}(t)$, whose average behavior—once sliding is achieved—converges to the so-called *equivalent injection* [233], [237]. This signal encapsulates the average correction effort required to maintain the output trajectory aligned with the measured output $y^{(i)}(t)$, effectively compensating for unknown disturbances such as sensor noise, exogenous perturbations, or adversarial attacks.

In this second stage, we leverage the equivalent injection term, together with structural knowledge of the system model, to render the attack signal observable. This enables the design of a distributed extended observer capable of jointly estimating the true plant state and the latent disturbance signal. The resulting scheme enhances situational awareness and enables real-time detection and mitigation of stealthy sensor attacks across the network.

5.4.1 Dynamic Constraint on the Sliding Layer

Once the sliding condition is enforced by the SMO, the output error signal $\boldsymbol{e}_y^{(i)}(t)$ is confined within a narrow boundary layer of radius $\phi^{(i)}$. For sufficiently small $\phi^{(i)}$, this regime approximates ideal sliding behavior, under which the following quasi-static approximations are valid:

$$\boldsymbol{e}_y^{(i)}(t) \approx 0, \qquad \dot{\boldsymbol{e}}_y^{(i)}(t) \approx 0, \qquad \dot{\boldsymbol{e}}_1^{(i)}(t) \approx 0.$$

In this setting, the discontinuous correction term $K^{(i)}(t)\boldsymbol{\nu}^{(i)}(t)$, originally introduced to maintain the sliding condition, behaves like an equivalent injection signal that captures the net effect of unknown inputs on the output channel.

To extract its meaningful low-frequency content while rejecting high-frequency chattering components, the signal is passed through a low-pass filter:

$$K^{(i)}\boldsymbol{\nu}_{\mathrm{eq}}^{(i)}(t) := \mathrm{LPF}\Big[K^{(i)}(t)\boldsymbol{\nu}^{(i)}(t)\Big],$$

a standard technique in sliding-mode observer design [233], which allows reliable reconstruction of disturbance profiles under bounded uncertainties.

Under these quasi-sliding conditions, the error dynamics derived in (5.9) simplify significantly. In particular, one can isolate the steady-state behavior of the internal error dynamics $\bar{e}_1^{(i)} := \mathrm{col}(\boldsymbol{e}_1^{(i)}, \boldsymbol{e}_1^{(-i)})$ using the structure of the transformed system matrices. Specifically, we obtain the following algebraic relations:

$$X_i\, \bar{\boldsymbol{e}}_1^{(i)} = \mathcal{F}_{12}^{(i)} \boldsymbol{d}^{(i)} + \sum_{j \in \mathcal{N}_i^{\mathrm{in}}} \mathcal{F}_{12}^{(i,j)} \boldsymbol{d}^{(j)} - \mathcal{D}_1^{(i)} \boldsymbol{w}^{(i)}, \tag{5.11}$$

$$K^{(i)} \boldsymbol{\nu}_{\mathrm{eq}}^{(i)} - \mathcal{D}_2^{(i)} \boldsymbol{w}^{(i)} = Y_i\, \bar{\boldsymbol{e}}_1^{(i)} - \mathcal{F}_{22}^{(i)} \boldsymbol{d}^{(i)} - \sum_{j \in \mathcal{N}_i^{\mathrm{in}}} \mathcal{F}_{22}^{(i,j)} \boldsymbol{d}^{(j)} + \dot{\boldsymbol{d}}^{(i)}, \tag{5.12}$$

where the composite matrices X_i and Y_i are defined as

$$X_i := \begin{bmatrix} \mathcal{F}_{11}^{(i)} & \mathcal{H}_{11}^{(i)} \end{bmatrix}, \qquad Y_i := \begin{bmatrix} \mathcal{F}_{21}^{(i)} & \mathcal{H}_{21}^{(i)} \end{bmatrix}.$$

The following assumption is needed to proceed with the elimination of $\bar{\boldsymbol{e}}_1^{(i)}$ from the system.

Assumption 5.4.1

For each $i \in \mathcal{V}$, the matrix $Y_i \in \mathbb{R}^{p_i \times m_i}$, where $m_i := \dim(\bar{\boldsymbol{e}}_1^{(i)})$, has full column rank. That is, $\mathrm{rank}(Y_i) = m_i$. □

Under this assumption, the matrix Y_i admits a left pseudoinverse $Y_i^{\dagger} \in \mathbb{R}^{m_i \times p_i}$, which can be used to express $\bar{\boldsymbol{e}}_1^{(i)}$ in terms of $K^{(i)} \boldsymbol{\nu}_{\mathrm{eq}}^{(i)}$, $\boldsymbol{d}^{(i)}$, $\boldsymbol{w}^{(i)}$, and neighbor signals. Specifically, by pre-multiplying both sides of (5.12) by $X_i Y_i^{\dagger}$ and comparing with (5.11), we can eliminate $\bar{e}_1^{(i)}$ and obtain the following dynamic constraint involving only known or estimable quantities:

$$M_i\, \dot{\boldsymbol{d}}^{(i)} = M_i K^{(i)} \boldsymbol{\nu}_{\mathrm{eq}}^{(i)} + \Theta_H^i \boldsymbol{d}^{(i)} + \sum_{j \in \mathcal{N}_i^{\mathrm{in}}} \Theta_N^{(i,j)} \boldsymbol{d}^{(j)} + \Theta_w^{(i)} \boldsymbol{w}^{(i)}, \tag{5.13}$$

where the matrices in the expression are defined as follows:

$$M_i := X_i Y_i^{\dagger}, \qquad \Theta_H^i := M_i \mathcal{F}_{22}^{(i)} - \mathcal{F}_{12}^{(i)},$$

$$\Theta_N^{(i,j)} := M_i \mathcal{F}_{22}^{(i,j)} - \mathcal{F}_{12}^{(i,j)}, \qquad \Theta_w^{(i)} := \mathcal{D}_1^{(i)} - M_i \mathcal{D}_2^{(i)}.$$

Equation (5.13) provides a key insight into the structure of the disturbance signal $\boldsymbol{d}^{(i)}(t)$. It shows that, on the sliding manifold, the derivative $\dot{\boldsymbol{d}}^{(i)}(t)$ is linearly related to the equivalent injection $\boldsymbol{\nu}_{\mathrm{eq}}^{(i)}(t)$, the current attack signals $\boldsymbol{d}^{(j)}(t)$, and the disturbance $\boldsymbol{w}^{(i)}(t)$. This relation will be exploited in the next section to construct an observer capable of estimating both $\boldsymbol{d}^{(i)}(t)$ and $\boldsymbol{x}^{(i)}(t)$.

5.4.2 Attack–State Coupling Dynamics

We now exploit the structural model of the disturbance signal acting on each agent. Recall that the unknown signal $\boldsymbol{d}^{(i)}$ entering the output channel of agent $i \in \mathcal{V}$ can be decomposed as

$$\boldsymbol{d}^{(i)} = E_i \boldsymbol{\delta}^{(i)} + \boldsymbol{v}^{(i)}, \qquad \dot{\boldsymbol{d}}^{(i)} = E_i \dot{\boldsymbol{\delta}}^{(i)} + \dot{\boldsymbol{v}}^{(i)},$$

where $\delta^{(i)} \in \mathbb{R}^{r_i^{\mathrm{a}}}$ is an unknown attack input evolving along the known columns of the attack signature matrix $E_i \in \mathbb{R}^{p^{(i)} \times r_i^{\mathrm{a}}}$, and $\boldsymbol{v}^{(i)}$ denotes an unstructured residual disturbance.

Substituting this decomposition into the dynamic constraint (5.13) yields:

$$G_i\,\dot{\boldsymbol{\delta}}^{(i)} = M_i K^{(i)} \boldsymbol{\nu}_{\mathrm{eq}}^{(i)} + \Theta_H^i (E_i \boldsymbol{\delta}^{(i)} + \boldsymbol{v}^{(i)}) + \sum_{j \in \mathcal{N}_i^{\mathrm{in}}} \Theta_N^{(i,j)} (E_j \boldsymbol{\delta}^{(j)} + \boldsymbol{v}^{(j)}) + \Theta_w^{(i)} \boldsymbol{w}^{(i)} - M_i \dot{\boldsymbol{v}}^{(i)}, \tag{5.14}$$

where the matrix $G_i := M_i E_i \in \mathbb{R}^{p^{(i)} \times r_i^{\mathrm{a}}}$ plays the role of a local regressor relating the attack derivative $\dot{\delta}^{(i)}$ to the observable quantities in the system.

To ensure well-posedness of the attack dynamics, we impose the following identifiability condition:

Assumption 5.4.2

For each $i \in \mathcal{V}$, the matrix $G_i = M_i E_i \in \mathbb{R}^{p^{(i)} \times r_i^{\mathrm{a}}}$ has full column rank, i.e., $\mathrm{rank}(G_i) = r_i^{\mathrm{a}}$. □

This condition guarantees that the attack components $\boldsymbol{\delta}^{(i)}(t)$ lie in a known and observable subspace and can be decoupled from the dynamics.

Premultiplying both sides of (5.14) by a left pseudoinverse $G_i^{\dagger} \in \mathbb{R}^{r_i^{\mathrm{a}} \times p^{(i)}}$, we obtain the explicit expression for the attack dynamics:

$$\dot{\delta}^{(i)} = G_i^{\dagger} \Big(M_i K^{(i)} \boldsymbol{\nu}_{\mathrm{eq}}^{(i)} + \Theta_H^i (E_i \boldsymbol{\delta}^{(i)} + \boldsymbol{v}^{(i)}) + \sum_{j \in \mathcal{N}_i^{\mathrm{in}}} \Theta_N^{(i,j)} (E_j \boldsymbol{\delta}^{(j)} + \boldsymbol{v}^{(j)}) + \Theta_w^{(i)} \boldsymbol{w}^{(i)} - M_i \dot{\boldsymbol{v}}^{(i)} \Big). \tag{5.15}$$

This equation describes the evolution of the attack signal $\boldsymbol{\delta}^{(i)}(t)$ as a function of: (i) the equivalent injection $\boldsymbol{\nu}_{\mathrm{eq}}^{(i)}(t)$, (ii) the known model matrices, (iii) the local and neighboring attack signals $\boldsymbol{\delta}^{(j)}(t)$, and (iv) both structured and unstructured disturbances.

In the next section, this model will serve as the basis for designing a distributed estimation scheme capable of reconstructing the attack signal and recovering the true state trajectory of each agent, exploiting the known structure of the attack and communication topology.

5.4.3 Detectability and Observer Synthesis for the Extended Dynamics

We now analyze the detectability and observer synthesis for the extended model obtained by combining the nominal dynamics with the attack dynamics derived in (5.15). Define the extended state as

$$\boldsymbol{z}^{(i)} := \mathrm{col}(\tilde{\boldsymbol{x}}^{(i)}, \boldsymbol{\delta}^{(i)}).$$

Then the local extended dynamics can be written as:

$$\dot{\boldsymbol{z}}^{(i)} = \mathcal{F}_z^{(i)} \boldsymbol{z}^{(i)} + \sum_{j \in \mathcal{N}_i^{\mathrm{in}}} \mathcal{H}_z^{(i,j)} \boldsymbol{z}^{(j)} + \mathcal{B}_z^{(i)} \boldsymbol{u}_z^{(i)} + \boldsymbol{\omega}^{(i)}, \qquad \boldsymbol{y}^{(i)} = C_z^{(i)} \boldsymbol{z}^{(i)} + \boldsymbol{v}^{(i)}, \tag{5.16}$$

where

$$\mathcal{F}_z^{(i)} := \mathrm{blkdiag}(\mathcal{F}^{(i,i)}, \mathcal{F}_\delta^{(i)}), \quad \mathcal{F}_\delta^{(i)} := G_i^{\dagger} \Theta_H^i E_i,$$

$$\mathcal{H}_z^{(i,j)} := \mathrm{blkdiag}(\mathcal{F}_j^{(-i)}, G_i^{\dagger} \Theta_N^{(i,j)} E_j), \quad C_z^{(i)} := [\, C_i \;\; E_i \,],$$

$$\mathcal{B}_z^{(i)} := \mathrm{blkdiag}(\mathcal{B}^{(i)}, G_i^{\dagger} M_i K^{(i)}), \quad \boldsymbol{u}_z^{(i)} := \mathrm{col}(\boldsymbol{u}^{(i)}, \boldsymbol{\nu}_{\mathrm{eq}}^{(i)}),$$

and residual perturbation term, which accounts for the model mismatch and collects the unmodeled effects from the algebraic substitution process, is given by:

$$\boldsymbol{\omega}^{(i)}(t) = \begin{bmatrix} \mathcal{D}^{(i)}\,\boldsymbol{w}^{(i)}(t) \\ G_i^{\dagger}\left(\Theta_H^i \boldsymbol{v}^{(i)}(t) + \sum_j \Theta_N^{(i,j)} \boldsymbol{v}^{(j)}(t) + \Theta_w^{(i)} \boldsymbol{w}^{(i)}(t) - M_i \dot{\boldsymbol{v}}^{(i)}(t)\right) \end{bmatrix}.$$

We now show that the augmented pair $(\mathcal{F}_z^{(i)}, C_z^{(i)})$ defining the local extended dynamics in (5.16) is detectable under mild assumption. This structural property ensures that the joint evolution of the nominal plant state and the unknown attack signal can be reconstructed from the available measurements, and forms the foundation for the design of a stable distributed observer.

Proposition 5.4.1

Consider the extended dynamics in (5.16)*. Under Assumptions 5.1.1–5.4.2, and assuming that the attack signature matrix E_i has full column rank, the pair $(\mathcal{F}_z^{(i)}, C_z^{(i)})$ is detectable for all $i \in \mathcal{V}$.* □

Proof. By Assumptions 5.4.1–5.4.2, the matrices $M_i = X_i Y_i^{\dagger}$ and $G_i = M_i E_i$ are well defined with $\operatorname{rank}(G_i) = r_i^{\mathrm{a}}$. This ensures that the attack dynamics are well-posed and that $\mathcal{F}_\delta^{(i)} = G_i^{\dagger}\Theta_H^i E_i = L_i E_i$, where $L_i := G_i^{\dagger}\Theta_H^i$.

We apply the PBH test to the extended pair. For any $s \in \mathbb{C}$, the PBH matrix reads:

$$\mathscr{P}(s) = \begin{bmatrix} sI - \mathcal{F}^{(i)} & 0 \\ 0 & sI - \mathcal{F}_\delta^{(i)} \\ C_i & E_i \end{bmatrix}.$$

The pair $(\mathcal{F}_z^{(i)}, C_z^{(i)})$ is detectable if $\operatorname{rank}(\mathscr{P}(s)) = n_i + r_i^{\mathrm{a}}$ for all s with $\Re(s) \geq 0$.

Suppose there exists s with $\Re(s) \geq 0$ and a nonzero pair (x, δ) such that:

$$(sI - \mathcal{F}^{(i)})\boldsymbol{x} = 0, \quad (sI - \mathcal{F}_\delta^{(i)})\boldsymbol{\delta} = 0, \quad C_i\boldsymbol{x} + E_i\boldsymbol{\delta} = 0. \tag{$*$}$$

If $\delta \neq 0$, the second condition implies $E_i\boldsymbol{\delta} = 0 \Rightarrow \mathcal{F}_\delta^{(i)}\boldsymbol{\delta} = 0$. Thus, $\delta \in \ker(E_i)$. Since E_i has full column rank, $\ker(E_i) = \{0\}$, which contradicts $\delta \neq 0$. Hence, $(\mathcal{F}_\delta^{(i)}, E_i)$ is observable.

If $\delta = 0$, then condition $(*)$ reduces to $(sI - \mathcal{F}^{(i,i)})\boldsymbol{x} = 0$, $C_i\boldsymbol{x} = 0$. Assumption 5.1.1 ensures that $(\mathcal{F}^{(i,i)}, C_i)$ is detectable, hence all unobservable modes are stable, and $(\boldsymbol{x}, \boldsymbol{\delta}) = (0, 0)$ is the only solution for $\Re(s) \geq 0$. Therefore, $\operatorname{rank}(\mathscr{P}(s)) = n_i + r_i^{\mathrm{a}}$, and the pair is detectable. □

The detectability of the extended pair $(\mathcal{F}_z^{(i)}, C_z^{(i)})$, established in Proposition 5.4.1, guarantees that it is structurally feasible to reconstruct the full extended state $\boldsymbol{z}^{(i)} = \operatorname{col}(\tilde{\boldsymbol{x}}^{(i)}, \boldsymbol{\delta}^{(i)})$, comprising both the plant estimation error and the unknown attack signal, from local output measurements $\boldsymbol{y}^{(i)}$ and known inputs.

Based on this structural property, we now synthesize a distributed observer for the extended dynamics (5.16), aimed at estimating $\boldsymbol{z}^{(i)}$ for each agent $i \in \mathcal{V}$. The proposed observer is distributed in nature, accounting for the coupling between neighboring agents through the interconnection terms $\mathcal{H}_z^{(i,j)}$, and takes the form:

$$\dot{\hat{\boldsymbol{z}}}^{(i)} = \mathcal{F}_z^{(i)}\,\hat{\boldsymbol{z}}^{(i)} + \sum_{j \in \mathcal{N}_i^{\mathrm{in}}} \mathcal{H}_z^{(i,j)}\,\hat{\boldsymbol{z}}^{(j)} + \mathcal{B}_z^{(i)}\,\boldsymbol{u}_z^{(i)} + K_z^{(i)}\,\boldsymbol{r}_z^{(i)} + \sum_{j \in \mathcal{N}_i^{\mathrm{in}}} K_z^{(i,j)}\,\boldsymbol{r}_z^{(j)}, \tag{5.17}$$

where the innovation signal is defined as

$$\boldsymbol{r}_z^{(i)} := \boldsymbol{y}^{(i)} - C_z^{(i)}\,\hat{\boldsymbol{z}}^{(i)}. \tag{5.18}$$

Letting $\boldsymbol{e}_z^{(i)} := \boldsymbol{z}^{(i)} - \hat{\boldsymbol{z}}^{(i)}$ denote the local estimation error, the corresponding error dynamics are given by:

$$\begin{aligned} \dot{\boldsymbol{e}}_z^{(i)} = {} & (\mathcal{F}_z^{(i)} - K_z^{(i)} C_z^{(i)})\,\boldsymbol{e}_z^{(i)} + \sum_{j \in \mathcal{N}_i^{\mathrm{in}}} (\mathcal{H}_z^{(i,j)} - K_z^{(i,j)} C_z^{(j)})\,\boldsymbol{e}_z^{(j)} \\ & + \boldsymbol{\omega}^{(i)} - K_z^{(i)}\boldsymbol{v}^{(i)} - \sum_{j \in \mathcal{N}_i^{\mathrm{in}}} K_z^{(i,j)}\boldsymbol{v}^{(j)}. \end{aligned} \tag{5.19}$$

The next result establishes sufficient conditions ensuring both exponential convergence of the distributed observer in the nominal (noise-free) case, and ultimate boundedness of the estimation error in the presence of bounded disturbances and measurement noise.

Proposition 5.4.2

Consider the extended model (5.16) *and the observer* (5.17)*, under Assumptions 5.1.1–5.4.2. For each agent $i \in \mathcal{V}$, define the local closed-loop and coupling matrices as*

$$F_i^{\mathrm{cl}} := \mathcal{F}_z^{(i)} - K_z^{(i)} H_z^{(i)}, \qquad \Delta_{ij} := \mathcal{F}_z^{(i,j)} - K_z^{(i,j)} H_z^{(j)},$$

and assume the existence of $P_i \succ 0$, $\eta_i > 0$ such that

$$(F_i^{\mathrm{cl}})^\top P_i + P_i F_i^{\mathrm{cl}} \preceq -\eta_i I. \tag{5.20}$$

Define the interconnection gains and disturbance scaling as

$$\gamma_{ij} := \frac{\|P_i\|\|\Delta_{ij}\|}{\eta_i}, \qquad [\Gamma]_{ij} := \begin{cases} \gamma_{ij}, & j \in \mathcal{N}_i^{\mathrm{in}}, \\ 0, & \text{otherwise}, \end{cases}$$

$$\bar{\ell}_i := \sup_t \left(\|\omega^{(i)}(t)\| + \|K_z^{(i)}\|\|v^{(i)}(t)\| + \sum_{j \in \mathcal{N}_i^{\mathrm{in}}} \|K_z^{(i,j)}\|\|v^{(j)}(t)\| \right),$$

$b_i := \frac{\|P_i\|}{\eta_i} \bar{\ell}_i$*, and introduce the diagonal matrices*

$$D := \mathrm{diag}\left(\frac{\eta_i}{2\|P_i\|} \right) \succ 0, \qquad S := \mathrm{diag}\left(\sqrt{\frac{\|P_i\|}{\lambda_{\min}(P_i)}} \right),$$

$b := \mathrm{col}(b_i)$*. If the spectral radius of the scaled interconnection matrix satisfies*

$$\rho(D^{-1}\Gamma) < 1, \tag{5.21}$$

then the distributed estimation error dynamics (5.19) *are exponentially stable. Specifically, there exist constants $c > 0$, $\lambda > 0$ such that, for all $t \geq t_0$,*

$$\|e_z(t)\| \leq c\, e^{-\lambda(t-t_0)} \|e_z(t_0)\| + \|S(I - D^{-1}\Gamma)^{-1} D^{-1} b\|, \tag{5.22}$$

and the asymptotic bound satisfies

$$\limsup_{t \to \infty} \|e_z(t)\| \leq \|S(I - D^{-1}\Gamma)^{-1} D^{-1} b\|. \tag{5.23}$$

□

Proof. Let $V_i := e_z^{i\top} P_i e_z^{(i)}$ and define the composite Lyapunov function $V := \sum_i V_i$. Along the trajectories of (5.19), using (5.20), we have

$$\dot{V}_i \leq -\eta_i \|e_z^{(i)}\|^2 + 2\|P_i\| \sum_{j \in \mathcal{N}_i^{\mathrm{in}}} \|\Delta_{ij}\|\|e_z^{(i)}\|\|e_z^{j}\| + 2\|P_i\|\|e_z^{(i)}\|\bar{\ell}_i.$$

Define $f_i := \sqrt{V_i} = \|P_i^{1/2} e_z^{(i)}\|$. By standard inequalities, it follows that

$$\dot{f}_i \leq -\frac{\eta_i}{2\|P_i\|} f_i + \sum_{j \in \mathcal{N}_i^{\mathrm{in}}} \gamma_{ij} f_j + \frac{\|P_i\|}{\eta_i} \bar{\ell}_i,$$

where $\gamma_{ij} := \frac{\|P_i\|\|\Delta_{ij}\|}{\eta_i}$. Stacking all f_i into $f := [f_1, \ldots, f_N]^\top$ yields the compact form

$$\dot{f} = (-D + \Gamma) f + b,$$

with $D := \mathrm{diag}(\eta_i/(2\|P_i\|)) \succ 0$, $\Gamma := [\gamma_{ij}] \geq 0$, and $b := [\|P_i\|\bar{\ell}_i/\eta_i]$.

Since D is diagonal and positive definite, we can write

$$F_c := -D + \Gamma = -D(I - D^{-1}\Gamma).$$

The matrix F_c is Hurwitz if all eigenvalues of $I - D^{-1}\Gamma$ have positive real parts. A sufficient condition ensuring this is

$$\rho(D^{-1}\Gamma) < 1.$$

Under this condition, the homogeneous system $\dot{f} = F_c f$ is exponentially stable, and the comparison system admits the solution

$$\|f(t)\| \leq c\, e^{-\lambda(t-t_0)}\|f(t_0)\| + \|(I - D^{-1}\Gamma)^{-1}D^{-1}b\|,$$

for some constants $c, \lambda > 0$. Since $\|e_z^{(i)}\| \leq s_i f_i$ with $s_i := \sqrt{\|P_i\|/\lambda_{\min}(P_i)}$, we obtain

$$\|e_z(t)\| \leq c\, e^{-\lambda(t-t_0)}\|e_z(t_0)\| + \|S(I - D^{-1}\Gamma)^{-1}D^{-1}b\|, \tag{5.24}$$

where $S := \mathrm{diag}(s_i)$. This establishes the exponential input-to-state stability of the distributed estimation error dynamics. □

The synthesis procedure implied by Proposition 5.4.2 provides a constructive framework for the design of a distributed observer with guaranteed stability and robustness properties. The design begins with the selection of the local observer gain $K_z^{(i)}$, which should be chosen to maximize the decay rate η_i appearing in the Lyapunov inequality (5.20). This can be achieved using classical tools such as pole placement or linear matrix inequality optimization. A larger η_i corresponds to a faster local convergence rate and enhances the robustness of each observer to both modeling errors and external disturbances.

Once the local dynamics are stabilized, attention shifts to the design of the coupling gains $K_z^{(i,j)}$, which determine the influence of neighboring agents on the local estimates. These gains should be chosen to minimize the interconnection mismatch $\|\Delta_{ij}\| = \|\mathcal{F}_z^{(i,j)} - K_z^{(i,j)}H_z^{(j)}\|$, for instance by least-squares fitting of $\mathcal{F}_z^{(i,j)}$ through the neighbor's output matrix $H_z^{(j)}$. Reducing this mismatch directly limits the propagation of estimation errors across the network, thereby improving the global resilience of the distributed estimator.

After all gains are selected, the final step is to verify the small-gain condition $\rho(D^{-1}\Gamma) < 1$, where Γ is the interconnection gain matrix defined in terms of the local quantities $\gamma_{ij} := \frac{\|P_i\|\|\Delta_{ij}\|}{\eta_i}$. This condition ensures that the global coupling between agents remains sufficiently weak relative to their local convergence rates, so that the composite matrix $-D + \Gamma$ is Hurwitz. Quantitatively, the requirement $\rho(D^{-1}\Gamma) < 1$ can be interpreted as demanding that, for each agent i,

$$\sum_{j \in \mathcal{N}_i^{\mathrm{in}}} \frac{\|P_i\|\|\Delta_{ij}\|}{\eta_i} < 1,$$

meaning that the total coupling influence received from the neighbors must be smaller than the local dissipative effect. In practice, this translates into a weak-coupling regime, where the information exchanged among agents introduces only moderate corrections to the locally stabilized dynamics. If this inequality is marginally violated, the condition can be recovered either by increasing the local decay rate η_i (via stronger feedback $K_z^{(i)}$) or by reducing the interconnection mismatch $\|\Delta_{ij}\|$ through appropriate retuning of $K_z^{(i,j)}$. Hence, the proposed synthesis method not only provides a distributed observer design procedure but also quantifies an explicit trade-off between local stability margins and inter-agent coupling strength. When this balance is achieved, the distributed estimation error dynamics remain input-to-state stable: even in the presence of bounded measurement noise and modeling uncertainties, the global estimation error remains uniformly bounded and converges exponentially to a small neighborhood of the true extended state.

5.5 Main Result

We now summarize the guarantees provided by the proposed modular observer architecture by integrating the convergence results established in Sections 5.3–5.4. The following theorem states sufficient conditions under which each agent can reliably estimate both its local plant state and the unknown attack input, using only local measurements and neighbor communication.

Theorem 5.5.1

Let Assumptions 5.1.1–5.4.2 hold. Consider the modular distributed observer composed of the sliding-mode output regulator (5.6) and the state–attack observer (5.17). Suppose:

(i) The adaptive gain law (5.7) guarantees finite-time convergence of the output error to the boundary layer, i.e., $\|\boldsymbol{e}_y^{(i)}(t)\| \le \phi^{(i)}$ for all $t \ge t_s^{(i)}$, $i \in \mathcal{V}$.

(ii) The equivalent injection is obtained as $\boldsymbol{\nu}_{\mathrm{eq}}^{(i)}(t) := \mathrm{LPF}[K^{(i)}(t)\boldsymbol{\nu}^{(i)}(t)]$, for all $t \ge t_s^{(i)}$.

(iii) The observer gains $K_z^{(i)}$, $K_z^{(i,j)}$ are such that the small-gain condition $\rho(\Gamma) < 1$ holds, with Γ defined as in Proposition 5.4.2.

Then, for each agent $i \in \mathcal{V}$, the estimation error $\boldsymbol{e}_z^{(i)}(t) := \hat{\boldsymbol{z}}^{(i)}(t) - \boldsymbol{z}^{(i)}(t)$, where $\boldsymbol{z}^{(i)}(t) := \mathrm{col}(\tilde{\boldsymbol{x}}^{(i)}(t), \boldsymbol{\delta}^{(i)}(t))$, satisfies (5.24). □

Proof. Follows by combining the results of Proposition 5.3.2 for finite-time convergence to sliding, and Proposition 5.4.2 for input-to-state stability of the distributed state–attack observer under bounded disturbances. □

Theorem 5.5.1 provides formal guarantees for robust state and attack estimation across the network. In particular, it ensures that each agent can reconstruct its local state and detect adversarial perturbations to its measurements within a bounded error margin that depends explicitly on system knowledge, design parameters, and external inputs.

Hence, in the presence of an attack on the measurement channel, the proposed scheme ensures that both the true system state and the unknown attack signal can be estimated with provable and explicitly characterizable accuracy.

5.6 Numerical Results

This section demonstrates the practical implementation and performance of the proposed modular observer architecture through a distributed simulation scenario involving adversarial sensor attacks. The objective is to validate the full estimation pipeline—including sliding-mode disturbance rejection and distributed attack reconstruction—under realistic noise, interconnection, and attack conditions.

5.6.1 Network Configuration and Setup

We consider a network of $N = 6$ agents interconnected according to the directed communication graph shown in Fig. 5.2. A directed edge from node j to node i implies that agent i receives information from agent j, i.e., $\mathcal{H}^{(i,j)} \neq 0$. The topology is sparse but connected, ensuring sufficient information propagation across the network for distributed estimation.

Each agent is modeled according to the local dynamics introduced in Section 5.1, instantiated in the canonical observer-compatible form (5.2)–(5.6). For each subsystem, the system matrix $\mathcal{F}^{(i)}$ is explicitly constructed in block form as

$$\mathcal{F}^{(i)} = \begin{bmatrix} \mathcal{F}_{11}^{(i)} & \mathcal{F}_{12}^{(i)} \\ \mathcal{F}_{21}^{(i)} & \mathcal{F}_{22}^{(i)} \end{bmatrix},$$

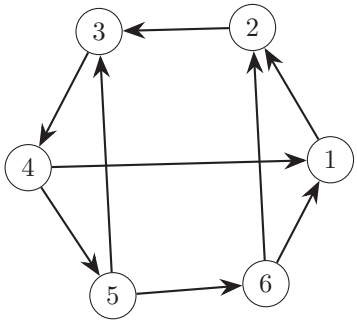


Figure 5.2: Directed interconnection topology among the six agents.

Table 5.1: Dimensions of the local subsystems.

Agent i	State dim. n_i	Output dim. p_i	Attacked outputs $r_{a,i}$
1	6	4	2
2	8	5	2
3	7	4	1
4	6	3	1
5	9	6	2
6	7	5	2

where the submatrix $\mathcal{F}_{22}^{(i)} \in \mathbb{R}^{p_i \times p_i}$ corresponds to the dynamics of the output channels. To challenge the observer's estimation capabilities and illustrate the benefits of the proposed architecture, the matrix $\mathcal{F}_{22}^{(i)}$ is designed to include unstable modes. The complete multi-agent system is implemented in MATLAB/SIMULINK, with automatic generation of modular components for each agent, including the sliding-mode observer, equivalent injection computation, and distributed state–attack observer.

The agents are heterogeneous in both state and output dimensions, as summarized in Table 5.1. Each output matrix is configured as $C_i = [0\ I_{p_i}]$, extracting the last p_i components of the local state vector. Sensor attacks are injected through the matrix E_i, which selects $r_{a,i} < p_i$ output channels as potential attack points.

The process and measurement noise terms $w^{(i)}(t)$ and $v^{(i)}(t)$ are modeled as independent zero-mean Gaussian disturbances with variance 10^{-2}, representing moderate stochastic uncertainty in both plant dynamics and sensor readings. The overall simulation horizon is set to $T = 80\,\text{s}$.

5.6.2 Observer Implementation

The full observer architecture described in Sections 5.3–5.4 is implemented locally by each agent as follows:

(1) Sliding–Mode Output Regulator. The SMO is realized according to the dynamics in (5.6), with adaptive gain evolution governed by (5.7). The design parameters are selected as $\phi^{(i)} = 0.5$, $\gamma_i = 200$, $\lambda_i = 0.05$, and initial gain $k_0 = 0.5$. The equivalent injection signal is computed by low-pass filtering the product $K^{(i)}(t)\boldsymbol{\nu}^{(i)}(t)$, providing a smooth surrogate for the total output disturbance.

(2) Distributed State–Attack Observer (SAO). The distributed observer (5.17) is implemented with locally synthesized gains. The internal observer gain $K_z^{(i)}$ is chosen via pole placement with desired eigenvalues in $[-10, -100]$, ensuring fast convergence. The coupling gains $K_z^{(i,j)}$ are computed via least-squares matching of the form $\mathcal{H}_z^{(i,j)} \approx K_z^{(i,j)} C_z^{(j)}$, and the small-gain condition $\rho(\Gamma) < 1$ (Proposition 5.4.2) is verified numerically to guarantee network stability.

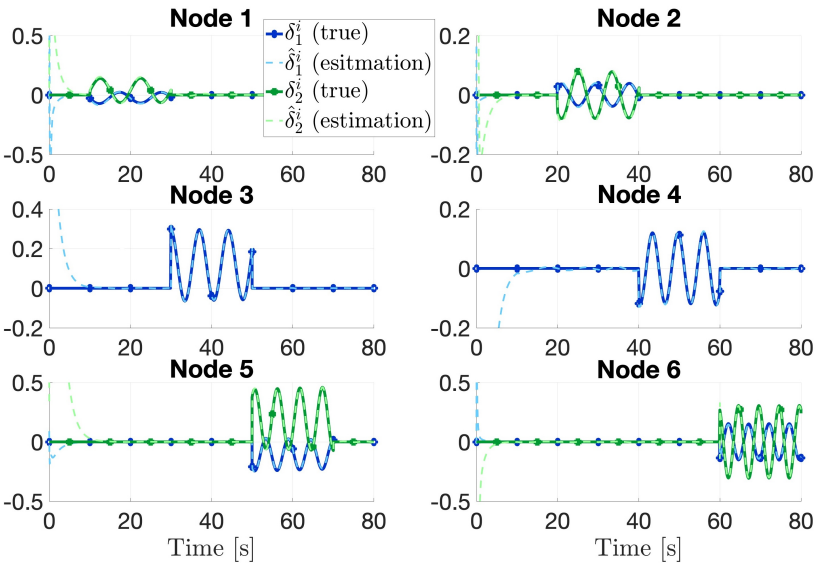


Figure 5.3: Estimated vs. true attack signals. Gray shading indicates active attack windows.

5.6.3 Attack Scenarios

To assess resilience, both Denial-of-Service (DoS) and replay attacks are simulated. Odd-indexed nodes ($i = 1, 3, 5$) are subjected to DoS events where the sensor output is suppressed: $\boldsymbol{\delta}^{(i)}(t) = -\boldsymbol{y}^{(i)}(t)$ for a specified time window. Even-indexed nodes ($i = 2, 4, 6$) are targeted with replay attacks of the form $\boldsymbol{\delta}^{(i)}(t) = \boldsymbol{y}^{(i)}(t - \tau_i) - \boldsymbol{y}^{(i)}(t)$, where $\tau_i \in [0.1, 1]$ s. Attack intervals are staggered across the network, starting at $t_{\text{on}}^{(i)} = 10i\,\text{s}$ and ending at $t_{\text{off}}^{(i)} = t_{\text{on}}^{(i)} + 20\,\text{s}$.

5.6.4 Results and Discussion

Figure 5.3 compares the true injected signals $\boldsymbol{\delta}^{(i)}(t)$ (solid) and their estimated counterparts $\hat{\boldsymbol{\delta}}^{(i)}(t)$ (dashed). Across all nodes, the estimates closely track the attack signals, with negligible delay and bounded steady-state error. Reconstruction is accurate even during overlapping attacks, demonstrating the scalability and resilience of the architecture.

Figure 5.4 shows the evolution of the attack estimation error norm $\|\boldsymbol{e}_\delta^{(i)}(t)\|$. Transient peaks at attack onset are quickly attenuated. In non-attacked agents, the error remains zero throughout, confirming that no false positives are introduced by the observer.

Figure 5.5 displays the norm of the state estimation error $\|\boldsymbol{e}_x^{(i)}(t)\|$. The error remains bounded even during active attacks, highlighting the decoupling between attack reconstruction and state estimation enabled by the augmented observer. These results are consistent with Proposition 5.4.2.

These simulations validate the theoretical results established in Section 5.5. The proposed two-layer observer architecture successfully reconstructs both the plant state and the unknown attack signals in a distributed setting. Estimation accuracy is preserved even under overlapping attacks, bounded noise, and sparse communication, confirming robustness and scalability of the scheme in realistic adversarial environments.

5.7 Discussion

This chapter presented a distributed observer framework based on Sliding Mode Observers for the reconstruction of attack signals targeting sensors and the isolation of compromised

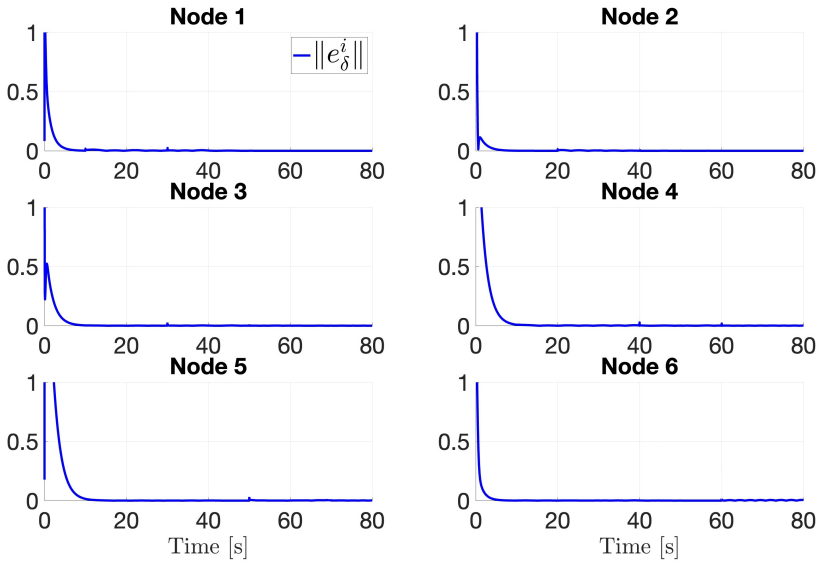


Figure 5.4: Norm of the attack estimation error $\|\boldsymbol{e}_\delta^{(i)}(t)\|$ across all agents.

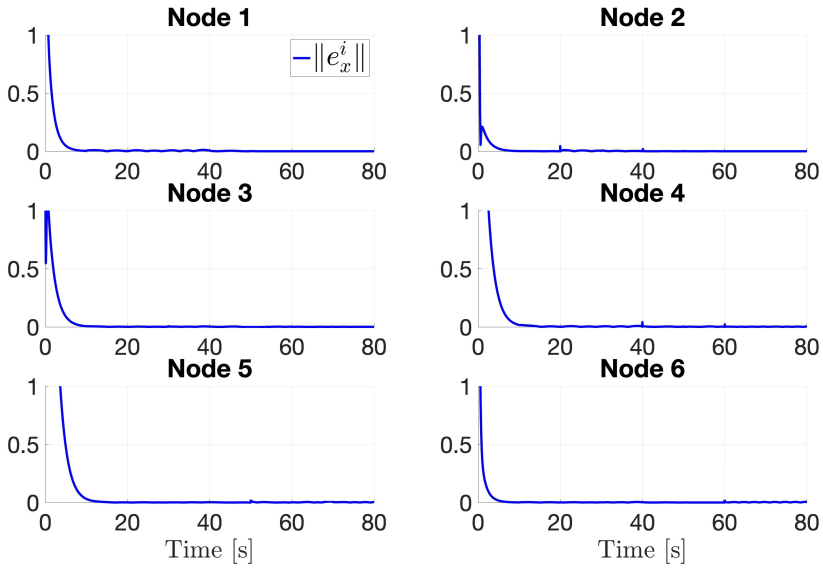


Figure 5.5: Norm of the state estimation error $\|e_x^{(i)}(t)\|$ for each agent.

nodes in large-scale linear cyber–physical systems. The proposed approach addresses a critical challenge in CPS security, namely the ability to maintain reliable state estimation in the presence of malicious sensor manipulations. By exploiting local measurements and limited inter-agent communication, each subsystem is equipped with an SMO capable of independently reconstructing the injected attack signals while preserving global consistency across the network.

From a theoretical standpoint, the proposed design establishes explicit bounds on the reconstruction error that depend on the disturbance and model uncertainties of each subsystem. Furthermore, the analytical results demonstrate that the isolation mechanism correctly identifies the attacked nodes without introducing false positives in non-compromised agents.

Beyond detection and isolation, this capability has important implications for secure control and resilience. Accurate reconstruction of attack signals allows for their direct compensation within the control loop, mitigating the effect of sensor corruption and

enabling continued safe operation under adversarial conditions. At the same time, the distributed architecture eliminates single points of failure and reduces communication overhead, making the scheme suitable for large-scale CPS deployments such as power grids, transportation networks, and industrial automation systems.

Overall, the proposed distributed SMO framework provides an analytically sound and practically scalable solution for attack reconstruction and isolation in large-scale CPS. It combines strong theoretical guarantees with demonstrated robustness in simulation, laying the groundwork for future extensions toward nonlinear systems, dynamic network topologies, and event-triggered communication strategies for resilient and secure estimation.

References — Observer Based Anomaly Detection and Isolation

[2] H. Lee, R. Nuqui, and J. Hong, "Cyber intrusion detection and sensor failure diagnosis using statistical hypothesis-based correlation: Application to FACTS AC filter banks," in *2024 IEEE/PES Transmission and Distribution Conference and Exposition (T&D)*, Journal Abbreviation: 2024 IEEE/PES Transmission and Distribution Conference and Exposition (T&D), May 6, 2024, pp. 1–5.

[10] Pasqualetti, F., Dörfler, F., and Bullo, F., "Attack detection and identification in cyber-physical systems," *IEEE transactions on automatic control*, vol. 58, no. 11, pp. 2715–2729, 2013.

[12] Weng, P., Chen, B., and Wang, R., "Secure fusion estimation against FDI sensor attacks in cyber–physical systems," *Franklin Open*, vol. 7, 2024. DOI: 10.1016/j.fraope.2024.100108. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85212601982&doi=10.1016%2fj.fraope.2024.100108&partnerID=40&md5=e2cc6cceade2ff9487294b998ff4ea39.

[15] Li, J., Yang, D., and Su, Q., "Sliding mode resilient control and application based on intermediate variable observer in smart grid," *International Journal of Control, Automation and Systems*, vol. 21, no. 6, pp. 1803–1815, 2023. DOI: 10.1007/s12555-021-0916-0. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85156221639&doi=10.1007%2fs12555-021-0916-0&partnerID=40&md5=5e9f10d1f575b855a6dab889f9a75cdd.

[17] Chen, G., Zhang, Y., Gu, S., and Hu, W., "Resilient state estimation and control of cyber-physical systems against false data injection attacks on both actuator and sensors," *IEEE Transactions on Control of Network Systems*, vol. 9, no. 1, pp. 500–510, 2022. DOI: 10.1109/TCNS.2021.3113265. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85115126198&doi=10.1109%2fTCNS.2021.3113265&partnerID=40&md5=c70f88330325683ae416e838d0f90786.

[19] Chattopadhyay, A. and Mitra, U., "Security against false data-injection attack in cyber-physical systems," *IEEE Transactions on Control of Network Systems*, vol. 7, no. 2, pp. 1015–1027, 2020. DOI: 10.1109/TCNS.2019.2927594. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85069439477&doi=10.1109%2fTCNS.2019.2927594&partnerID=40&md5=c1f7af7cb157c776bb7e05c26d346374.

[31] Teixeira, A., Sandberg, H., and Johansson, K., "Networked control systems under cyber attacks with applications to power networks," presented at the Proceedings of the 2010 American Control Conference, ACC 2010, 2010, pp. 3690–3696. DOI: 10.1109/acc.2010.5530638. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-77957775842&doi=10.1109%2facc.2010.5530638&partnerID=40&md5=5097e13dfcc486051b09d9ced37a0cbd.

[32] Chowdhury, N., Belikov, J., Baimel, D., and Levron, Y., "Observer-based detection and identification of sensor attacks in networked CPSs," *Automatica*, vol. 121, 2020. DOI: 10.1016/j.automatica.2020.109166. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85088921633&doi=10.1016%2fj.automatica.2020.109166&partnerID=40&md5=d01180439d1af8f1e18094afbfefc239.

[33] Ahmad, S. and Ahmed, H., "Robust intrusion detection for resilience enhancement of industrial control systems: An extended state observer approach," *IEEE Transactions on Industry Applications*, vol. 59, no. 6, pp. 7735–7743, 2023. DOI: 10.1109/TIA.2023.3305361. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85168281024&doi=10.1109%2fTIA.2023.3305361&partnerID=40&md5=0aadf0e4dddca4fcbd379c3cf8c5faf2.

[34] Y. Mo, R. Chabukswar, and B. Sinopoli, "Detecting integrity attacks on SCADA systems," *IEEE Transactions on Control Systems Technology*, vol. 22, no. 4, pp. 1396–1407, Jul. 2014. DOI: 10.1109/TCST.2013.2280899.

[35] K. Manandhar, X. Cao, F. Hu, and Y. Liu, "Detection of faults and attacks including false data injection attack in smart grid using kalman filter," *IEEE Transactions on Control of Network Systems*, vol. 1, no. 4, pp. 370–379, Dec. 2014. DOI: 10.1109/TCNS.2014.2357531.

[36] Zhang, H., He, L., Wang, Z., Gao, S., and Yan, H., "The detection mechanism for false data injection attack via the ellipsoidal set-membership approach," *Asian Journal of Control*, vol. 25, no. 6, pp. 4853–4863, 2023. DOI: 10.1002/asjc.3166. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85164323819&doi=10.1002%2fasjc.3166&partnerID=40&md5=e0d42f40a1398ce244db93233ac19307.

[37] Li, J., Wang, Z., Shen, Y., and Xie, L., "Attack detection for cyber-physical systems: A zonotopic approach," *IEEE Transactions on Automatic Control*, vol. 68, no. 11, pp. 6828–6835, 2023. DOI: 10.1109/TAC.2023.3240383. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85148462989&doi=10.1109%2fTAC.2023.3240383&partnerID=40&md5=0ec1842e79fa72b3241e8601a323c516.

[38] Forti, N., Battistelli, G., Chisci, L., and Sinopoli, B., "Joint attack detection and secure state estimation of cyber-physical systems," *International Journal of Robust and Nonlinear Control*, vol. 30, no. 11, pp. 4303–4330, 2020. DOI: 10.1002/rnc.4724. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85071986226&doi=10.1002%2frnc.4724&partnerID=40&md5=c64ef74d02845a7ffa03b43962dd0eff.

[39] Leko, D. and Vasak, M., "Minimum robust invariant sets and kalman filtering in cyber attacking and defending," presented at the Proceedings of the American Control Conference, vol. 2022-June, 2022, pp. 2924–2931. DOI: 10.23919/ACC53348.2022.9867272. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85138492583&doi=10.23919%2fACC53348.2022.9867272&partnerID=40&md5=5cb4fd3346b494777218c47a362a5266.

[40] Su, Q., Li, S., Gao, Y., Huang, X., and Li, J., "Observer-based detection and reconstruction of dynamic load altering attack in smart grid," *Journal of the Franklin Institute*, vol. 358, no. 7, pp. 4013–4027, 2021. DOI: 10.1016/j.jfranklin.2021.02.008. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85102077025&doi=10.1016%2fj.jfranklin.2021.02.008&partnerID=40&md5=f3f5033ebd2f30654f193b26053deaca.

[41] Li, J., Yang, D., and Su, Q., "Reliable control strategy based on sliding mode observer against FDI attacks in smart grid," *Asian Journal of Control*, vol. 25, no. 2, pp. 910–920, 2023. DOI: 10.1002/asjc.2839. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85129260072&doi=10.1002%2fasjc.2839&partnerID=40&md5=8034b5a649f3cebf8d0c518b7e03cc82.

[43] Liu, H., Li, Y., Han, Q.-L., Raissi, T., and Chai, T., "Secure estimation, attack isolation, and reconstruction based on zonotopic unknown input observer," *IEEE Transactions on Automatic Control*, vol. 68, no. 12, pp. 7312–7325, 2023. DOI: 10.1109/TAC.2023.3275965. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85160221136&doi=10.1109%2fTAC.2023.3275965&partnerID=40&md5=3dfe8540d677cbb541e9a028c2633f16.

[44] Yang, H. and Ma, R., "Sparse attack reconstruction for cyber–physical systems via descriptor reduced-order observer," *Nonlinear Analysis: Hybrid Systems*, vol. 50, 2023. DOI: 10.1016/j.nahs.2023.101380. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85166240334&doi=10.1016%2fj.nahs.2023.101380&partnerID=40&md5=6c4f534825b72b07cac859cf96bb0986.

[45] Kachhwaha, M., Modi, H., Nehra, M., and Fulwani, D., "Robust observer-based defense strategy against actuator and sensor cyber-attacks in DCMGs," *IEEE Transactions on Industrial Informatics*, vol. 20, no. 10, pp. 11 687–11 696, 2024. DOI: 10.1109/TII.2024.3412211. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85206516903&doi=10.1109%2fTII.2024.3412211&partnerID=40&md5=d0e7331a2ad7d1793579a40eb1e27078.

[46] Y. Mo and B. Sinopoli, "On the performance degradation of cyber-physical systems under stealthy integrity attacks," *IEEE Transactions on Automatic Control*, vol. 61, no. 9, pp. 2618–2624, Sep. 2016. DOI: 10.1109/TAC.2015.2498708.

[47] Bonczek, P. and Bezzo, N., "Memoryless cumulative sign detector for stealthy CPS sensor attacks," presented at the IFAC-PapersOnLine, vol. 53, 2020, pp. 838–844. DOI: 10.1016/j.ifacol.2020.12.840. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85107537377&doi=10.1016%2fj.ifacol.2020.12.840&partnerID=40&md5=f3fcaf93fd070d910adca02236b651c7.

[48] Chen, Y., Liang, Z., Zhao, Z., Wang, H., and Dong, J., "Super-twisting nonsingular terminal sliding mode control for cyber physical system under FDI attacks," *Nonlinear Dynamics*, vol. 113, no. 3, pp. 2545–2558, 2025. DOI: 10.1007/s11071-024-10358-8. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85205535663&doi=10.1007%2fs11071-024-10358-8&partnerID=40&md5=27f3b222d0a7ef7cbbbb7d26809fd71e.

[49] Nateghi, S., Shtessel, Y., and Edwards, C., "Resilient control of cyber-physical systems under sensor and actuator attacks driven by adaptive sliding mode observer," *International Journal of Robust and Nonlinear Control*, vol. 31, no. 15, pp. 7425–7443, 2021. DOI: 10.1002/rnc.5694. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85110016740&doi=10.1002%2frnc.5694&partnerID=40&md5=76e7a5942040f084365b2a59a6513c47.

[53] Liu, H., Wang, S., and Li, Y., "Event-triggered control and proactive defense for cyber-physical systems," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 52, no. 10, pp. 6305–6313, 2022. DOI: 10.1109/TSMC.2022.3144337. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85124075239&doi=10.1109%2fTSMC.2022.3144337&partnerID=40&md5=023f58b4b0c7cb5e1636d8b04ee52e37.

[55] Zhang, H., Cheng, P., Shi, L., and Chen, J., "Optimal denial-of-service attack scheduling with energy constraint," *IEEE Transactions on Automatic Control*, vol. 60, no. 11, pp. 3023–3028, 2015. DOI: 10.1109/TAC.2015.2409905. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-84945315428&doi=10.1109%2fTAC.2015.2409905&partnerID=40&md5=715aacd03ce9a4c2a096d193d6ad790f.

[56] Li, G., Wang, Y., Min, Z., and Zhang, D., "Faults and attacks collaborative detection for cyber-physical systems with mixed network attacks," presented at the Chinese Control Conference, CCC, 2024, pp. 4859–4864. DOI: 10.23919/CCC63176.2024.10661834. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85205509751&doi=10.23919%2fCCC63176.2024.10661834&partnerID=40&md5=a2aa3e0dc924d40489d3e3eae0ddfa79.

[57] J. Xiao and M. Feroskhan, "Cyber attack detection and isolation for a quadrotor UAV with modified sliding innovation sequences," *IEEE Transactions on Vehicular Technology*, vol. 71, no. 7, pp. 7202–7214, Jul. 2022. DOI: 10.1109/TVT.2022.3170725.

[58] Piao, X. and Lin, C., "Sensor attack and state estimation for cyber-physical systems with state delay via reduced-order observer," presented at the Chinese Control Conference, CCC, vol. 2023-July, 2023, pp. 5158–5162. DOI: 10.23919/CCC58697.2023.10240583. [Online]. Available: https://www.scopus.com/inward/record.

uri?eid=2-s2.0-85175578064&doi=10.23919%2fCCC58697.2023.10240583&partnerID=40&md5=bfaf5b418aa8373844a692baa649992d.

[65] X. Huang, J. Li, and Q. Su, "An observer with cooperative interaction structure for biasing attack detection and secure control," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 53, no. 4, pp. 2543–2553, Apr. 2023. DOI: 10.1109/TSMC.2022.3213516.

[66] Teixeira, A., Sou, K. C., Sandberg, H., and Johansson, K. H., "Secure control systems: A quantitative risk management approach," *IEEE Control Systems Magazine*, vol. 35, no. 1, pp. 24–45, 2015.

[67] G. Lu and D. Feng, "Network security situation awareness for industrial control system under integrity attacks," in *2018 21st International Conference on Information Fusion (FUSION)*, Journal Abbreviation: 2018 21st International Conference on Information Fusion (FUSION), Jul. 10, 2018, pp. 1808–1815. DOI: 10.23919/ICIF.2018.8455208.

[68] Ren, G., Liu, H., Zhang, Y., and Li, Y., "Attack detection based on encoding-decoding approach for cyber-physical systems," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 53, no. 10, pp. 6293–6302, 2023. DOI: 10.1109/TSMC.2023.3283414. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85162858130&doi=10.1109%2fTSMC.2023.3283414&partnerID=40&md5=7fc527c6f9b8b94d096dcd9868f1dfab.

[69] Khorrami, L. and Afshar, A., "Attack detection in active queue management within large-scale networks control system with information of network and physical system," presented at the 2016 24th Iranian Conference on Electrical Engineering, ICEE 2016, 2016, pp. 714–719. DOI: 10.1109/IranianCEE.2016.7585614. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-84994085527&doi=10.1109%2fIranianCEE.2016.7585614&partnerID=40&md5=920b9a99ee3553e3b1e5d592968e1f81.

[70] Zheng, J., Liu, D., Hua, J., and Ning, X., "Optimal two-channel switching false data injection attacks against remote state estimation of the unmanned aerial vehicle cyber-physical system," *Defence Technology*, vol. 47, pp. 319–332, 2025. DOI: 10.1016/j.dt.2024.12.025. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85215929359&doi=10.1016%2fj.dt.2024.12.025&partnerID=40&md5=23042d033060c30901ab22d750b45114.

[72] Sun, D., Cho, M., and Hwang, I., "On attack detection and identification for the cyber–physical system using lifted system model," *Automatica*, vol. 176, 2025. DOI: 10.1016/j.automatica.2025.112243. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-105000842080&doi=10.1016%2fj.automatica.2025.112243&partnerID=40&md5=1ca057402feab68dd00c03218dcb52d6.

[73] I. Jovanov and M. Pajic, "Relaxing integrity requirements for attack-resilient cyber-physical systems," *IEEE Transactions on Automatic Control*, vol. 64, no. 12, pp. 4843–4858, Dec. 2019. DOI: 10.1109/TAC.2019.2898510.

[74] Fan, X.-Y. and Lin, W.-J., "Reachable set control for cyber-physical systems with false data injection attacks," presented at the Proceedings - 2023 IEEE 6th International Conference on Industrial Cyber-Physical Systems, ICPS 2023, 2023. DOI: 10.1109/ICPS58381.2023.10127989. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85163058876&doi=10.1109%2fICPS58381.2023.10127989&partnerID=40&md5=aed699ad761792c4420ac427ed32e081.

[75] Qiu, S. and Liu, H., "A new zonotope-based attack detection method for UAV," presented at the Chinese Control Conference, CCC, vol. 2022-July, 2022, pp. 4276–4280. DOI: 10.23919/CCC55666.2022.9902124. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85140482486&doi=10.23919%2fCCC55666.2022.9902124&partnerID=40&md5=9bd3f93303cc73d974a6f6e05a5e12c4.

[76] Wang, J. et al., "Covert attack detection and resilient control of quadcopter," *IEEE Transactions on Industrial Informatics*, vol. 21, no. 5, pp. 3924–3933, 2025. DOI: 10.1109/TII.2025.3528579. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85218769026&doi=10.1109%2fTII.2025.3528579&partnerID=40&md5=61e799c8e3383266419e2df8cf94152b.

[77] Eslami, A. and Khorasani, K., "Cyber-attack detection and isolation in event-based cyber-physical systems," presented at the Proceedings of the American Control Conference, 2024, pp. 3460–3466. DOI: 10.23919/ACC60939.2024.10644303. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85204468219&doi=10.23919%2fACC60939.2024.10644303&partnerID=40&md5=5ffadeedb440bd91d84c5410e5dd3a8d.

[78] Bahrami, M. and Jafarnejadsani, H., "Privacy-preserving stealthy attack detection in multi-agent control systems," presented at the Proceedings of the IEEE Conference on Decision and Control, vol. 2021-December, 2021, pp. 4194–4199. DOI: 10.1109/CDC45484.2021.9683427. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85125691405&doi=10.1109%2fCDC45484.2021.9683427&partnerID=40&md5=92a261c8ba4acb98043cd749ddd967eb.

[79] Q. Liu, Y. Long, T. Li, and C. L. P. Chen, "Attack tolerant fault detection for CPSs: An unknown input interval observer approach," *IEEE Transactions on Automation Science and Engineering*, vol. 22, pp. 1163–1172, 2025.

[80] Wang, J., Zheng, Y., Ding, J., Xie, X., and Zhang, W., "Multiasynchronous extended dissipative sliding mode control of LC circuits in grid-connected system under actuator attacks," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 72, no. 4, pp. 1609–1620, 2025. DOI: 10.1109/TCSI.2024.3517703. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-105001685778&doi=10.1109%2fTCSI.2024.3517703&partnerID=40&md5=06fbc2750170491b32fb7e942b49a47c.

[81] Y. Yan, P. Antsaklis, and V. Gupta, "A resilient design for cyber physical systems under attack," in *2017 American Control Conference (ACC)*, Journal Abbreviation: 2017 American Control Conference (ACC), May 24, 2017, pp. 4418–4423. DOI: 10.23919/ACC.2017.7963635.

[82] Ahmad, S. and Ahmed, H., "Robust intrusion detection for resilience enhancement of industrial control systems: An extended state observer approach," presented at the 2022 IEEE Texas Power and Energy Conference, TPEC 2022, 2022. DOI: 10.1109/TPEC54980.2022.9750751. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85128775901&doi=10.1109%2fTPEC54980.2022.9750751&partnerID=40&md5=3842adab1600e49b6a8c1d612c000eb4.

[84] Liu, H., Zhang, Y., Li, Y., and Niu, B., "Proactive attack detection scheme based on watermarking and moving target defense," *Automatica*, vol. 155, 2023. DOI: 10.1016/j.automatica.2023.111163. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85163519547&doi=10.1016%2fj.automatica.2023.111163&partnerID=40&md5=e53ad9c7782d7717a52ddc44030d9d4c.

[85] T. Rhouma, J. Y. Keller, K. Chabir, D. Sauter, and M. N. Abdelkrim, "Coding control signals and switching LQG controller for secure fault-tolerant control against stealthy false data injection," in *2016 3rd Conference on Control and Fault-Tolerant Systems (SysTol)*, Journal Abbreviation: 2016 3rd Conference on Control and Fault-Tolerant Systems (SysTol), Sep. 7, 2016, pp. 750–755. DOI: 10.1109/SYSTOL.2016.7739838.

[86] Y. Chen, S. Kar, and J. M. F. Moura, "Optimal attack strategies subject to detection constraints against cyber-physical systems," *IEEE Transactions on Control of Network Systems*, vol. 5, no. 3, pp. 1157–1168, Sep. 2018. DOI: 10.1109/TCNS.2017.2690399.

[87] Pei, C., Xiao, Y., Liang, W., and Han, X., "A deviation-based detection method against false data injection attacks in smart grid," *IEEE Access*, vol. 9, pp. 15 499–15 509, 2021. DOI: 10.1109/access.2021.3051155. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85099594437&doi=10.1109%2faccess.2021.3051155&partnerID=40&md5=265dd819e2aac2707de79a68e40c8ee5.

[88] Li, J. and Zhang, Y., "Detection of false data injection attack in CPS by adaptive unscented kalman filter," presented at the IECON Proceedings (Industrial Electronics Conference), 2023. DOI: 10.1109/IECON51785.2023.10312054. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85179524773&doi=10.1109%2fIECON51785.2023.10312054&partnerID=40&md5=7c3d49e3ce97d0bf72db7a4987a8ec21.

[92] J. Pisani, G. Cavone, L. Giarré, and F. Pascucci, "Industrial control systems attack detection by hybrid digital twin," in *2023 European Control Conference (ECC)*, Journal Abbreviation: 2023 European Control Conference (ECC), Jun. 13, 2023, pp. 1–6. DOI: 10.23919/ECC57647.2023.10178183.

[94] Huang, X., Ding, D.-W., Ren, Y., and Xie, X., "Attack detection of complex cyber-physical networks via set-membership theory," *IEEE Transactions on Network Science and Engineering*, vol. 10, no. 4, pp. 2283–2293, 2023. DOI: 10.1109/TNSE.2023.3245214. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85149414924&doi=10.1109%2fTNSE.2023.3245214&partnerID=40&md5=01edab9a98e6b02f6721bfe411831439.

[95] Wang, P., Ren, X., and Wang, D., "Nonlinear cyber-physical system security control under false data injection attack," presented at the Chinese Control Conference, CCC, vol. 2022-July, 2022, pp. 4311–4316. DOI: 10.23919/CCC55666.2022.9902555. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85140481263&doi=10.23919%2fCCC55666.2022.9902555&partnerID=40&md5=e0191709514598d09e012ecedfdbae8e.

[96] Yang, H., Yin, S., Han, H., and Sun, H., "Sparse actuator and sensor attacks reconstruction for linear cyber-physical systems with sliding mode observer," *IEEE Transactions on Industrial Informatics*, vol. 18, no. 6, pp. 3873–3884, 2022. DOI: 10.1109/TII.2021.3111221. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85114735782&doi=10.1109%2fTII.2021.3111221&partnerID=40&md5=253061fb0d55528459bb27f4e2c0e662.

[97] L. Bin Bashar et al., "Observer-based attack and disturbance mitigation for heavy-duty vehicles in an autonomous docking application," *IEEE Access*, vol. 12, pp. 142 488–142 504, 2024. DOI: 10.1109/ACCESS.2024.3467263.

[98] X. Wang, X. Luo, M. Zhang, Z. Jiang, and X. Guan, "Detection and isolation of false data injection attacks in smart grid via unknown input interval observer," *IEEE Internet of Things Journal*, vol. 7, no. 4, pp. 3214–3229, Apr. 2020. DOI: 10.1109/JIOT.2020.2966221.

[99] Yang, H., Han, H., Yin, S., Han, H., and Wang, P., "Sliding mode-based adaptive resilient control for markovian jump cyber–physical systems in face of simultaneous actuator and sensor attacks," *Automatica*, vol. 142, 2022. DOI: 10.1016/j.automatica.2022.110345. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85129516600&doi=10.1016%2fj.automatica.2022.110345&partnerID=40&md5=e3bdbcb9b072d221358879e9dd319460.

[100] Fattore, G. and Valcher, M. E., "A data-driven approach to UIO-based fault diagnosis," en, in *2024 IEEE 63rd Conference on Decision and Control (CDC)*, Milan, Italy: IEEE, Dec. 2024, pp. 6831–6836. DOI: 10.1109/CDC56724.2024.10886303. Accessed: Jul. 2, 2025. [Online]. Available: https://ieeexplore.ieee.org/document/10886303/.

[101] Yang, G., Herrera, L., and Yao, X., "False Data Injection Attack Detection in DC Microgrids Based on Data-Driven Unknown Input Observers," *IEEE Journal of Emerging and Selected Topics in Power Electronics*, 2025. DOI: 10.1109/JESTPE.2025.3539958. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85217523555&doi=10.1109%2fJESTPE.2025.3539958&partnerID=40&md5=d26fb059848670a429ec371032343010.

[102] Wang, X., Wang, X., Luo, X., Guan, X., and Wang, S., "Novel cyber-physical collaborative detection and localization method against dynamic load altering attacks in smart energy grids," *Global Energy Interconnection*, vol. 7, no. 3, pp. 362–376, 2024. DOI: 10.1016/j.gloei.2024.06.003. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85197504950&doi=10.1016%2fj.gloei.2024.06.003&partnerID=40&md5=cc5d027f9f498e13b1ca396961f04a50.

[103] Gao, B., Bu, B., and Wang, X., "A comprehensive resilient control strategy for CBTC systems through train-to-train communications under malicious attacks," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 11, pp. 21 015–21 033, 2022. DOI: 10.1109/TITS.2022.3192510. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85135756590&doi=10.1109%2fTITS.2022.3192510&partnerID=40&md5=f7efcb044835ada297b0139fd389b045.

[104] Menelaou, C., Zhang, K., Timotheou, S., Panayiotou, C. G., and Parisini, T., "Stealthy attack detection of controlled ramp meters in freeway networks," in *2024 IEEE 63rd Conference on Decision and Control (CDC)*, IEEE, 2024, pp. 1289–1294.

[140] Forti, N., Battistelli, G., Chisci, L., and Sinopoli, B., "Worst-case analysis of joint attack detection and resilient state estimation," presented at the 2017 IEEE 56th Annual Conference on Decision and Control, CDC 2017, vol. 2018-January, 2017, pp. 182–188. DOI: 10.1109/CDC.2017.8263663. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85046148151&doi=10.1109%2fCDC.2017.8263663&partnerID=40&md5=389d4d199081822ff9f999d05135007f.

[141] Giraldo, J. et al., "A survey of physics-based attack detection in cyber-physical systems," *ACM Computing Surveys*, vol. 51, no. 4, pp. 1–36, Jul. 31, 2019. Accessed: May 20, 2025. [Online]. Available: https://dl.acm.org/doi/10.1145/3203245.

[142] Lukens, J., Passian, A., Yoginath, S. B., Law, K., and Dawson, J. A., "Bayesian estimation of oscillator parameters: Toward anomaly detection and cyber-physical system security," *Sensors*, vol. 22, no. 16, 2022. DOI: 10.3390/s22166112.

[143] Yang, X., Beg, O., Kenigsberg, M., and Johnson, T. T., "A framework for identification and validation of affine hybrid automata from input-output traces," *ACM Transactions on Cyber-Physical Systems (TCPS)*, vol. 6, no. 1, pp. 1–24, 2022. DOI: 10.1145/3470455.

[144] Babunski, D., Poposki, F., Koleva, R., Zaev, E., and Rath, G., "Kalman filter based detection of fdia on a hybrid nonlinear hydro-power plant control system model," *2022 11th Mediterranean Conference on Embedded Computing (MECO)*, pp. 1–4, 2022.

[145] Zhang, K., Keliris, C., Parisini, T., Jiang, B., and Polycarpou, M. M., "Passive attack detection for a class of stealthy intermittent integrity attacks," *IEEE/CAA Journal of Automatica Sinica*, vol. 10, no. 4, pp. 898–915, 2023.

[146] Suo, Y., Chai, R., Chai, S., Farhan, I., Xia, Y., and Liu, G.-P., "Attack detection and secure state estimation of collectively observable cyber-physical systems under false data injection attacks," *IEEE Transactions on Automatic Control*, vol. 69, no. 3, pp. 2067–2074, 2024. DOI: 10.1109/TAC.2023.3316160. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85172996372&doi=10.1109%2fTAC.2023.3316160&partnerID=40&md5=7bd908468e3504be1d4e90bc65a79e4e.

[147] Chen, Y., Long, Y., and Li, T., "Attacks detection and security control against false data injection attacks based on interval type-2 fuzzy system," presented at the IECON Proceedings (Industrial Electronics Conference), vol. 2022-October, 2022. DOI: 10.1109/IECON49645.2022.9968628. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85143887361&doi=10.1109%2fIECON49645.2022.9968628&partnerID=40&md5=231323f8c6fb5872d3e1f747ef46a363.

[148] Meng, Q., Kasis, A., Yang, H., and Polycarpou, M. M., "Secure state estimation of networked switched systems under denial-of-service attacks," *European Journal of Control*, vol. 80, p. 101 037, 2024.

[149] Chen, X., Hu, S., Zhang, W., Xie, X., and Yue, D., "DoS-resilient time varying estimators and controllers co-design for NCSs under sensor and actuator attacks," *IEEE Transactions on Information Forensics and Security*, 2025. DOI: 10.1109/TIFS.2025.3576587. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-105007555944&doi=10.1109%2fTIFS.2025.3576587&partnerID=40&md5=3a1d9c43e8c4d4a9fd762ff9f76c43cf.

[150] Guan, Z., Sun, N., Xu, Y., and Yang, T., "A comprehensive survey of false data injection in smart grid," *International Journal of Wireless and Mobile Computing*, vol. 8, no. 1, pp. 27–33, 2015. DOI: 10.1504/IJWMC.2015.066756.

[151] Oliveira, S., Leal, A. B., Teixeira, M., and Lopes, Y. K., "A classification of cybersecurity strategies in the context of discrete event systems," *Annual Reviews in Control*, vol. 56, p. 100 907, 2023. Accessed: May 22, 2025.

[152] Cetinkaya, A., Ishii, H., and Hayakawa, T., "Event-triggered output feedback control resilient against jamming attacks and random packet losses," *IFAC-PapersOnLine*, vol. 48, no. 22, pp. 270–275, 2015.

[153] Liu, S., Li, S., and Xu, B., "Event-triggered resilient control for cyber-physical system under denial-of-service attacks," *International Journal of Control*, vol. 93, no. 8, pp. 1907–1919, 2020.

[154] Wakaiki, M., Cetinkaya, A., and Ishii, H., "Quantized output feedback stabilization under dos attacks," in *2018 Annual American Control Conference (ACC)*, IEEE, 2018, pp. 6487–6492.

[155] Chu, H., Gorbachev, S., Yue, D., and Dou, C., "Output formation containment for multiagent systems under multipoint multipattern FDI attacks: A resilient impulsive compensation control approach," *IEEE Transactions on Cybernetics*, vol. 54, no. 4, pp. 2606–2617, 2024. DOI: 10.1109/TCYB.2023.3319647. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85174824102&doi=10.1109%2fTCYB.2023.3319647&partnerID=40&md5=6d11c085a801b2ec16b01b9b14c1ba0d.

[156] Guadarrama-Estrada, A., Osorio-Gordillo, G., Vargas-Méndez, R., Reyes-Reyes, J., and Astorga-Zaragoza, C., "Cyber–physical system attack detection and isolation: A takagi–sugeno approach," *Mathematical and Computational Applications*, vol. 30, no. 1, 2025. DOI: 10.3390/mca30010012. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85219191853&doi=10.3390%2fmca30010012&partnerID=40&md5=558bbf2d388b178bb1451fa8797933f4.

[157] Cao, X., Liu, S., and Cen, J., "Observer-based adaptive neural asynchronous h∞ control for fuzzy markov jump systems under FDI attacks," *Journal of the Franklin Institute*, vol. 361, no. 16, 2024. DOI: 10.1016/j.jfranklin.2024.107147. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85201009507&doi=10.1016%2fj.jfranklin.2024.107147&partnerID=40&md5=60d7a1911616371b8192bf50d28206b6.

[158] X. Hu, X. Liu, and P. Wang, "Intermediate estimator-based attack reconstruction for cyber-physical systems," in *2023 5th International Conference on Industrial Artificial Intelligence (IAI)*, Journal Abbreviation: 2023 5th International Conference on Industrial Artificial Intelligence (IAI), Aug. 21, 2023, pp. 1–6. DOI: 10.1109/IAI59504.2023.10327621.

[159] He, K., Li, T., Long, Y., Park, J., and Chen, C., "Secure state estimation and actuator attack reconstruction for cyber-physical systems based on sliding-mode observer," *International Journal of Robust and Nonlinear Control*, vol. 33, no. 14, pp. 8508–8523, 2023. DOI: 10.1002/rnc.6833. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85162005459&doi=10.1002%2frnc.6833&partnerID=40&md5=e6b9b56d0616345e914aba76993b6162.

[160] K. Manandhar and X. Cao, "Attacks/faults detection and isolation in the smart grid using kalman filter," in *2014 23rd International Conference on Computer Communication and Networks (ICCCN)*, Journal Abbreviation: 2014 23rd International Conference on Computer Communication and Networks (ICCCN), Aug. 4, 2014, pp. 1–6. DOI: 10.1109/ICCCN.2014.6911809.

[161] Amin, S., Litrico, X., Sastry, S., and Bayen, A., "Cyber security of water scada systems-part II: Attack detection using enhanced hydrodynamic models," *IEEE Transactions on Control Systems Technology*, vol. 21, no. 5, pp. 1679–1693, 2013. DOI: 10.1109/TCST.2012.2211874. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-84897648424&doi=10.1109%2fTCST.2012.2211874&partnerID=40&md5=52bba36c07b352a034ae451253dfb515.

[162] Khalghani, M., Solanki, J., Solanki, S., and Sargolzaei, A., "Resilient and stochastic load frequency control of microgrids," presented at the IEEE Power and Energy Society General Meeting, vol. 2019-August, 2019. DOI: 10.1109/PESGM40551.2019.8974111. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85079035544&doi=10.1109%2fPESGM40551.2019.8974111&partnerID=40&md5=6cca604bbd63169457dd4a6f060486b9.

[163] Xie, C.-H. and Yang, G.-H., "Observer-based attack-resilient control for linear systems against FDI attacks on communication links from controller to actuators," *International Journal of Robust and Nonlinear Control*, vol. 28, no. 15, pp. 4382–4403, 2018. DOI: 10.1002/rnc.4236. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85053349940&doi=10.1002%2frnc.4236&partnerID=40&md5=972e9f32999a65676e5579edb09eab02.

[164] A. O. Aluko, D. G. Dorrell, and E. E. Ojo, "Observer-based detection and mitigation scheme for isolated microgrid under false data injection attack," in *2021 IEEE Southern Power Electronics Conference (SPEC)*, Journal Abbreviation: 2021 IEEE Southern Power Electronics Conference (SPEC), Dec. 6, 2021, pp. 1–6.

[165] Chen, L., Edwards, C., and Alwi, H., "Sensor redundancy based FDI using an LPV sliding mode observer," *IET Control Theory and Applications*, vol. 12, no. 14, pp. 1956–1963, 2018. DOI: 10.1049/iet-cta.2017.0818. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85052644224&doi=10.1049%2fiet-cta.2017.0818&partnerID=40&md5=53fc505be181a017cf7ee48cf93131f3.

[166] Zhao, L., Li, W., Li, Y., Han, N., and Zheng, N., "Integrated security control for nonlinear CPS with actuator fault and FDI attack: An active attack-tolerant approach," *Actuators*, vol. 12, no. 5, 2023. DOI: 10.3390/act12050216. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85160340432&doi=10.3390%2fact12050216&partnerID=40&md5=350f6081a86f12e30f4bce799f48163d.

[167] Gallo, A., C. Anand, S., Teixeira, A., and Ferrari, R., "Switching multiplicative watermark design against covert attacks," *Automatica*, vol. 177, 2025. DOI: 10.1016/j.automatica.2025.112301. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-105002132409&doi=10.1016%2fj.automatica.2025.112301&partnerID=40&md5=3ee3e7b61fd1ce64b732304b0e4eadae.

[168] Sasahara, H., Ishizaki, T., Imura, J.-I., and Sandberg, H., "Disconnection-aware attack detection in networked control systems," presented at the IFAC-PapersOnLine, vol. 53, 2020, pp. 3515–3520. DOI: 10.1016/j.ifacol.2020.12.1710. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85098368830&doi=10.1016%2fj.ifacol.2020.12.1710&partnerID=40&md5=2a28095dc3be0c9833252a447ac25b01.

[169] Mao, Y., Jafarnejadsani, H., Zhao, P., Akyol, E., and Hovakimyan, N., "Novel stealthy attack and defense strategies for networked control systems," *IEEE Transactions on Automatic Control*, vol. 65, no. 9, pp. 3847–3862, 2020. DOI: 10.1109/TAC.2020.2997363. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85090948945&doi=10.1109%2fTAC.2020.2997363&partnerID=40&md5=832124b75669ef53f7a7cfdffabd8ce3.

[170] Ahmed, Z., Nasir, M., Alsekait, D., Shah, M., AbdElminaam, D., and Ahmad, F., "Control conditions for equal power sharing in multi-area power systems for resilience against false data injection attacks," *Energies*, vol. 17, no. 22, 2024. DOI: 10.3390/en17225757. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85210312143&doi=10.3390%2fen17225757&partnerID=40&md5=766aedb418cb13039cceb49996d67df2.

[171] Yan, J.-J., Yang, G.-H., and Wang, Y., "Dynamic reduced-order observer-based detection of false data injection attacks with application to smart grid systems," *IEEE Transactions on Industrial Informatics*, vol. 18, no. 10, pp. 6712–6722, 2022. DOI: 10.1109/TII.2022.3144445. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85123721940&doi=10.1109%2fTII.2022.3144445&partnerID=40&md5=0c78173af9f990b989518b551bed8ed3.

[172] Wang, X. and Zolghadri, A., "Covert attack detection and secure control for cyber physical systems," presented at the IFAC-PapersOnLine, Issue: 2, vol. 56, 2023, pp. 7184–7189. DOI: 10.1016/j.ifacol.2023.10.145. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85183631952&doi=10.1016%2fj.ifacol.2023.10.145&partnerID=40&md5=090a25dac6e0eec712112c301eaec355.

[173] Dutta, R., Hu, Y., Yu, F., Zhang, T., and Jin, Y., "Design and analysis of secure distributed estimator for vehicular platooning in adversarial environment," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 4, AR, 2022. DOI: 10.1109/TITS.2020.3036376. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85097165996&doi=10.1109%2fTITS.2020.3036376&partnerID=40&md5=b75df2e845690faee5a38783c9283069.

[174] Hwang, S., Cho, M., Wu, G., and Hwang, I., "Resilient tracking control for leader-follower multi-agent systems against sinusoidal sensor attacks: An LMI-based framework," *IEEE Control Systems Letters*, 2025. DOI: 10.1109/LCSYS.2025.3576284. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-105007304551&doi=10.1109%2fLCSYS.2025.3576284&partnerID=40&md5=119e3a6deca42f76cc979dde6528bdf1.

[175] Wang, J., Zhang, K., Gao, Q., and Peng, P., "Worst-case stealthy linear attacks on distributed kalman filter under kullback-leibler divergence," presented at the Chinese Control Conference, CCC, vol. 2023-July, 2023, pp. 5168–5173. DOI: 10.23919/CCC58697.2023.10240621. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85175527675&doi=10.23919%2fCCC58697.2023.10240621&partnerID=40&md5=6f9f87c78320d3082c12ae9967ce30cf.

[176] Jin, K. and Ye, D., "False-data injection attacks against distributed filters with improved complete-stealthiness," *IEEE Transactions on Industrial Informatics*, vol. 20, no. 8, pp. 10 344–10 353, 2024. DOI: 10.1109/TII.2024.3393501. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85192748956&doi=10.1109%2fTII.2024.3393501&partnerID=40&md5=b05ffee1dcaf47b8d82d88b3cc3b4e4d.

[177] F. Miao and Q. Zhu, "A moving-horizon hybrid stochastic game for secure control of cyber-physical systems," in *53rd IEEE Conference on Decision and Control*, Journal Abbreviation: 53rd IEEE Conference on Decision and Control, Dec. 15, 2014, pp. 517–522. DOI: 10.1109/CDC.2014.7039433.

[178] D. Shi, R. J. Elliott, and T. Chen, "On finite-state stochastic modeling and secure estimation of cyber-physical systems," *IEEE Transactions on Automatic Control*, vol. 62, no. 1, pp. 65–80, Jan. 2017. DOI: 10.1109/TAC.2016.2541919.

[179] L. Yang, C. Wen, and T. Wen, "Multilevel fine fingerprint authentication method for key operating equipment identification in cyber-physical systems," *IEEE Transactions on Industrial Informatics*, vol. 19, no. 2, pp. 1217–1226, Feb. 2023. DOI: 10.1109/TII.2022.3193955.

[180] F. Akbarian, A. Ramezani, M. -T. Hamidi-Beheshti, and V. Haghighat, "Intrusion detection on critical smart grid infrastructure," in *2018 Smart Grid Conference (SGC)*, Journal Abbreviation: 2018 Smart Grid Conference (SGC), Nov. 28, 2018, pp. 1–6.

[181] C. Yang, L. Feng, H. Zhang, S. He, and Z. Shi, "A novel data fusion algorithm to combat false data injection attacks in networked radar systems," *IEEE Transactions on Signal and Information Processing over Networks*, vol. 4, no. 1, pp. 125–136, Mar. 2018.

[182] L. Liang and S. Liu, "Event-triggered distributed attack detection and fault diagnosis," *IEEE Transactions on Instrumentation and Measurement*, vol. 72, pp. 1–11, 2023. DOI: 10.1109/TIM.2022.3216674.

[183] A. J. Gallo, M. S. Turan, F. Boem, T. Parisini, and G. Ferrari-Trecate, "A distributed cyber-attack detection scheme with application to DC microgrids," *IEEE Transactions on Automatic Control*, vol. 65, no. 9, pp. 3800–3815, Sep. 2020. DOI: 10.1109/TAC.2020.2982577.

[184] Barboni, A., Rezaee, H., Boem, F., and Parisini, T., "Detection of covert cyber-attacks in interconnected systems: A distributed model-based approach," *IEEE Transactions on Automatic Control*, vol. 65, no. 9, pp. 3728–3741, 2020.

[185] Wang, Y., Li, Y., Parisini, T., and Zhao, D., "Resilient distributed control for uncertain nonlinear interconnected systems under replay cyber-attacks," *IEEE Transactions on Automatic Control*, 2025.

[186] Akbari, E., Tabatabaei, S., Barkhordari Yazdi, M., Arefi, M., and Cao, J., "Resilient backstepping control for a class of switched nonlinear time-delay systems under hybrid cyber-attacks," *Engineering Applications of Artificial Intelligence*, vol. 122, 2023. DOI: 10.1016/j.engappai.2023.106128. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85150417800&doi=10.1016%2fj.engappai.2023.106128&partnerID=40&md5=a61cb010c8c705b3f07b36834e237ff0.

[187] Chen, W., Lin, F., and Wang, L., "Simultaneous detection and estimation of false data injection attacks in cyber-physical battery systems using a learning observer," presented at the 2023 International Conference on Control, Automation and Diagnosis, ICCAD 2023, 2023. DOI: 10.1109/ICCAD57653.2023.10152377. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85164165056&doi=10.1109%2fICCAD57653.2023.10152377&partnerID=40&md5=b643a5098f638c12752721aec8b00553.

[188] Y. Xuan and M. Naghnaeian, "Detection and identification of CPS attacks with application in vehicle platooning: A generalized luenberger approach," in *2021 American Control Conference (ACC)*, Journal Abbreviation: 2021 American Control Conference (ACC), May 25, 2021, pp. 4013–4020. DOI: 10.23919/ACC50511.2021.9483074.

[189] M. Kazemi and K. Khorasani, "Detection and identification of cyber-attacks in switched cyber-physical systems," in *2024 10th International Conference on Control, Decision and Information Technologies (CoDIT)*, Journal Abbreviation: 2024 10th International Conference on Control, Decision and Information Technologies (CoDIT), Jul. 1, 2024, pp. 738–744. DOI: 10.1109/CoDIT62066.2024.10708518.

[190] Lu, G., Li, K., and Li, Y., "Predefined time fuzzy adaptive tracking control for nonlinear cyber physical systems with unmeasurable states and FDI attacks," *Nonlinear Dynamics*, vol. 113, no. 4, pp. 3389–3404, 2025. DOI: 10.1007/s11071-024-10374-8. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85205588776&doi=10.1007%2fs11071-024-10374-8&partnerID=40&md5=ac884a33b153d3971d058d6fb5ad675b.

[191] De, S. and Sodhi, R., "A simple cyber attack detection scheme for smart grid cyber security enhancement," presented at the 2020 21st National Power Systems Conference, NPSC 2020, 2020. DOI: 10.1109/NPSC49263.2020.9331837. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85100891740&doi=10.1109%2fNPSC49263.2020.9331837&partnerID=40&md5=f10bdaee625f883eb99a06225fc6b2b0.

[192] Palleti, V., Tan, Y., and Samavedham, L., "A mechanistic fault detection and isolation approach using kalman filter to improve the security of cyber physical systems," *Journal of Process Control*, vol. 68, pp. 160–170, 2018. DOI: 10.1016/j.jprocont.2018.05.005. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85048792612&doi=10.1016%2fj.jprocont.2018.05.005&partnerID=40&md5=b5c2176cd50285bdef99991000dc7f2b.

[193] Y. Li and T. Chen, "Stochastic detector against linear deception attacks on remote state estimation," in *2016 IEEE 55th Conference on Decision and Control (CDC)*, Journal Abbreviation: 2016 IEEE 55th Conference on Decision and Control (CDC), Dec. 12, 2016, pp. 6291–6296. DOI: 10.1109/CDC.2016.7799237.

[194] Yu, F., Dutta, R., Zhang, T., Hu, Y., and Jin, Y., "Fast attack-resilient distributed state estimator for cyber-physical systems," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 39, no. 11, pp. 3555–3565, 2020. DOI: 10.1109/TCAD.2020.3013072. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85096037132&doi=10.1109%2fTCAD.2020.3013072&partnerID=40&md5=948001f6ac6baff5cf1247daaaa84e22.

[195] Weng, P., Chen, B., Dong, S., and Yu, L., "Fusion estimation for FDI sensor attacks in distributed systems," presented at the IEEE International Conference on Control and Automation, ICCA, vol. 2020-October, 2020, pp. 260–265. DOI: 10.1109/ICCA51439.2020.9264334. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85098058586&doi=10.1109%2fICCA51439.2020.9264334&partnerID=40&md5=1e0205eb56c1cc8e7801977f47729d1b.

[196] Torabi, A., Zarei, J., Razavi-Far, R., and Saif, M., "Decentralized resilient output-feedback control design for networked control systems under denial-of-service," *IEEE Systems Journal*, vol. 16, no. 4, pp. 5620–5629, Dec. 2022. Accessed: Aug. 21, 2025.

[197] Zhang, K., Hu, S., and Jiang, B., "Sliding mode integral observers for sensor faults detection and isolation in nonlinear systems," 2007.

[198] Guha, D., Roy, P., and Banerjee, S., "Fractional-order cyber-tolerant frequency controller for renewables-integrated power system," presented at the 2023 IEEE 3rd International Conference on Smart Technologies for Power, Energy and Control, STPEC 2023, 2023. DOI: 10.1109/STPEC59253.2023.10430913. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85186518164&doi=10.1109%2fSTPEC59253.2023.10430913&partnerID=40&md5=74b6f4e63e7d6aadd06af03b5aabf399.

[199] Li, J. and Zhang, Y., "FDI attack detection based on l2 gain performance adaptive integral sliding mode observer," presented at the 2024 IEEE 7th International Conference on Industrial Cyber-Physical Systems, ICPS 2024, 2024. DOI: 10.1109/ICPS59941.2024.10640058. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85203721438&doi=10.1109%2fICPS59941.2024.10640058&partnerID=40&md5=a1e0ba2b3d9d02c27af204875d8c6be7.

[200] Taoufik, A., Defoort, M., Busawon, K., Dala, L., and Djemai, M., "A distributed observer-based cyber-attack identification scheme in cooperative networked systems under switching communication topologies," *Electronics (Switzerland)*, vol. 9, no. 11, pp. 1–20, 2020. DOI: 10.3390/electronics9111912. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85096069234&doi=10.3390%2felectronics9111912&partnerID=40&md5=d0477de9cf9873b6757bfafd667df636.

[201] Li, M., Chen, Y., Ma, J., and Liu, Y., "Reduced-order-asynchronous-based control for nonlinear jump networked systems with attacks," *Information Sciences*, vol. 577, pp. 180–195, 2021. DOI: 10.1016/j.ins.2021.07.013. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85109952823&doi=10.1016%2fj.ins.2021.07.013&partnerID=40&md5=4a98de4901514ae23aedd3c827799c80.

[202] Wu, J., Peng, C., Zhang, J., and Tian, E., "A sampled-data-based secure control approach for networked control systems under random dos attacks," *IEEE Transactions on Cybernetics*, vol. 54, no. 8, pp. 4841–4851, 2024.

[203] Li, D., Li, D., Liu, J., Song, Y., and Ji, Y., "Backstepping sliding mode control for cyber-physical systems under false data injection attack," presented at the 2022 IEEE International Conference on Mechatronics and Automation, ICMA 2022, 2022, pp. 357–362. DOI: 10.1109/ICMA54519.2022.9855988. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85137821864&doi=10.1109%2fICMA54519.2022.9855988&partnerID=40&md5=8a96153615439abc1513fc716e4cf156.

[204] Taheri, M., Khorasani, K., Shames, I., and Meskin, N., "Cyberattack and machine-induced fault detection and isolation methodologies for cyber-physical systems," *IEEE Transactions on Control Systems Technology*, vol. 32, no. 2, pp. 502–517, 2024. DOI: 10.1109/TCST.2023.3324870. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85181843630&doi=10.1109%2fTCST.2023.3324870&partnerID=40&md5=ff2b8bae9fbaaebde4798fca4b4d68dd.

[205] Barboni, A., Rezaee, H., Boem, F., and Parisini, T., "Detection of covert cyber-attacks in interconnected systems: A distributed model-based approach," *IEEE Transactions on Automatic Control*, vol. 65, no. 9, pp. 3728–3741, Sep. 2020. Accessed: Aug. 21, 2025.

[206] Y. Shi, C. Liu, H. Chen, and B. Jiang, "Intelligent detection of distinguishability between faults and attacks in cyber-physical DC microgrids," in *2023 6th International Conference on Robotics, Control and Automation Engineering (RCAE)*, Journal Abbreviation: 2023 6th International Conference on Robotics, Control and Automation Engineering (RCAE), Nov. 3, 2023, pp. 371–376. DOI: 10.1109/RCAE59706.2023.10398817.

[207] Jabbarnejad, A., Vaez-Zadeh, S., and Eslahi, M., "An Analytical Parameter-Free Cyberattack Detection Method for Grid-Connected Converters," *IEEE Transactions on Power Electronics*, vol. 39, no. 12, pp. 15 770–15 784, 2024. DOI: 10.1109/TPEL.2024.3430542. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85199505221&doi=10.1109%2fTPEL.2024.3430542&partnerID=40&md5=0b81446d2525fabe9b06c921efc0fac8.

[208] Machina, V., Koduru, S., Madichetty, S., Mishra, S., and Kamel, A., "Sensor attack detection and mitigation using hybrid physics informed neural networks - a real-time implementation for DC-DC converter," *IEEE Transactions on Industry Applications*, vol. 60, no. 4, pp. 5702–5713, 2024. DOI: 10.1109/TIA.2024.3383806.

[209] Zhao, L., Li, W., and Li, Y.-J., "Dual security control based on fusion of mechanism analysis and data-driven under adaptive event-triggered communication scheme for ICPS," *Kongzhi yu Juece/Control and Decision*, vol. 39, no. 1, pp. 206–218, 2024. DOI: 10.13195/j.kzyjc.2022.0557.

[210] Yang, C. and Che, W.-W., "Observer-based attack detection and data-driven resilient control for heterogeneous nonlinear nonaffine mass under FDI attacks," *International Journal of Robust and Nonlinear Control*, vol. 35, no. 3, pp. 1079–1090, 2025. DOI: 10.1002/rnc.7699.

[211] Scruggs, D. et al., "Neural network-based cyber-threat detection strategy in four motor-drive autonomous electric vehicles," *IEEE Access*, vol. 12, pp. 124 220–124 230, 2024.

[212] Qi, R., Zheng, J., Luo, Z., and Li, Q., "A novel unsupervised data-driven method for electricity theft detection in AMI using observer meters," *IEEE Transactions on Instrumentation and Measurement*, vol. 71, 2022. DOI: 10.1109/TIM.2022.3189748.

[213] Ye, G., Shen, L., Feng, Y., Fan, L., Zhang, C., and Li, Y., "Data-driven-based event-triggered resilient control for cigarette weight control system under denial of service attacks and false data injection attacks," *IEEE Access*, vol. 12, pp. 53 633–53 645, 2024. DOI: 10.1109/ACCESS.2024.3384413.

[214] Li, Y., Li, G., and Li, W., "Research on comprehensive security control of ICPS based on DoS attacks energy-grading under ADETCS," presented at the Proceedings of 2023 CAA Symposium on Fault Detection, Supervision and Safety for Technical Processes, SAFEPROCESS 2023, 2023.

[215] Wang, X., Luo, X., Pan, X., and Guan, X., "Detection and location of bias load injection attack in smart grid via robust adaptive observer," *IEEE Systems Journal*, vol. 14, no. 3, pp. 4454–4465, 2020. DOI: 10.1109/JSYST.2020.2967126. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85090902341&doi=10.1109%2fJSYST.2020.2967126&partnerID=40&md5=d70acb7f44c636423cfb97715de28f29.

[216] M. Badr, H. A. Talebi, and M. A. Khosravi, "A novel approach for discriminating faults and replay attacks in hybrid systems," *IEEE Access*, vol. 12, pp. 40 064–40 074, 2024. DOI: 10.1109/ACCESS.2024.3368398.

[217] Zhang, R. C., Li, W., and Li, N., "On the optimal control of network lqr with spatially-exponential decaying structure," in *2023 American Control Conference (ACC)*, IEEE, 2023, pp. 1775–1780.

[218] Grewal, M. S. and Andrews, A. P., *Kalman filtering: Theory and Practice with MATLAB*. John Wiley & Sons, 2014.

[219] Akhlaghi, S., Zhou, N., and Huang, Z., "Adaptive adjustment of noise covariance in kalman filter for dynamic state estimation," in *2017 IEEE power & energy society general meeting*, IEEE, 2017, pp. 1–5.

[220] Fabry, J., Kühl, S., and Jax, P., "On the steady state performance of the kalman filter applied to acoustical systems," *IEEE Signal Processing Letters*, vol. 27, pp. 1854–1858, 2020. [Online]. Available: https://www.semanticscholar.org/paper/bd656f63f7486f5bbed335de8779c5bf0a3ef0db.

[221] Shu, H., Simon, E. P., and Ros, L., "Third-order kalman filter: Tuning and steady-state performance," 11, vol. 20, IEEE, 2013, pp. 1082–1085.

[222] Assimakis, N. and Adam, M., "Iterative and algebraic algorithms for the computation of the steady state kalman filter gain," *International Scholarly Research Notices*, vol. 2014, pp. 1–10, 2014. [Online]. Available: https://www.semanticscholar.org/paper/90a08ecb296a78ba37595e75ca98f54697099df4.

[223] Dehghannasiri, R., Esfahani, M. S., and Dougherty, E. R., "Intrinsically bayesian robust kalman filter: An innovation process approach," *IEEE Transactions on Signal Processing*, vol. 65, no. 10, pp. 2531–2546, 2017.

[224] Boutayeb, M., Rafaralahy, H., and Darouach, M., "Convergence analysis of the extended kalman filter used as an observer for nonlinear deterministic discrete-time systems," *IEEE Transactions on Automatic Control*, vol. 42, no. 4, pp. 581–586, 1997. DOI: 10.1109/9.566674.

[225] Boulkroune, B., Geebelen, K., Wan, J., and Nunen, E. van, "Auto-tuning extended kalman filters to improve state estimation," in *2023 IEEE Intelligent Vehicles Symposium (IV)*, IEEE, 2023, pp. 1–6.

[226] Kwak, S. G. and Kim, J. H., "Central limit theorem: The cornerstone of modern statistics," *Korean journal of anesthesiology*, vol. 70, no. 2, p. 144, 2017.

[227] Miura, C. K., "Tests for the mean of the inverse gaussian distribution," *Scandinavian Journal of Statistics*, pp. 200–204, 1978.

[228] Bullo, F., *Lectures on network systems*. Kindle Direct Publishing Seattle, DC, USA, 2020, vol. 1.

[229] Fawzi, H., Tabuada, P., and Diggavi, S., "Secure estimation and control for cyber-physical systems under adversarial attacks," *IEEE Transactions on Automatic control*, vol. 59, no. 6, pp. 1454–1467, 2014.

[230] Shoukry, Y. and Tabuada, P., "Event-triggered state observers for sparse sensor noise/attacks," 8, vol. 61, IEEE, 2015, pp. 2079–2091.

[231] Mo, Y. et al., "Cyber–physical security of a smart grid infrastructure," *Proceedings of the IEEE*, vol. 100, no. 1, pp. 195–209, 2011.

[232] Chen, Y., Kar, S., and Moura, J. M., "Resilient distributed estimation: Sensor attacks," *IEEE Transactions on Automatic Control*, vol. 64, no. 9, pp. 3772–3779, 2018.

[233] Edwards, C. and Spurgeon, S. K., *Sliding mode control: theory and applications*. Crc Press, 1998.

[234] Hui, S., Sudhoff, S. D., and Zak, S. H., "On estimating regions of stability of the estimation error of sliding mode observers for uncertain systems," in *2006 American Control Conference*, IEEE, 2006, 6–pp.

[235] Utkin, V. I., *Sliding modes in control and optimization*. Springer Science & Business Media, 2013.

[236] Tan, C. P. and Edwards, C., "An lmi approach for designing sliding mode observers," *International Journal of Control*, vol. 74, no. 16, pp. 1559–1568, 2001.

[237] Utkin, V. I., *Sliding modes in optimization and control problems*, 1992.

Information Freshness in Adversarial Multi-Agent Systems

Chapter 6

Introduction to Information Freshness in Adversarial Multi-Agent Systems

In the first part of this thesis, we developed a methodology for sensor attack detection and reconstruction in networked control systems. Under suitable structural and observability assumptions, we further derived a reconstruction procedure that can, in principle, recover the nominal state despite adversarial interference. However, in many practical scenarios the attacker cannot be physically removed, quarantined, or blocked—for example, due to legacy constraints, safety requirements, or limited actuation over the communication stack. Consequently, even after successful detection and isolation, the adversary's presence may continue to degrade performance.

This part of the thesis addresses precisely this gap: we develop a *mitigation* framework that limits the impact of an identified—but unremovable— strategic adversary. The goal is *reconfigure communication policies* so that the residual effect of the attack on monitoring and control remains bounded and possibly acceptable.

In industrial control architectures, information flows through a hierarchy of layers that bridge the physical process and the supervisory decision-making level. At the field layer, dense networks of sensors and actuators interface directly with the process, measuring variables such as temperatures, pressures, or flow rates and enforcing low-level control laws through *Programmable Logic Controllers* (PLCs). Above this layer, the *Supervisory Control and Data Acquisition* (SCADA) system aggregates measurements from distributed PLCs, monitors plant-wide performance, and supports maintenance and optimization tasks. Although it does not act directly on the process, the supervisory layer depends critically on the timely and reliable delivery of data from the field network to maintain an accurate representation of the system state.

The efficiency and safety of this information hierarchy rely on the freshness of the data exchanged between its layers. Each sensor reading represents a snapshot of the process at the time of acquisition, yet by the time it reaches the control center it may already be outdated due to network delays, congestion, or packet loss. When communication is unreliable or bandwidth-limited, the supervisory system's perception of the process diverges from reality, degrading the performance of monitoring, control, and fault-detection functions.

Classical communication metrics—such as throughput, latency, or jitter—capture network performance but not necessarily the operational relevance of the transmitted data. A message may arrive with low delay yet still represent an obsolete measurement, while frequent transmissions can carry redundant information that offers little improvement in situational awareness. To address this limitation, the concept of *information freshness*, formalized by the *Age of Information* (AoI), was introduced as a higher-level performance metric that quantifies the time elapsed since the generation of the most recently received and useful update [54]. AoI and its extensions link communication dynamics with control performance, providing a principled way to reason about how network behavior affects the timeliness and utility of information.

The importance of freshness becomes evident in distributed automation scenarios such as smart manufacturing, remote process supervision, and large-scale IoT ecosystems. In these contexts, supervisory systems must integrate asynchronous data streams from hundreds of heterogeneous sensors operating over shared or wireless networks. Scheduling, prioritization, and transmission policies determine which subsystems remain "visible" in real time, directly influencing estimation accuracy and responsiveness. Maintaining a low average age of information is thus essential to ensure consistent state awareness and reliable decision-making.

The problem grows more complex when the communication network is exposed to malicious interference. Modern industrial control systems are increasingly interconnected with corporate networks and cloud infrastructures, exposing them to cyber threats such as denial-of-service , jamming, or false-data injection attacks. Unlike random losses, these disruptions are intentional: attackers may deliberately delay or block updates to increase staleness, inject falsified data to mislead operators, or intercept transmissions to gain strategic advantage. Under these conditions, traditional latency-based metrics fail to distinguish between "fresh but incorrect" and "stale but trustworthy" information. A comprehensive security-aware analysis must therefore account for the interplay between timeliness, accuracy, and confidentiality.

This chapter investigates *information freshness under adversarial conditions*, focusing on how network imperfections and malicious actions affect the temporal validity of the data available to supervisory systems. Building on the ADI tools developed earlier, our objective is twofold: (i) to quantify the degradation of freshness caused by attacks and delays, and (ii) to design mitigation policies—at the estimator and communication layers—that preserve timely situational awareness even when an identified adversary cannot be blocked. Such policies are essential for the next generation of distributed industrial systems, where control decisions increasingly depend on remote sensing and cloud-based analytics.

Relation to the Author's Publications. The conceptual formulation of *information freshness under adversarial conditions* originate from the study reported in [J2]. This work established the first game-theoretic link between data timeliness and correctness, characterizing equilibrium strategies for rational attackers and legitimate transmitters in cyber–physical systems. The underlying modeling approach, based on continuous-time Markov chains and strategic optimization, was initially presented and validated through the conference contribution [C8], where the AoII metric was formally introduced and analyzed.

Complementary results were disseminated through the conference papers [C15] and [C14], both of which refined the stochastic characterization of adversarial interference and explored its impact on freshness-aware decision-making in multi-sensor networks. These studies formalized the analytical tractability of AoII-based modeling and its applicability to a wide range of adversarial scenarios, including false data injection, denial-of-service, and substitution attacks.

Subsequent extensions of this framework were developed in the journal article [J3], currently under review, which introduces data drift and dynamic attack adaptation into the correctness model. In that study, three representative cyber attack classes—*data substitution*, *denial of service (DoS)*, and *data deletion*—are systematically compared to evaluate their impact on model integrity, detection accuracy, and system resilience over time.

The multi-agent and distributed extensions of the proposed framework draw upon the collaborative works [C13] and [C12], which investigated strategic communication and sensing under resource constraints and shared network conditions. In particular, [C12] demonstrated the use of AoII in multi-source coordination games, highlighting the trade-offs among correctness, freshness, and channel occupancy in adversarial settings.

Organization of the chapter. The remainder of this chapter is structured as follows:

- **Section 6.1** establishes the theoretical background for the study of information freshness in cyber–physical and networked control systems. It introduces the principal performance metrics used to quantify the timeliness, correctness, and secrecy of information, with particular emphasis on how these quantities evolve in adversarial settings.

- **Section 6.2** situates the thesis within this broader landscape. Building upon the literature analysis, it introduces the concept of the *Age of Incorrect Information* as a unifying measure that jointly captures timeliness and correctness, and it outlines the game-theoretic framework developed in this work to model strategic adversarial interaction. The section summarizes the main scientific contributions, emphasizing the analytical tractability, interpretability, and generality of the proposed approach.
- **Section 6.3** formalizes the mathematical tools and assumptions employed throughout the thesis. It introduces the stochastic models that describe the evolution of correctness, the incorporation of adversarial actions, and the resulting strategic equilibrium formulations. This section serves as the methodological bridge between the conceptual foundations presented earlier in the chapter and the detailed analytical developments that follow in later parts of the thesis.

6.1 Literature Review on Information Freshness under Adversarial Conditions

The capacity of a system to operate effectively in dynamic environments depends critically on how promptly it can acquire, process, and disseminate information. In modern networked and cyber-physical infrastructures, decisions are increasingly made based on real-time data flows rather than static models. This paradigm shift has elevated the concept of *information freshness*—the degree to which received information reflects the current state of the system—from a secondary performance indicator to a central design concern. Unlike latency or throughput, freshness captures a temporal relationship between the physical world and its digital representation, quantifying not only how fast data travels, but how up-to-date it remains once it arrives.

The notion of freshness is particularly salient in environments where actions are tightly coupled to observations, such as distributed control, autonomous mobility, and edge intelligence. In these systems, outdated information can degrade situational awareness, induce instability, or propagate erroneous decisions. Maintaining freshness is therefore both a technical and epistemic challenge: it concerns not only the design of efficient communication and computation processes, but also the preservation of trust in what is known about the system at any given moment. As devices and agents act upon streaming data, the timeliness of information becomes a form of control—an operational property that determines how accurately the cyber layer can mirror and influence the physical one.

This challenge acquires new depth in the presence of adversaries. Modern networks operate in contested environments where communication links may be disrupted, monitored, or manipulated. Under such conditions, freshness is no longer degraded merely by stochastic delays or congestion, but by intentional interference. An adversary may selectively jam transmissions, inject stale or falsified updates, or infer private states through timing analysis. Consequently, the concept of information freshness must evolve beyond purely temporal considerations to encompass aspects of *integrity*, *confidentiality*, and *resilience*. This expansion bridges traditionally separate fields—such as control theory, information security, and network optimization—into a unified discourse around the *trustworthiness of real-time information*.

The growing literature on this topic reflects an interdisciplinary convergence. Queueing-theoretic studies formalize freshness through stochastic age processes; control-theoretic works treat freshness as a feedback property; and security-focused analyses reinterpret freshness as a function of exposure or leakage. Collectively, these perspectives transform freshness from a descriptive measure of delay into a normative design criterion—one that defines how systems ought to behave under uncertainty, contention, and attack.

Understanding how information freshness can be maintained, quantified, and optimized under adversarial influence thus becomes essential for the design of next-generation networked systems. The following sections examine how this concept has been formalized through different classes of metrics, how agent interactions shape its optimization, and

how algorithmic strategies have evolved to balance freshness with other critical system objectives in collaborative, competitive, and adversarial environments.

6.1.1 Metrics Classification

The quantification of information freshness has become a cornerstone in the analysis and optimization of modern communication and cyber-physical systems. A broad family of metrics has emerged to capture how timely and relevant the information available at a destination remains over time. These metrics range from the classical *Age of Information* to more advanced formulations that account for context, semantics, and even adversarial interference. In adversarial settings, freshness can no longer be viewed as a purely temporal quantity: it must also incorporate aspects of integrity, confidentiality, and resilience against malicious manipulation. This section reviews and classifies the principal families of metrics used to represent and optimize freshness in both benign and hostile environments, emphasizing their mathematical foundations and their relevance to real-world systems.

Classical Age-Based Metrics The foundational concept in this domain is the *Age of Information* , which measures the staleness of information at a receiver relative to the generation time of the most recently received update. Let t denote the current time and $u(t)$ the generation time of the latest received packet. The instantaneous AoI is defined as:

$$\Delta(t) = t - u(t). \tag{6.1}$$

Over time, AoI increases linearly and drops when a new update is received, producing the characteristic sawtooth pattern used in many analytical models. The *time-averaged AoI* over an interval $[0, T]$ is:

$$\bar{\Delta} = \frac{1}{T} \int_0^T \Delta(t)\, dt, \tag{6.2}$$

while the *Peak AoI*—a measure of the maximum information staleness before an update—can be expressed as:

$$\Delta_{\text{peak}}^{(n)} = \max_{t \in (t_{n-1}, t_n)} \Delta(t), \tag{6.3}$$

where t_n is the reception time of the n-th update.

These classical AoI formulations remain the backbone of freshness analysis in wireless telemetry, UAV communication [238], networked control [239], and secure telemetry systems [52], [240]. However, they assume benign environments and do not capture intentional delay or data corruption caused by adversarial actors. This limitation has led to several extensions that adapt AoI to more complex and security-aware scenarios.

Context-Aware and Demand-Driven Metrics In many modern systems, information is not continuously consumed but rather accessed on demand, or its value depends on contextual factors such as task urgency or user intent. To account for these aspects, several context-aware and demand-driven metrics extend the classical AoI framework.

The *Query Age of Information (Q-AoI)* [241], [242] is designed for pull-based systems where data is requested by queries rather than pushed periodically. It measures the age of the most recently available update at the time of query:

$$\Delta_Q(t_q) = t_q - u(t_q), \tag{6.4}$$

where t_q is the query time and $u(t_q)$ denotes the generation time of the latest update available by then. Q-AoI is particularly relevant in cloud-assisted monitoring and database-driven IoT systems, where information freshness is meaningful only at query instants.

The *Urgency of Information (UoI)* [243], [244] introduces a decay function to model how the relevance of information decreases with time:

$$U(t) = \gamma(\Delta(t)), \tag{6.5}$$

where $\gamma(\cdot)$ is a non-increasing function, typically modeled as an exponential decay $\gamma(\delta) = e^{-\lambda\delta}$. This metric captures scenarios where delayed updates quickly lose their value, such as event detection or fault monitoring.

The *Value of Information (VoI)* [245], [246] incorporates data semantics into freshness measurement, weighting the age of data by its intrinsic utility:

$$V(t) = \phi(x(t)) \cdot \gamma(\Delta(t)), \tag{6.6}$$

where $x(t)$ represents the information content and $\phi(x)$ denotes its associated utility. VoI has become particularly important in distributed and resource-constrained systems, where bandwidth or energy limitations require prioritization of high-utility updates—an issue that becomes even more pressing under adversarial interference, where communication resources must be used strategically to maximize system performance.

Security-Oriented and Adversarial Metrics In adversarial environments, freshness must account not only for how recent information is but also for whether it is trustworthy, secure, and uncompromised. This has led to the formulation of several security-augmented AoI metrics.

The *Secrecy AoI* quantifies the age of the most recent update that remains confidential to an adversary:

$$\Delta_s(t) = t - u_s(t), \tag{6.7}$$

where $u_s(t)$ is the generation time of the latest securely received update. The *Age of Leaked Information (AoLI)* measures the freshness of information from the adversary's perspective:

$$\Delta_e(t) = t - u_e(t), \tag{6.8}$$

where $u_e(t)$ denotes the generation time of the last intercepted or compromised message. Finally, the *Age of Incorrect Information (AoII)* evaluates the time elapsed since the receiver last held an accurate system state:

$$\Delta_i(t) = t - u_c(t), \tag{6.9}$$

where $u_c(t)$ corresponds to the generation time of the last verified correct update.

These adversarially aware metrics are essential in contexts such as jamming-resistant communication [247], systems under false data injection [248], and privacy-preserving IoT applications [240], [249]. They allow system designers to quantify not only freshness from a legitimate user's viewpoint but also the potential informational advantage gained by an adversary—an increasingly relevant consideration in secure CPS design.

Collaborative and Distributed Metrics In distributed and multi-agent systems, information freshness is inherently a collective property rather than an individual one. Agents often share overlapping or correlated observations, making it necessary to consider metrics that jointly reflect freshness across the network.

The *Correlation-Aware AoI (CAAoI)* captures redundancy among sources by weighting each node's AoI according to its correlation with others:

$$\text{CAAoI}(t) = \sum_{i \in \mathcal{S}} w_i(t) \cdot \Delta_i(t), \quad w_i(t) = 1 - \max_{j \neq i} \rho_{ij}, \tag{6.10}$$

where ρ_{ij} represents the correlation coefficient between data streams i and j. The *Age of Collection (AoC)* applies to mobile data-gathering scenarios and measures, for each source i, the time elapsed since the last sampling by the collector:

$$\text{AoC}_i(t) = t - s_i, \tag{6.11}$$

where s_i is the last sampling time. The *Age of Federated Information (AoFI)* is used in federated learning or collaborative estimation systems, representing the average freshness of updates aggregated from a selected subset of contributors:

$$\text{AoFI}(t) = \frac{1}{|\mathcal{K}|} \sum_{i \in \mathcal{K}} \Delta_i(t), \tag{6.12}$$

where $\mathcal{K} \subseteq \mathcal{S}$ denotes the set of participating nodes. Finally, *Freshness Variance* measures fairness in freshness distribution:

$$\sigma^2(t) = \frac{1}{N} \sum_{i=1}^{N} \left(\Delta_i(t) - \mu(t)\right)^2, \quad \mu(t) = \frac{1}{N} \sum_{i=1}^{N} \Delta_i(t). \tag{6.13}$$

These distributed formulations are vital for coordinated decision-making in federated learning, collaborative UAV operations, and underwater sensor networks [246], [250], [251], [252]. They are particularly relevant in adversarial environments, where correlated redundancy and cooperative updates can increase robustness against localized attacks or communication failures.

Multi-Objective Optimization with Information Freshness Although AoI is a fundamental metric, it is rarely optimized in isolation. In practical systems, freshness interacts with other performance criteria such as energy consumption, latency, control quality, secrecy, and learning convergence. Multi-objective formulations are therefore central to designing real-world systems that balance competing goals under resource constraints and potential adversarial threats.

The *AoI–Energy* trade-off captures the conflict between frequent transmissions, which improve freshness, and limited energy budgets in wireless or UAV networks [52], [253], [254]. The *AoI–Latency* trade-off recognizes that minimizing delay does not necessarily minimize AoI, particularly in queue-based or bursty traffic systems [247], [255]. In networked control systems, AoI is co-optimized with control performance metrics such as mean-square error or control cost to ensure stable and responsive behavior even under uncertain or adversarial delays [245], [256]. The *AoI–Secrecy* trade-off becomes crucial in secure communications, balancing timely updates with confidentiality under eavesdropping risk [240], [249], [257]. In federated learning, the *AoI–Learning* trade-off addresses the tension between model convergence and communication cost [242], [246], [251], while the *AoI–Economic* trade-off links freshness optimization to incentive mechanisms in data markets [241].

These formulations are solved using a variety of analytical and algorithmic approaches, including Lyapunov optimization, convex programming, constrained Markov decision processes, game-theoretic analysis, and reinforcement learning. Each provides a structured framework to capture how information freshness interacts with system constraints and adversarial uncertainty. Together, they illustrate that freshness is no longer a passive performance indicator but an active system variable—one that must be managed jointly with other objectives to ensure efficiency, resilience, and security in modern cyber-physical and networked environments.

6.1.2 Models of Agent Interaction

The dynamics of information freshness in networked systems cannot be understood in isolation from the behaviors of the agents that generate, transmit, and consume data. Freshness emerges as a collective property—shaped not only by communication protocols and physical constraints, but by the strategic relationships among interacting entities. In distributed environments, agents may cooperate toward a shared goal, act independently in pursuit of individual objectives, or behave adversarially to disrupt the information flow. These interaction modes define distinct analytical frameworks and system behaviors, ranging from joint optimization and equilibrium analysis to defense against intelligent interference.

Collaborative Interaction Models Collaboration represents the most constructive form of interaction among agents sharing a communication or sensing environment. In this setting, agents coordinate their actions—such as scheduling transmissions, allocating bandwidth, or dividing sensing tasks—to minimize global performance metrics like the network-wide Age of Information. Cooperation transforms what would otherwise be a set of competing, locally optimal processes into a cohesive system capable of sustaining freshness across spatially and functionally distributed components.

This paradigm is particularly valuable when no single agent possesses full knowledge or control of the system. Through information sharing and task division, collaborative agents can collectively achieve timeliness levels unattainable in isolation. In federated learning systems, for example, client devices jointly determine when and how to transmit updates so that both data relevance and representativeness are preserved during aggregation [246]. Similarly, in data marketplaces, cooperative allocation mechanisms balance sensing costs and update frequency to optimize collective freshness under budget constraints [241].

Multi-UAV systems exemplify the benefits of explicit coordination: by jointly planning trajectories and communication schedules, UAVs avoid redundant coverage and reduce overall staleness of collected data. Works such as [251], [253], [258] show that cooperative path design leads to substantial reductions in average AoI across large spatial regions, particularly in missions requiring synchronized coverage or low-latency feedback. Similar effects are observed in collaborative perception for autonomous driving and industrial control, where nodes share sensory updates to maintain consistent situational awareness and prevent stale information from propagating across control loops [259], [260].

Even in severely resource-constrained settings, such as LoRa or underwater sensor networks, collaborative strategies improve freshness by leveraging relay selection and joint parameter adaptation [252], [261]. Cooperation at the edge—through shared computational infrastructure or coordinated spectrum access—further extends this principle, enabling systems to maintain low AoI without overloading communication backbones [262], [263], [264]. Across these contexts, collaboration acts as a stabilizing mechanism: it mitigates interference, balances resource consumption, and reinforces the temporal coherence of distributed information systems.

Competitive Interaction Models Competition arises when agents share communication resources but optimize freshness individually rather than collectively. Unlike collaborative frameworks, where coordination is explicit, competitive systems are characterized by decentralized decision-making and self-interest. Each agent seeks to minimize its own AoI, often unaware—or indifferent—to the resulting impact on others. The absence of coordination can lead to contention, oscillatory scheduling, and inefficiencies, even though every participant acts rationally.

Such dynamics naturally appear in multi-user networks, distributed wireless channels, and edge computing environments where access opportunities are scarce. Agents compete for time slots, transmission power, or spectral resources, and the resulting equilibria reflect a balance of selfish optimization rather than social optimality. These scenarios are typically analyzed through non-cooperative game theory or multi-agent learning frameworks, where the existence and efficiency of equilibria can be studied formally. The classical inefficiency of decentralized behavior is often quantified through the *price of anarchy*, denoting the performance gap between individual equilibrium strategies and the globally optimal configuration.

Representative studies highlight this tension between autonomy and efficiency. In [265], [266], users independently compete for multi-channel access under power limitations, producing interference patterns that elevate average AoI. In [267], agents with uncertain mobility and channel states select transmission schedules that minimize their own Peak AoI but collectively degrade overall freshness. Market-inspired and incentive-driven frameworks, such as those proposed in [241], attempt to reconcile this conflict by embedding mechanism design principles that encourage truthful behavior and resource-efficient

updates. Similar competitive dynamics are studied in learning-based environments, where independent agents adapt transmission or sensing policies based on local rewards [268].

In essence, competitive models reveal the fragility of freshness in decentralized systems: when agents act independently, the network's temporal coherence depends on emergent equilibria rather than centralized coordination. This framework is invaluable for understanding systems where competition is inevitable—such as unlicensed spectrum access, vehicular communication, and self-organizing IoT networks—and where limited cooperation must be engineered through incentives or regulation.

Adversarial Interaction Models Adversarial interaction represents the most disruptive and strategically complex regime of multi-agent freshness dynamics. Here, some agents deliberately act to degrade the freshness, integrity, or confidentiality of information maintained by legitimate entities. The objectives are asymmetric: while the defender seeks to minimize its AoI and preserve accurate state information, the adversary aims to maximize uncertainty, induce misinformation, or exploit timing dependencies to extract knowledge. This asymmetry introduces elements of deception, observation, and control that go beyond the equilibria of competitive systems.

Adversarial interference can manifest in several ways. In jamming scenarios, attackers selectively target transmissions or communication slots to induce stale updates, as demonstrated in [52], [238], [247]. Eavesdropping attacks, explored in [240], [269], [270], exploit transmission timing and channel characteristics to infer system states, leading to the development of metrics such as Secrecy AoI and Age of Leaked Information (AoLI). False data injection and timestamp manipulation introduce subtler threats: as shown in [248] and [245], adversaries can exploit trust mechanisms to reclassify outdated data as fresh or inject corrupted information that propagates within the control loop.

Advanced attacks often combine stealth and adaptation. In covert observation scenarios, adversaries probabilistically monitor packet transmissions to infer sensitive events without overt interference [257]. Legitimate agents respond by enforcing covertness constraints—balancing freshness with undetectability—as studied in [249], [271]. These dynamics turn freshness preservation into a game of timing and concealment, where even the act of transmitting information can reveal private state trajectories.

To capture this complexity, game-theoretic and stochastic decision frameworks are frequently employed. Attacker-defender dynamics have been modeled using static and sequential games [254], [272], [273], while adaptive defense policies are learned through reinforcement and online optimization methods [242], [244]. These approaches formalize the interaction as a feedback process between information flow and strategic interference, revealing that the stability of freshness under attack depends as much on intelligent response as on communication efficiency.

Adversarial models thus extend the study of information freshness beyond the realm of performance optimization. They expose its deeper nature as a dynamic expression of trust, control, and information asymmetry—where the preservation of temporal validity becomes inseparable from the protection of informational integrity.

6.1.3 Solution Strategies

Designing mechanisms that preserve information freshness under dynamic and potentially adversarial conditions requires more than the definition of suitable metrics or the understanding of agent interactions. It demands concrete strategies—architectural, algorithmic, and theoretical—that govern how systems sense, decide, and react over time. Across the literature, these strategies can be broadly interpreted as different forms of *control organization* and *decision methodology*. The former determines how authority and knowledge are distributed among agents; the latter defines how these agents adapt their behavior to uncertain or hostile environments. Together, they form the operational backbone of freshness-aware design.

Architectural Paradigms The way control is structured within a system—whether centralized, decentralized, or distributed—fundamentally determines how freshness can be optimized and protected. Each paradigm embodies a distinct philosophy of coordination and trust, reflecting different trade-offs between performance, scalability, and resilience.

Centralized architectures assume the existence of a global controller with complete visibility of the system state and the ability to coordinate decisions across all entities. They are particularly effective in static or well-connected environments where communication is reliable and latency predictable. Centralized formulations dominate early studies on AoI optimization, where the focus lies on finding globally optimal scheduling or resource-allocation policies under full observability. Examples include UAV-enabled IoT frameworks designed for joint energy–AoI minimization [238], [274], satellite-terrestrial networks for covert communication planning [249], and mobile edge computing (MEC) systems managed through centralized queueing-based optimization [262]. Centralized orchestration is also employed in fog and edge computing systems [260], [275], multi-UAV coordination based on actor–critic reinforcement learning [253], and data acquisition markets operating under mechanism design principles [241]. Despite their near-optimal performance, such architectures remain vulnerable to scalability limits, communication bottlenecks, and single points of failure—vulnerabilities that adversaries can exploit through targeted attacks or misinformation injection.

Decentralized architectures, in contrast, distribute intelligence among autonomous agents that make decisions based on local observations or limited feedback. This design philosophy enhances robustness and privacy, allowing systems to operate even when connectivity is intermittent or trust is partial. Decentralized scheduling has proven effective in federated learning environments, where clients independently manage update timing to balance computational load and convergence speed without exchanging raw data [246]. Similar paradigms guide vehicle platooning [264] and local timestamp defense mechanisms against manipulation attacks [248]. Although decentralization enhances resilience and autonomy, it also introduces new challenges—most notably the lack of global coordination, which can lead to suboptimal freshness distributions and temporal inconsistency among agents.

Distributed architectures occupy an intermediate space between these extremes, combining localized autonomy with structured coordination. Agents exchange partial information—such as model parameters, channel states, or scheduling signals—to jointly optimize a shared freshness objective. Distributed frameworks are particularly suited to large-scale or adversarial scenarios, where central control is impractical yet cooperation is beneficial. Notable examples include deep reinforcement learning–based anti-jamming routing in satellite networks [239], distributed Q-learning for relay and spreading factor selection in LoRa systems [261], and stochastic task allocation in Internet of Everything (IoE) sensing [276]. Hybrid 6G architectures combine centralized training with device-side inference to achieve both coordination and adaptability [255]. Under adversarial stress—such as eavesdropping or false data injection—distributed game-theoretic schemes offer resilience through redundancy and cooperative adaptation [243], [277]. These systems exemplify a broader trend: the migration from monolithic control to adaptive, networked intelligence as a means to sustain freshness and trust in uncertain conditions.

Analytical and Algorithmic Methodologies Beyond architecture, the strategies used to maintain freshness depend on the mathematical and computational tools that govern decision-making. The literature reveals a rich landscape of solution techniques—each reflecting a different epistemological stance toward uncertainty and adversarial behavior.

Game-theoretic frameworks provide a natural language for capturing the strategic interplay between legitimate and malicious agents. In static games of complete information, both defender and adversary select strategies simultaneously under perfect knowledge, as in jamming or eavesdropping models [52], [245], [265], [266], [277]. Sequential formulations, often modeled as Stackelberg games, capture leader–follower dynamics in which the legitimate system acts first and the adversary responds [247], [272], [278]. These

frameworks formalize the idea of anticipatory defense—where freshness is maintained not by reacting to attacks, but by predicting and shaping adversarial behavior. Dynamic and Markov games extend this reasoning to temporally correlated attacks [245], [270], revealing how the co-evolution of strategies affects long-term freshness equilibria. Collectively, game-theoretic models provide a structured foundation for reasoning about control under contestation, enabling explicit trade-offs between timeliness, secrecy, and resource expenditure [238], [256], [269].

Reinforcement learning (RL) approaches, by contrast, replace explicit modeling with adaptive experimentation. Agents learn freshness-preserving policies by interacting with the environment and observing long-term returns, making RL particularly suitable for partially observable or stochastic systems. Centralized RL methods, applied to UAV networks [253], MEC task scheduling [268], or satellite routing [239], exploit full state information for globally coordinated decision-making. Distributed RL variants, such as multi-agent Q-learning, enable autonomous nodes to adapt locally in federated or vehicular networks [246], [261], [264]. More sophisticated deep reinforcement learning (DRL) frameworks extend scalability and generalization through hybrid schemes that blend centralized training with decentralized execution [255]. By learning directly from experience, RL systems can track and counter evolving adversarial tactics—transforming freshness maintenance into a form of continual adaptation.

Constrained optimization techniques occupy the analytical core of freshness research. These models formalize the problem of minimizing AoI subject to physical, energetic, or security-related limitations. They provide closed-form or convex solutions that balance freshness against secrecy, energy, and latency constraints. Representative studies include joint AoI–energy optimization under selective jamming [274], [279], secrecy-constrained scheduling [244], [269], and covert transmission design under detection thresholds [249], [271]. Similar formulations extend to timestamp manipulation [248] and probabilistic eavesdropping [257]. In cooperative contexts, convex and quasi-convex programming frameworks support resource allocation and offloading optimization in fog and edge systems [260], [262], [275]. The appeal of these methods lies in their tractability and interpretability: they yield explicit design rules for freshness preservation even under worst-case adversarial assumptions [267].

Online learning and scheduling approaches extend these ideas to non-stationary environments where system parameters and adversarial strategies evolve over time. By continuously updating decisions in response to observed outcomes, these methods guarantee sub-linear regret with respect to an ideal offline scheduler [242]. They also allow robustness analysis under worst-case dynamics using competitive-ratio frameworks [248], [267]. Drawing on concepts from adversarial bandits and online convex optimization, such strategies emphasize adaptation over prediction—seeking not to foresee the future, but to perform near-optimally despite its uncertainty.

Markov models and queueing theory provide the analytical underpinning for much of the freshness literature. By modeling packet arrivals, service times, and interference as stochastic processes, they allow formal derivations of AoI, AoII, and AoLI under various operational and adversarial conditions [245], [257]. Markov renewal and queueing models support both theoretical insight and validation of algorithmic methods, linking statistical variability to freshness degradation [252], [263]. These frameworks reveal how randomness and contention translate into measurable staleness—and how adversaries can exploit, or defenders can mitigate, these effects through temporal control.

Heuristic and bio-inspired algorithms represent a more exploratory, computationally flexible approach. Techniques such as Ant Colony Optimization (ACO) [258] mimic collective intelligence to design near-optimal trajectories or scheduling sequences without explicit gradients or full models. Their adaptability and scalability make them attractive for large, nonlinear, or combinatorial problems, where traditional optimization may be infeasible. While they lack formal guarantees, their empirical effectiveness in real-time and resource-constrained contexts underscores their practical relevance.

Finally, specialized tools such as *mechanism design* and *martingale theory* bridge game-theoretic and probabilistic reasoning. Mechanism design introduces incentive compatibility

into freshness optimization, ensuring that self-interested agents behave truthfully and efficiently under private cost information [241]. Martingale-based analyses, in turn, provide probabilistic performance guarantees under stochastic task allocation and peak-AoI constraints [276]. Together, these methods illustrate how formal reasoning—economic or probabilistic—can complement algorithmic adaptation, yielding systems that are not only efficient but verifiably fair and reliable.

6.2 Thesis Contributions

This thesis develops a rigorous mathematical framework for analyzing the *strategic interaction* between a legitimate transmitter and a malicious adversary in a networked cyber–physical system. At the core of this framework lies the concept of the *Age of Incorrect Information* (AoII), a metric designed to jointly capture both the *timeliness* and the *correctness* of the information available to a receiver. In supervisory and control applications, decision quality depends critically on the validity and freshness of received data: outdated or corrupted information can propagate erroneous commands, trigger unsafe responses, or erode system efficiency. The transmitter therefore seeks to minimize the AoII by issuing frequent and accurate updates, ensuring that the receiver's view of the system remains both current and trustworthy.

The adversary, fully aware of this objective, acts with the opposite intent. By selectively corrupting, blocking, or substituting packets, it aims to *maximize* the receiver's AoII—prolonging intervals of incorrect information and thereby impairing situational awareness and decision-making accuracy. This antagonistic interaction naturally forms a *strategic game* in which both agents behave rationally but pursue conflicting goals: the transmitter minimizes, while the attacker maximizes, the same information-theoretic performance index. The equilibrium of this game reveals the operating point at which the system achieves robustness against economically rational adversarial behavior.

Within this analytical framework, the thesis investigates several classes of adversarial policies, each reflecting a different mechanism of information compromise:

- **False Data Injection (FDI):** the adversary forges seemingly legitimate updates and injects them into the communication channel, causing the receiver to store erroneous information even when transmission timing remains nominal;
- **Data Substitution (DS):** the attacker intercepts valid packets and replaces them with corrupted counterparts, thereby combining the effects of informational deception and delivery delay;
- **Denial-of-Service (DoS):** the adversary disrupts communication by selectively blocking, jamming, or discarding updates, resulting in information staleness and loss of situational awareness.

Among these, FDI attacks receive particular attention due to their prevalence and sophistication in modern industrial and IoT networks. While DS and DoS attacks are analyzed under a single-sensor configuration—formalized as a *static game of complete information*—the FDI scenario is investigated in greater depth. In this case, the interaction between transmitter and adversary is modeled as a *Bayesian game*, where the transmitter possesses only partial knowledge of the attacker's operational cost, represented through a probability distribution. The framework is further extended to a *multi-sensor* setting that captures resource-constrained architectures in which multiple data sources share a common communication channel. This generalization exposes intrinsic trade-offs between network capacity, information correctness, and resilience to adversarial interference, providing analytical insight into the optimal coordination of communication resources in contested environments.

Furthermore, a *multi-sensor* extension is developed to capture resource-constrained architectures, where multiple sources compete for a shared communication channel. This

configuration exposes fundamental trade-offs between network capacity, information correctness, and adversarial resilience, providing insight into optimal resource allocation in contested environments.

The principal scientific contributions of this thesis are summarized below:

1. *Correctness-aware freshness modeling:* formulation and analytical characterization of the *Age of Incorrect Information* as a joint indicator of data timeliness and semantic validity at the receiver, extending classical freshness metrics to adversarial settings;
2. *Stochastic modeling of correctness dynamics:* development of a continuous-time Markov chain framework describing natural and adversarial transitions between correct and incorrect information states, thereby unifying the effects of deception, drift, and denial-of-service within a single probabilistic formalism;
3. *Game-theoretic analysis of adversarial interaction:* design and solution of static and Bayesian strategic games between the transmitter and the attacker, including explicit derivation of Nash equilibria, feasibility thresholds, and equilibrium utilities for various attack models;
4. *Extension to distributed and resource-constrained architectures:* generalization of the theoretical model to multi-sensor networks with shared communication resources, illustrating how optimal transmission policies balance freshness, correctness, and bandwidth efficiency in adversarial contexts;
5. *Analytical tractability and interpretability:* derivation of closed-form equilibrium strategies and steady-state performance measures that offer actionable design principles for secure, efficient, and resilient information-update mechanisms in cyber–physical systems under attack.

Together, these contributions establish a unified foundation for analyzing and mitigating adversarial impacts on information freshness. They bridge concepts from control theory, communication networks, and game theory into a coherent framework capable of quantifying how rational strategic behavior—both defensive and offensive—shapes the temporal reliability of information in distributed systems.

6.3 Modeling Formalism Used in This Thesis

6.3.1 Problem Setting and Architecture

The system under study consists of two communicating entities that exchange information about the state of a physical process over a networked channel. In an industrial control context, these roles are typically played by a *Programmable Logic Controller (PLC)* acting as the transmitter and a *SCADA* system acting as the receiver. More generally, we refer to them simply as the *transmitter* and the *receiver*, since the subsequent analysis applies to any cyber–physical architecture in which information about a physical process is periodically communicated between distributed components.

The physical process is modeled generically as a nonlinear, time-varying system

$$\dot{\boldsymbol{x}}(t) = f\big(\boldsymbol{x}(t), \boldsymbol{u}(t), t\big) + \boldsymbol{w}(t), \qquad \boldsymbol{y}(t) = h\big(\boldsymbol{x}(t), t\big) + \boldsymbol{v}(t), \tag{6.14}$$

where $\boldsymbol{x}(t)$ is the plant state, $\boldsymbol{u}(t)$ the (local) control input, and $\boldsymbol{y}(t)$ the measured output; $\boldsymbol{w}(t)$ and $\boldsymbol{v}(t)$ capture process and measurement uncertainties. The transmitter operates in close contact with the plant, acquires $\boldsymbol{y}(t)$, and possibly applies $\boldsymbol{u}(t)$ via a local feedback policy; the *only* quantity sent over the network is the measurement (or a function thereof). Let $\{t_k\}$ denote the update times and define the transmitted packets as $\tilde{\boldsymbol{y}}_k \coloneqq \boldsymbol{y}(t_k)$.

The transmitter periodically sends $\tilde{\boldsymbol{y}}_k$ to the receiver, which performs higher-level supervision, diagnostics, or decision support based on the *received* sequence. Before

reaching the receiver, however, packets may be affected by natural distortions, random losses, or deliberate manipulations by an adversary; thus the receiver's view may deviate from the true process state. Overall reliability hinges on two properties of the received data: their *timeliness* (how up-to-date they are) and their *correctness* (whether they reflect the true plant output).

Only the communication channel between transmitter and receiver is considered vulnerable to interference (e.g., injection, jamming, selective blocking). The local actuation loop (transmitter $\rightarrow$ plant) is assumed secure and not subject to external manipulation, reflecting common industrial deployments where fast control is confined to a trusted network while supervisory links may be exposed. Consequently, we focus on attacks that alter what the receiver *perceives* (delay, corruption, substitution) without directly changing the applied control $\boldsymbol{u}(t)$.

We assume the transmission delay is negligible relative to the process time scale: whenever a packet is not blocked or lost, it is considered instantaneously available at the receiver. This isolates the effect of adversarial mechanisms on information *correctness/freshness* from pure latency effects.

Each successful transmission incurs a cost, modeled as a constant $C > 0$ per update. This cost abstracts practical constraints such as energy expenditure (battery-powered devices), bandwidth usage (shared networks), monetary fees (cellular or leased links), and congestion or duty-cycle limitations. In the ensuing analysis, C penalizes overly aggressive update policies, forcing a principled trade-off between keeping information fresh/correct at the receiver and expending scarce communication resources.

6.3.2 Drift Rate: Physical Origin and Modeling Role

The *drift rate* is a key concept introduced to bridge the continuous evolution of the physical process with the discrete perception of information at the supervisory level. In essence, it quantifies how quickly the information stored by the receiver becomes outdated as the underlying process naturally evolves over time. Even in the absence of attacks or communication delays, the physical state of a plant changes continuously, while the receiver relies on discrete-time updates. As a consequence, the last received measurement—though initially accurate—gradually loses relevance as the system departs from the condition it represented. The drift rate captures this intrinsic *information aging* by expressing the average frequency with which stored data cease to reflect the true system state.

Formally, we assume that the information available to the receiver remains valid as long as the process output does not deviate from its last known correct value beyond a prescribed tolerance $\omega > 0$. Let $\boldsymbol{y}(t)$ denote the true output of the plant, and $y(t_u)$ the last correct value known to the receiver. The instant when the deviation $|y(t) - y(t_u)|$ first exceeds the threshold ω marks the loss of correctness. The corresponding random variable

$$T_\omega = \inf\{t > 0 : |y(t_u + t) - y(t_u)| \geq \omega\}$$

represents the time interval during which the stored information remains valid. Assuming that this process follows a memoryless evolution—consistent with many stochastic or approximately Markovian physical systems—T_ω can be modeled as exponentially distributed with rate parameter $d > 0$:

$$\mathbb{P}(T_\omega > t) = e^{-dt}, \qquad \mathbb{E}[T_\omega] = \frac{1}{d}.$$

Under this formulation, the drift rate d serves as the *instantaneous hazard of correctness loss*: the higher its value, the faster the system's physical state diverges from the last known measurement, and hence the shorter the validity of stored information. Conversely, a small drift rate characterizes slow or quasi-stationary processes, where measurements remain representative of the true state for longer periods.

This abstraction is powerful because it allows the complex and potentially nonlinear dynamics of a physical process to be condensed into a single, statistically meaningful

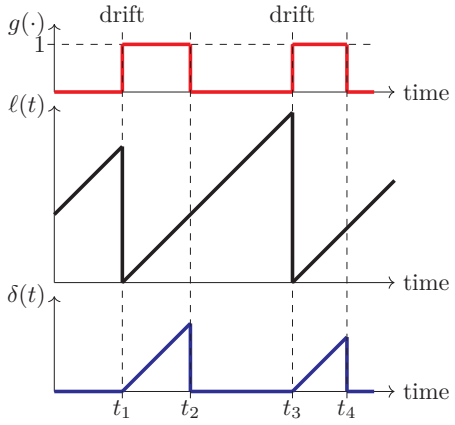


Figure 6.1: AoII dynamics. AoII grows whenever information is incorrect (e.g., at t_1 and t_3), and resets after correct updates (at t_2 and t_4).

parameter. Rather than modeling the full state-space evolution, one only needs to characterize how rapidly information becomes obsolete on average. The drift rate thus provides the fundamental link between physical time evolution and informational freshness, enabling a tractable yet expressive way to describe how natural dynamics—and later, adversarial actions—affect the correctness of supervisory data.

6.3.3 Markov Modeling of Correctness and Attacks

To evaluate how cyberattacks affect the quality of the information available to a supervisory system, it is necessary to quantify not only *how old* the received data are, but also *whether they remain correct*. Classical freshness metrics, such as the *Age of Information*, capture the time elapsed since the last update was received, but implicitly assume that all received packets are correct. This assumption no longer holds in adversarial environments, where an attacker may deliberately corrupt, delay, or block transmissions. For this reason, this thesis adopts the *Age of Incorrect Information* as the central performance metric. The AoII measures the average time during which the receiver's knowledge of the system is *incorrect*—that is, either outdated or wrong due to false data injection. It therefore extends the AoI concept by explicitly incorporating the *correctness dimension*, making it particularly suitable for modeling cyber–physical systems under attack.

Definition 6.3.1 (*Age of Incorrect Information*)

Let $y(t)$ be the true signal measured by a sensor, and $y(t^u)$ the last update at the monitor. Define the discrepancy indicator

$$g\big(y(t), y(t^u)\big) \triangleq \begin{cases} 1, & \text{if } y(t^u) \text{ is outdated or corrupted,} \\ 0, & \text{otherwise,} \end{cases}$$

and let t^d be the last time correct information was available. The penalty is $\ell(t) \triangleq t - t^d$, and the instantaneous AoII is

$$\delta(t) \triangleq \ell(t)\, g\big(y(t), y(t^u)\big). \tag{6.15}$$

□

Figure 6.1 illustrates the temporal behavior of the AoII. The metric increases linearly whenever the receiver holds incorrect information—such as after a process drift or data corruption—and resets to zero upon a correct update. This evolution is consistent with the renewal structure typical of two–state continuous–time Markov chain (CTMC) models [280]. Unlike the classical AoI, which grows continuously with data staleness, the AoII

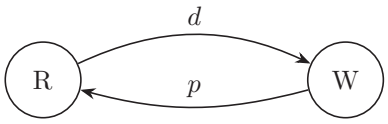


Figure 6.2: Nominal correctness dynamics: transitions from correct (R) to incorrect (W) occur due to process drift at rate d, while updates at rate p restore correctness.

increases *only* while the monitored information is wrong, thus capturing both temporal and semantic degradation.

To formalize this behavior, we model the evolution of information correctness as a two–state CTMC, where the system alternates between periods in which the receiver's knowledge is correct and periods in which it is incorrect.

- **Right (R)** — the information stored at the receiver correctly reflects the plant's current state;
- **Wrong (W)** — the stored information is incorrect, either because the physical process has changed (natural drift) or because an adversary has interfered with communication.

Nominal case. In the absence of adversarial interference, correctness naturally degrades as the physical process evolves, while each successful transmission restores an accurate and up-to-date measurement. Transitions from R to W occur at rate d, corresponding to the *drift rate*—the average frequency with which the true system state deviates beyond an acceptable tolerance, rendering the last received measurement obsolete. Conversely, legitimate updates from the transmitter arrive at rate $p > 0$, immediately restoring correctness by refreshing the stored value. The pair (d, p) thus fully characterizes the stochastic correctness dynamics under nominal, attack-free operation, as illustrated in Figure 12.2a.

Using renewal theory, the steady–state average AoII can be obtained by analyzing one full correctness cycle, composed of a correct period followed by an incorrect one. These periods are exponentially distributed with parameters d and p, respectively. By computing the expected area under the AoII curve in each cycle and dividing by the expected cycle duration, one obtains:

$$\Delta(p) = \frac{\frac{1}{2p^2}}{\frac{1}{p} + \frac{1}{d}}. \tag{6.16}$$

This expression captures the trade–off between the transmission rate p and the natural drift rate d. A larger p reduces the average duration for which the information is incorrect, improving freshness, but at the cost of higher communication and energy expenditure.

The transmitter thus faces an optimization problem: selecting its transmission rate to balance *information correctness* against *transmission cost*. This trade–off can be formalized as:

$$u_N(p) = -\Delta(p) - Cp,$$

where $C > 0$ denotes the cost per update. The optimal transmission policy in the nominal case is then obtained by maximizing this utility:

$$p^* = \arg\max_p u_N(p),$$

which identifies the transmission rate that minimizes the average AoII while maintaining efficient use of communication resources. This optimization expresses the fundamental trade-off between information freshness and transmission effort: increasing p reduces

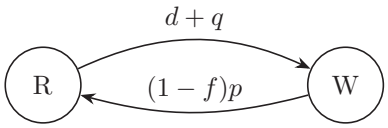


Figure 6.3: Adversarial correctness dynamics: corruption increases the degradation rate $(d+q)$, while blocking reduces the effective restoration rate $(1-f)p$.

the time during which the receiver holds incorrect data but also increases energy and bandwidth consumption.

By solving $\frac{\partial u_N(p)}{\partial p} = 0$, it is possible to show that the utility function admits a unique stationary point corresponding to the global maximum. The resulting optimal transmission rate is given in closed form by:

$$p^* = \frac{1}{6}\left(-3d + \sqrt{3}\,\sqrt{s} + \frac{3}{C}\sqrt{\frac{d^2}{3} + \frac{s}{3} + \frac{2\sqrt{3}d}{C\sqrt{s}}}\right), \tag{6.17}$$

where

$$s = d^2 + \frac{d^4 C}{r} + \frac{r}{C},$$

$$r = \left(\frac{27d^2C}{2} + d^6C^3 + \frac{3\sqrt{3}}{2}\sqrt{d^4C^2\left(27 + 4d^4C^2\right)}\right)^{\frac{1}{3}}.$$

This expression represents the unique feasible transmission rate that maximizes the transmitter's utility in nominal conditions. It captures the balance between the rate of physical evolution, characterized by the drift parameter d, and the communication cost C. For rapidly varying processes (large d), the optimal rate p^* increases, reflecting the need for more frequent updates to maintain correctness. Conversely, for higher transmission costs (large C), the transmitter reduces its update frequency to remain economically efficient.

This closed-form solution defines the nominal operating point of the system, serving as the analytical benchmark against which the impact of adversarial actions will be assessed in the following sections.

Adversarial case. When the communication channel is exposed to cyber threats, the correctness dynamics are altered by the attacker's behavior. Two primary adversarial mechanisms are considered. First, an attacker may inject falsified data that misleads the receiver, effectively accelerating the degradation of correctness. This is modeled as an additional corruption rate $q \geq 0$, which increases the total transition rate from R to W to $(d+q)$. Second, the attacker may block or jam legitimate transmissions, preventing updates from restoring correctness. This is represented by a reduction in the effective restoration rate, modeled as $(1-f)p$, where $f \in [0,1)$ denotes the fraction of updates that are lost or corrupted in transit. The resulting adversarial correctness model is shown in Figure 6.3.

Following the same renewal–theoretic reasoning, the average AoII in the presence of adversarial interference is given by:

$$\Delta(p,q,f) = \frac{\frac{1}{2(1-f)^2p^2}}{\frac{1}{(1-f)p} + \frac{1}{d+q}}. \tag{6.18}$$

This closed–form metric explicitly quantifies how both the attack intensity (q,f) and the transmission policy p affect the receiver's average time under incorrect information. As expected, increasing p helps mitigate the effect of attacks, but the benefit diminishes

as f grows, since a larger fraction of updates fail to reach the receiver. Conversely, higher corruption rates q make correctness loss more frequent, directly amplifying Δ.

The AoII thus provides a unified measure that captures both *timeliness* and *semantic integrity* of information, making it an appropriate performance index for evaluating adversarial effects in cyber–physical systems.

Strategic layer. To capture the interaction between the legitimate transmitter and the adversarial agent, we introduce a strategic layer in which both entities act as rational decision-makers with opposing objectives. The legitimate transmitter (T) aims to minimize the expected *Age of Incorrect Information* while controlling its communication expenditure, whereas the malicious agent (M) seeks to maximize AoII while minimizing its own operational cost. This strategic interplay forms the foundation of a two-player non-cooperative game.

We denote by r the *attacker's activity rate*, representing the effort required by the adversary to induce a given level of disruption in the communication process—such as a failure probability f, a false data injection rate q, or a combined effect in hybrid attacks. In general, the activity rate can be expressed as:

$$r = \gamma_{\mathsf{A}}(q, f, p), \tag{6.19}$$

where the function $\gamma_{\mathsf{A}}(\cdot)$ characterizes the specific attack mechanism and its dependence on both the attacker's and the transmitter's strategies.

Both agents are assumed to incur costs proportional to their level of activity. The transmitter pays a unit transmission cost $C > 0$ per unit rate p, while the adversary pays a unit operational cost $K > 0$ per unit of activity rate r. These costs reflect energy expenditure, bandwidth allocation, or computational load associated with each agent's actions.

Based on these definitions, we introduce the utility functions for the two players:

$$u_{\mathrm{T}}(p, q, f) = -\Delta(p, q, f) - C\,p, \qquad u_{\mathrm{M}}(p, q, f) = \Delta(p, q, f) - K\,r, \tag{6.20}$$

where $\Delta(p, q, f)$ denotes the expected steady-state AoII as a function of the transmitter's and attacker's strategies. The transmitter seeks to minimize its utility loss, while the attacker seeks to maximize its own.

The strategic equilibrium of this interaction is defined by the pair $(p^\star, r^\star)$ satisfying the Nash conditions:

$$\frac{\partial u_{\mathrm{T}}}{\partial p} = 0, \qquad \frac{\partial u_{\mathrm{M}}}{\partial r} = 0.$$

At this point, neither player can improve its outcome unilaterally, representing a stable balance between defensive and offensive strategies. This equilibrium captures the trade-off between system resilience and attack cost, and it provides insight into the parameter regimes where adversarial actions become either economically viable or self-defeating.

Chapter 7

False Data Injection

In this chapter, we focus on one of the most insidious and technically challenging forms of adversarial interference in networked estimation systems: *False Data Injection* (FDI). Unlike denial-based attacks, which disrupt communication by suppressing or delaying packet delivery, FDI attacks preserve the apparent normality of network operations while corrupting the informational content of transmitted data. From the perspective of the communication layer, the system continues to function nominally—packet rates, delays, and connectivity remain within expected ranges—yet the receiver's knowledge of the monitored process gradually diverges from reality. This discrepancy, hidden behind an illusion of freshness and stability, renders FDI particularly dangerous in cyber–physical and IoT infrastructures, where decisions rely on the integrity of streamed information.

Building on the stochastic framework introduced in the previous chapter, we specialize the continuous-time Markov model of information correctness to explicitly capture the mechanisms of false update injection. The alternation between correct and incorrect information states at the receiver is driven by two competing processes: legitimate updates that restore correctness, and adversarial injections that introduce falsified data. We derive closed-form expressions quantifying the impact of these processes on the *Age of Incorrect Information* , which serves as a fundamental metric for evaluating both the temporal and semantic reliability of the received data. This metric provides a principled means to quantify how FDI attacks degrade information trustworthiness over time and to analyze the trade-offs faced by both legitimate and malicious agents.

The chapter proceeds from analytical modeling toward strategic characterization. We first establish the stochastic representation of FDI-induced corruption and then interpret the resulting interactions between transmitter and adversary through the lens of non-cooperative game theory. This formulation enables a systematic study of equilibrium behaviors, resource allocation strategies, and stability properties in both complete and incomplete information settings. Finally, the analysis is extended to multi-sensor architectures, revealing how limited communication and attack resources distribute across interconnected nodes and how such coupling shapes system-level correctness and resilience.

Organization of the chapter. The structure of the chapter is as follows:

- **Section 7.1** introduces the continuous-time Markov model for False Data Injection attacks and derives closed-form expressions for the expected Age of Incorrect Information (AoII). This section establishes the quantitative foundation for assessing the informational impact of adversarial updates.
- **Section 7.2** formulates the interaction between transmitter and adversary as a static game with complete information. Analytical expressions for the Nash Equilibrium are obtained, characterizing how optimal update and injection rates depend on system parameters and operational costs.
- **Section 7.3** extends the model to incomplete information by introducing uncertainty in the attacker's cost structure. The resulting Bayesian game formulation yields the Bayesian Nash Equilibrium, capturing how belief and uncertainty shape defensive and offensive strategies.
- **Section 7.4** generalizes the analysis to multi-sensor systems, where multiple data sources and attack targets coexist under shared resource constraints. The section derives equilibrium conditions and explores the collective dynamics of freshness, correctness, and adversarial allocation in distributed architectures.

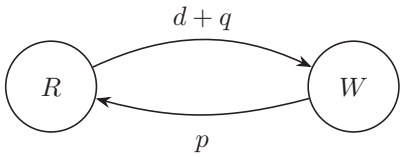


Figure 7.1: Continuous-time Markov chain under False Data Injection . The adversary increases the transition rate from the correct to the incorrect state by injecting falsified updates at rate q.

7.1 False Data Injection: System-Level Impact

We now specialize the generic framework introduced in Section 6.3.1 to the case of *False Data Injection* attacks. In practical terms, an FDI adversary may: (i) alter sensor readings before transmission by compromising a field device or communication node; (ii) modify packets in transit between transmitter and receiver; or (iii) impersonate legitimate sensors by transmitting fabricated updates.

Unlike denial-based attacks, FDI does not necessarily disrupt packet delivery. Communication continues to appear nominal at the network level—packet rates, delays, and connectivity remain unaffected—making these attacks particularly insidious. Instead, their impact manifests as a progressive deviation between the receiver's perceived and actual process states, thereby corrupting the information content while preserving apparent freshness.

When an FDI attack is active, the adversary injects falsified updates at rate q. Each injected update drives the system into an incorrect-information state, even though a new packet is successfully received. Hence, the overall rate of transitions from the correct state (R) to the incorrect state (W) increases from the natural drift rate d to $d+q$, as shown in Figure 7.1. Meanwhile, legitimate updates transmitted at rate p restore correctness, inducing the transition $W \to R$.

In the terminology of the general framework, this scenario corresponds to an attacker whose *activity rate function* is:

$$\gamma_{\mathsf{F}}(q, f, p) = q, \tag{7.1}$$

meaning that the adversary's effort is directly proportional to the injection rate q, since each falsified update requires one unit of activity.

The long-term behavior of this two-state continuous-time Markov chain determines the receiver's exposure to incorrect information, quantified through the *Age of Incorrect Information* (AoII). By integrating over the steady-state distribution of the chain, the expected AoII under FDI can be derived as:

$$\Delta_{\mathsf{F}}(p, q) = \frac{\frac{1}{2p^2}}{\frac{1}{p} + \frac{1}{d+q}} = \frac{d+q}{2p(p+d+q)}. \tag{7.2}$$

This closed-form expression highlights two opposing effects:

- Increasing the legitimate update rate p reduces Δ_{F}, thereby improving information freshness and correctness;
- Increasing the injection rate q worsens Δ_{F}, since the receiver transitions to the incorrect state more frequently.

As the attack intensity q grows, the AoII increases monotonically, indicating that the receiver spends more time holding erroneous information. This metric therefore provides a unified quantitative measure of the degradation induced by FDI attacks, capturing both the temporal and informational dimensions of performance loss.

The analytical formulation of $\Delta_{\mathsf{F}}(p, q)$ in (7.2) constitutes the foundation for the subsequent game-theoretic analysis. There, both the transmitter and the adversary are modeled as rational agents: the transmitter seeks to minimize AoII while limiting its transmission cost, whereas the attacker aims to maximize AoII while containing its attack expenditure. The resulting equilibrium analysis reveals how optimal strategies emerge from the balance between communication effort, attack intensity, and cost efficiency.

7.2 Static Game-Theoretic Modeling with Complete Information

To analyze the strategic interaction between the legitimate transmitter and the adversarial agent, we begin with the simplest yet fundamental setting: a *static game with complete information*. We focus on a single sensing unit, so that the process variable is scalar, $y(t) \in \mathbb{R}$. This assumption allows us to capture the essential dynamics of information freshness and correctness without the additional complexity introduced by multiple sensors or coupling effects.

In this framework, both players choose their actions simultaneously, and each has full knowledge of the system parameters (e.g., drift rate d, cost coefficients C and K), the admissible action set of the opponent, and the utility functions that describe both agents' objectives.

In other words, complete information means there is no uncertainty about the opponent's cost of action or preferences. This assumption enables a clear characterization of the strategic coupling between transmission and attack decisions before introducing information asymmetries in later chapters.

7.2.1 Game-Theoretic formulation

We formalize the interaction as a two-player static non-cooperative game

$$\mathcal{G}_{\mathsf{F}} = (\mathcal{P}, \mathcal{A}, \mathcal{U}),$$

where:

- $\mathcal{P} = \{\mathrm{N}, \mathrm{M}\}$ denotes the set of players: the legitimate transmitter (N) and the malicious agent (M);
- the continuous action spaces are $p \in [0, \infty)$ for N and $q \in [0, \infty)$ for M, representing transmission and injection rates, respectively;
- and $\mathcal{U} = \{u_{\mathrm{N}}, u_{\mathrm{M}}\}$ denotes the pair of utility functions defined below.

The players' utilities are built upon the *Age of Incorrect Information* , $\Delta_{\mathrm{F}}(p, q)$, which quantifies how long the receiver holds erroneous or outdated information. As derived in Section 7.1, the expected KoII under an FDI attack is:

$$\Delta_{\mathrm{F}}(p, q) = \frac{1/(2p^2)}{1/p + 1/(d+q)} = \frac{d+q}{2p(p+d+q)}. \tag{7.3}$$

The transmitter aims to maintain information freshness (i.e., minimize Δ_{F}) while reducing transmission effort. Conversely, the adversary seeks to maximize Δ_{F}, thereby worsening the receiver's situational awareness, but must also pay a cost proportional to its attack activity. These competing objectives are captured by the following utility functions:

$$u_{\mathrm{N}}(p, q) = -\Delta_{\mathrm{F}}(p, q) - C\,p, \qquad u_{\mathrm{M}}(p, q) = \Delta_{\mathrm{F}}(p, q) - K\,q, \tag{7.4}$$

where $C > 0$ represents the cost per unit transmission rate for the defender, which can model energy expenditure, bandwidth limitations, or scheduling penalties, and $K > 0$ represents the cost per unit injection rate for the attacker, capturing the energy, risk, or complexity associated with generating false data.

The first term in each utility corresponds to the informational effect (AoII), while the second term represents the resource expenditure or operational penalty. This structure reflects a *zero-sum tendency*—the attacker's gain in AoII corresponds to a loss for the defender—yet each agent incurs its own cost independently, leading to a non-zero-sum game.

7.2.2 Derivation of the Nash Equilibrium

A central goal in the game-theoretic analysis is the identification of the *Nash Equilibrium* — a state in which no player can unilaterally improve its own utility by changing its strategy, given the strategy of the opponent. In this context, the NE represents a stable operating point for the transmitter–attacker pair: once reached, neither agent has an incentive to deviate from its chosen rate.

Before deriving the explicit equilibrium strategies, we first establish that the game $\mathcal{G}$, as defined in Section 7.2, admits a *unique* Nash Equilibrium in pure strategies. This theoretical result guarantees that the strategic interaction between the transmitter and the adversary leads to a single, well-defined steady-state behavior, ensuring both existence and predictability of the equilibrium outcome.

Theorem 7.2.1 (*Existence and Uniqueness of Nash Equilibrium*)

The static game $\mathcal{G}_{\mathrm{F}}$ admits a unique pure-strategy Nash Equilibrium. □

Proof. We begin by observing that the utility functions $u_{\mathrm{N}}(p,q)$ and $u_{\mathrm{M}}(p,q)$ are continuously differentiable in p and q, respectively.

For fixed q, the transmitter's utility $u_{\mathrm{N}}(p,q)$ is strictly concave in p. This can be shown by computing the second derivative:

$$\frac{\partial^2 u_{\mathrm{N}}}{\partial p^2} = -\frac{1}{p^3(p+d+q)^2} - \frac{2(d+q)}{p^2(p+d+q)^3},$$

which is strictly negative for all $p > 0$ and $q \geq 0$.

Similarly, for fixed p, the adversary's utility $u_{\mathrm{M}}(p,q)$ is strictly concave in q, since:

$$\frac{\partial^2 u_{\mathrm{M}}}{\partial q^2} = -\frac{1}{p(p+d+q)^3} < 0.$$

Because both utilities are strictly concave in their respective strategy variables and the action spaces are convex (and can be assumed compact without loss of generality), Glicksberg's extension of Nash's Theorem guarantees the existence of at least one pure-strategy Nash Equilibrium. The strict concavity further ensures that this equilibrium is unique. □

For the specific formulation at hand, the NE conditions can be numerically derived by looking at the critical points of the analytical expressions, which implies to solve the following conditions [281]

$$\frac{\partial u_{\mathrm{M}}(p,q)}{\partial q} = 0 \qquad \frac{\partial u_{\mathrm{N}}(p,q)}{\partial p} = 0\,, \tag{7.5}$$

which imply

$$\frac{\partial \Delta_{\mathrm{F}}}{\partial q} = K \qquad \frac{\partial \Delta_{\mathrm{F}}}{\partial p} = -C. \tag{7.6}$$

Rearranging the terms in $\frac{\partial \Delta_{\mathrm{F}}}{\partial q} = K$ results in a second-degree equation with respect to q:

$$\frac{1}{2\,(d+p+q)^2} = K\,.$$

Solving this equation is straightforward and yields two solutions, one negative for all possible values of K. Therefore, the only viable solution to our problem is given by

$$q = -d - p + \frac{1}{\sqrt{2K}}.$$

Substituting this expression for q into $\frac{\partial \Delta_{\mathrm{F}}}{\partial p} = -C$ and rearranging the terms, we obtain a second-degree equation with respect to p:

$$K - \frac{1}{2p^2} = C.$$

Once again, it is straightforward that this equation has two solutions, one of which is negative for all possible values of C. We conclude that at the equilibrium

$$p = \frac{1}{\sqrt{2K + 2C}}.$$

Thus, the NE is

$$\begin{aligned} p &= \frac{1}{\sqrt{2K + 2C}}, \\ q &= -d - p + \frac{1}{\sqrt{2K}}. \end{aligned} \tag{7.7}$$

In (7.7), K must be sufficiently small to ensure that q is positive. If the mathematical solution of (7.7) yields a value of q less than zero, it implies that the attacker gains no benefit from the manipulated data and therefore chooses to remain silent. In this case the NE still exists, but the condition degenerates to the border points of the feasibility intervals, in particular q is equal to 0 and the optimal update frequency p is reduced to a single-agent optimization problem that maximizes $u_N(p, 0)$ as per (6.17).

Hence, in (7.7), the injection cost term K is balanced against the natural drift d and the transmission rate p. If these terms become too large, q can only remain positive if K is low enough. This highlights the trade-off between the different factors influencing the system dynamics.

Taking a broader perspective, we can explore the range of appropriate values for K, ensuring that player M actively participates in the game and (7.7) accurately reflects the NE of the system. From (7.7), if $K > d^{-2}/2$, it is impossible to get $q > 0$.

Even if $K < d^{-2}/2$, it might be inconvenient for the adversary to transmit, as it must also hold

$$-d - \frac{1}{\sqrt{2K + 2C}} + \frac{1}{\sqrt{2K}} > 0. \tag{7.8}$$

Considering that K and C are positive and so is d, the following theorem states that a threshold value $K_{\mathsf{F}}^{\star}$ exists that defines the border between the conditions for the NE to lie in an inner point or to degenerate at the border. In more detail, if $K > K_{\mathsf{F}}^{\star}$, the cost is too high for the adversary and it does not really partake in the interaction, leaving the only drift to the natural rate d.

Theorem 7.2.2 (*Existence and Uniqueness of the Critical Cost $K_{\mathsf{F}}^{\star}$*)

For all $C \in (0, \infty)$ and $d \in (0, \infty)$, a value $K_{\mathsf{F}}^{\star}$ exists such that for $K < K_{\mathsf{F}}^{\star}$,

$$\begin{cases} p^{\star} > 0,\ q^{\star} > 0, & \text{if } K < K_{\mathsf{F}}^{\star}, \\ p^{\star} = p_{\mathrm{N}},\ q^{\star} = 0, & \text{if } K \geq K_{\mathsf{F}}^{\star}. \end{cases}$$

□

The proof is provided in Appendix C.1.

This result implies that player M will avoid transmitting if $K > K_{\mathsf{F}}^{\star}$ since the cost is too high. According to the obtained equations at NE, a higher transmission cost for

the controller lowers the transmission rate p. Thus, the theorem implies that for the adversary to conveniently transmit, the controller's update rate must be low enough to undermine the system.

To sum up, the NE conditions are

$$p^\star = \begin{cases} \dfrac{1}{\sqrt{2K+2C}}, & \text{if } K < K_{\mathsf{F}}^\star, \\ p_{\mathrm{N}} & \text{otherwise,} \end{cases} \tag{7.9}$$

$$q^\star = \begin{cases} -d - p^\star + \dfrac{1}{\sqrt{2K}}, & \text{if } K < K_{\mathsf{F}}^\star, \\ 0, & \text{otherwise.} \end{cases} \tag{7.10}$$

The first pair of expressions corresponds to the Nash Equilibrium of the static game $\mathcal{G}$, in which both players—N and M—actively participate. The second case, on the other hand, arises when the attacker's cost exceeds the threshold $K_{\mathsf{F}}^\star$, making the adversary inactive ($q^\star = 0$). In this regime, the transmitter optimizes its utility function becomes $u_{\mathrm{N}}(p, 0)$, and the optimization

$$arg\max_{p} \{-\Delta_{\mathrm{F}}(p, 0) - Cp\}$$

admits a single feasible solution, which is p_{N} defined in (6.17).

7.2.3 Numerical Results

This section illustrates the strategic interaction between the transmitter and the adversary through numerical simulations, highlighting the equilibrium behavior and sensitivity of the system to key parameters. All numerical results are obtained using the analytical expressions derived in the previous sections.

Figure 7.2 shows the adversary's utility $u_{\mathrm{M}}(p, q)$ as a function of its own injection rate q for several fixed transmission rates p. As the transmitter increases its update frequency, the adversary's payoff decreases, reflecting the deterrent effect of frequent updates: timely transmissions limit the window during which falsified data can persist in the receiver's memory. This inverse relationship captures the core strategic tension of the problem—while higher update rates reduce vulnerability, they also entail greater communication cost for the transmitter.

Figure 7.3 presents the transmitter's utility $u_{\mathrm{N}}(p, q)$ as a function of the adversarial injection rate q. As q increases, the transmitter's performance deteriorates due to the rising frequency of corrupted updates. This behavior confirms that adversarial activity directly penalizes estimation accuracy and freshness, reinforcing the need for optimal response strategies that balance communication cost and resilience.

Figures 7.4 and 7.5 compare the equilibrium transmission rate $p^\star$ in the presence and absence of an adversary. For small values of the attack cost parameter C, the adversary remains inactive, and the equilibrium coincides with the non-adversarial case. As C increases, the attacker becomes more active, prompting the transmitter to compensate by increasing its update rate to preserve information correctness. This transition highlights the sensitivity of equilibrium strategies to cost structures and the existence of a critical threshold $K_{\mathsf{F}}^\star$, beyond which the adversary's actions are no longer profitable.

Finally, Figure 7.6 reports the equilibrium values of $p^\star$ and $q^\star$ as functions of the natural drift rate d. As the underlying process evolves more rapidly, the transmitter increases its update effort to maintain correctness, whereas the adversary reduces its injection activity: when the system's natural dynamics already induce significant information decay, further attack effort yields diminishing returns. This inverse adjustment demonstrates how the

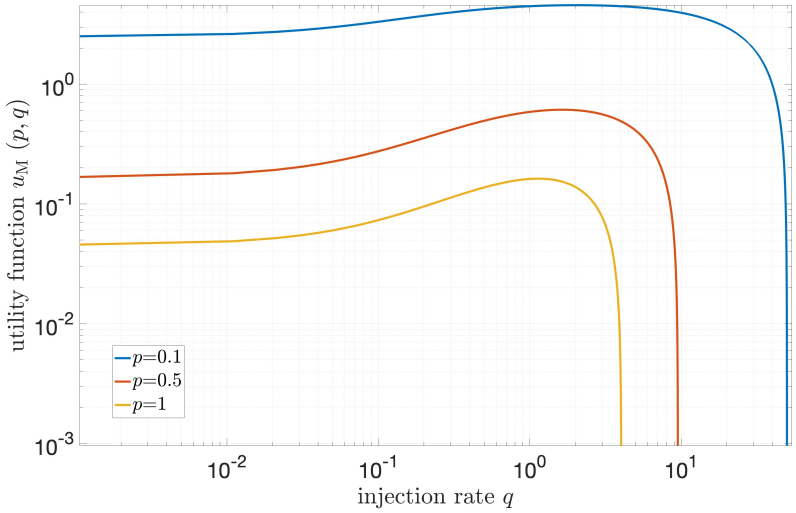


Figure 7.2: Adversary's utility $u_{\mathrm{M}}(p,q)$ for various transmission rates p ($C = 1$, $K = 0.1$, $d = 0.1$).

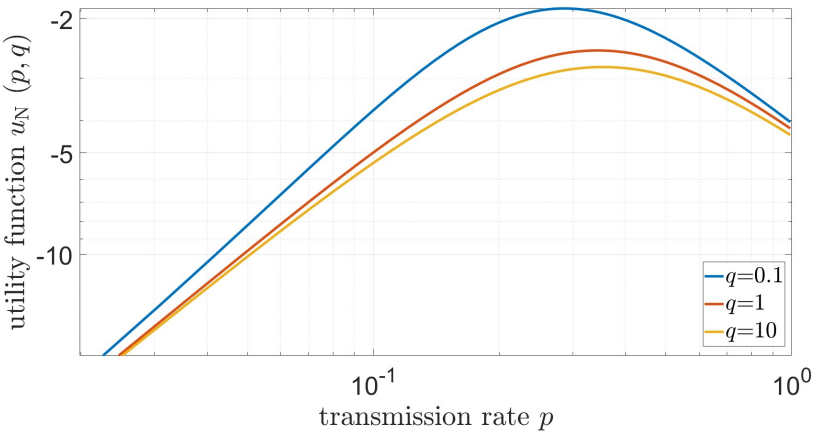


Figure 7.3: Transmitter's utility $u_{\mathrm{N}}(p,q)$ for different values of q ($C = 1$, $K = 0.1$, $d = 1$).

Nash Equilibrium internalizes the joint effects of environmental dynamics and adversarial interference.

The numerical analysis confirms the theoretical findings of the static game formulation. The existence of a unique Nash Equilibrium provides a robust analytical foundation for predicting rational behaviors in contested communication environments. The results emphasize how optimal strategies are shaped by both cost asymmetries and system dynamics, and how appropriate tuning of update policies can effectively neutralize rational adversaries. These insights offer practical guidance for the design of resilient communication protocols that sustain correctness and freshness even under strategic interference.

7.3 Bayesian Game-Theoretic Modeling with Incomplete Information

In realistic adversarial scenarios, it is reasonable to assume that the attacker has both the time and the incentive to thoroughly analyze the target system, thereby acquiring accurate

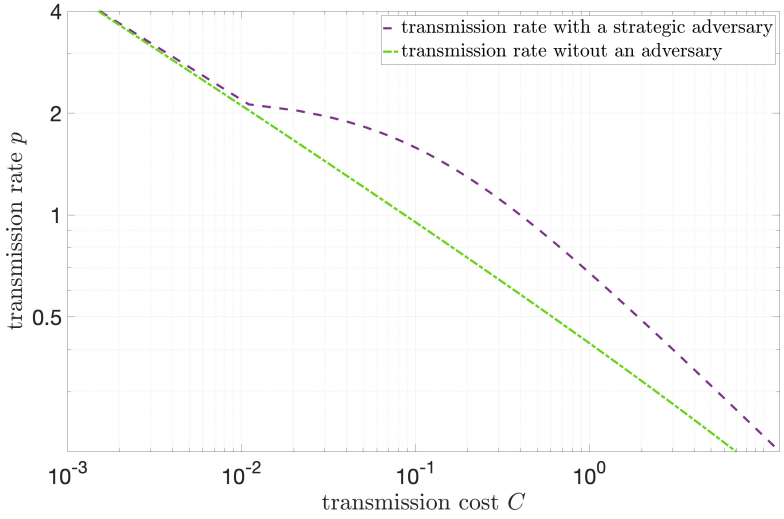


Figure 7.4: Equilibrium transmission rate $p^\star$ with and without adversary ($K = 0.1$, $d = 0.5$).

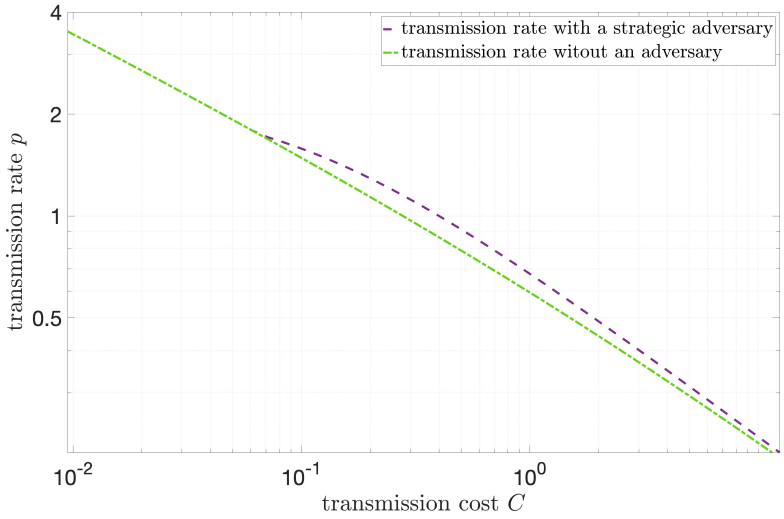


Figure 7.5: Equilibrium transmission rate $p^\star$ with and without adversary ($K = 0.1$, $d = 1$).

knowledge of its parameters and structural properties. In this section, we focus on the single–sensor case, where the process output is scalar, $y(t) \in \mathbb{R}$. This simplification allows us to isolate the effect of uncertainty in the adversary's behavior on the transmitter's optimal policy, before generalizing to multi–sensor settings in later chapters.

Conversely, the legitimate transmitter—often constrained by time, observability, or computational resources—may not possess precise information about the attacker's internal cost structure or operational capabilities. To capture this inherent asymmetry of knowledge, we extend the previous formulation to a *Bayesian* setting in which the attacker's cost parameter K is modeled as a random variable with known probability distribution. This probabilistic representation enables the transmitter to plan its strategy based on beliefs about the attacker's characteristics rather than exact knowledge, naturally leading to the concept of a *Bayesian Nash Equilibrium*.

In this framework, the injection cost K is privately known only to the attacker, while the

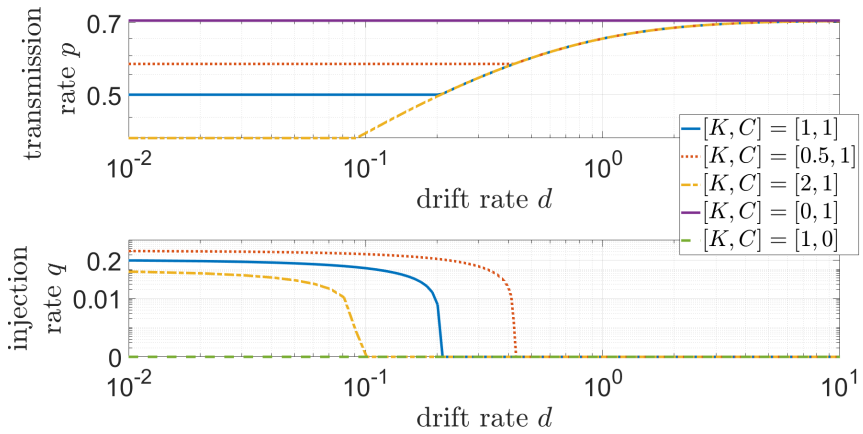


Figure 7.6: Equilibrium values of $p^\star$ and $q^\star$ as functions of drift rate d.

transmitter has access solely to its probability density function (pdf) $f_K(k)$. Uncertainty is thus associated exclusively with the attacker's cost, whereas the defender's cost parameter C remains fixed and known. This modeling choice mirrors typical adversarial settings: the attacker can study the system in advance and act under heterogeneous or uncertain resource costs, while the defender reacts using fixed, well-defined communication or energy expenditures. Such an asymmetry aligns with standard formulations in security games (e.g., [282]), where uncertainty is typically linked to the adversary's private information.

7.3.1 Bayesian Game- Theoretic Formulation

The strategic interaction under incomplete information is modeled as a Bayesian game

$$\mathcal{G}_{\mathrm{F}}^{B} = (\mathcal{P}, \mathcal{A}, \mathcal{U}, f_K(k)),$$

where:

- The set of players is $\mathcal{P} = \{\mathrm{N}, \mathrm{M}\}$, representing the legitimate transmitter (defender) and the malicious attacker;
- The action sets are continuous: the transmitter selects a transmission rate $p \in [0, \infty)$, while the attacker chooses a false data injection rate $q \in [0, \infty)$;
- The utility functions $\mathcal{U} = \{u_{\mathrm{N}}, u_{\mathrm{M}}\}$ describe the opposing objectives of the two agents:

$$u_{\mathrm{N}}(p, q) = -\Delta_{\mathrm{F}}(p, q) - Cp, \qquad u_{\mathrm{M}}(p, q; K) = \Delta_{\mathrm{F}}(p, q) - Kq, \tag{7.11}$$

  where $\Delta_{\mathrm{F}}(p, q)$ denotes the average Age of Incorrect Information (AoII), while $C > 0$ and $K > 0$ represent, respectively, the transmitter's communication cost and the attacker's resource cost.

Since the attacker's cost parameter K is privately known only to itself, the transmitter evaluates its performance in expectation with respect to its belief about K, represented by the probability density function $f_K(k)$:

$$\mathbb{E}_K[u_{\mathrm{N}}(p, q(K))] = \int_0^\infty [-\Delta_{\mathrm{F}}(p, q(K)) - Cp]\, f_K(k)\, dk,$$

where $q(K)$ denotes the attacker's best response to the chosen transmission rate p for a given cost realization $K = k$.

The main objective of this analysis is to determine the *Bayesian Nash Equilibrium* (BNE), which characterizes the stable operating point of the game under asymmetric information. A strategy pair $(p^\star, q^\star(K))$ constitutes a Bayesian Nash Equilibrium if neither player can improve its expected utility by unilaterally deviating from it.

7.3.2 Derivation of the Bayesian Nash Equilibrium

To determine the BNE, where each player optimally responds based on their beliefs, we can proceed as follows. Since the strategies are the choices of p and q for players N and M, respectively, we shall derive their best responses $p^\star(q)$ and $q^\star(p)$. A Nash equilibrium is a strategic choice where no player can improve their utility by unilaterally changing their strategy, given that the strategy of the other player remains unchanged [283]. This corresponds to a fixed point of the best responses, where neither player desires to deviate. However, player M has complete information on the value of K; thus, its computation of $q^\star(p)$, which is a function of K and therefore will be denoted as $q^\star_K(p)$ is immediate. The same does not apply to player N, who does not directly know K, only its distribution. Thus, the best response $p^\star(q^\star_K)$ in the BNE is the one maximizing the *expected* utility $\mathbb{E}_K[u_\mathrm{N}(p, q^\star_K)] = \int_0^\infty u_\mathrm{N}(p, q^\star_K(p)) f_K(k) dK$.

Theorem 7.3.1 (*Existence and Uniqueness of BNE*)

For game $\mathcal{G} = (\mathcal{P}, \mathcal{A}, \mathcal{U}, f_K(k))$, there exists a unique BNE. □

Proof. The utilities in (7.5) are built from $\Delta_\mathrm{F}(p, q) = (d + q)/(2\,p\,(d + q + p))$, a smooth rational function. Hence u_M is strictly concave in q and u_N is strictly concave in p. The same holds for $\mathbb{E}_K[u_\mathrm{N}(p, q^\star_K)]$ as a function of p, if $q^\star_K$ is fixed. Hence, since the (expected) utilities have a unique maximizer, the best response can be seen as a function, i.e., it is always unique, which implies that Nash theorem holds in a stronger version (Brouwer's theorem instead of Kakutani's) to prove the existence of the Nash equilibrium [283]. Also, the game is adversarial and constant-sum, since $u_N + u_\mathrm{M} = -Cp - Kq$ as Δ_F cancels out. This implies diagonal strict concavity, which according to Rosen's theorem, gives uniqueness of the Nash equilibrium [284]. □

Theorem 7.3.1 establishes existence and uniqueness of the Bayesian Nash equilibrium. By invoking Rosen's theorem on diagonally strictly concave games, the equilibrium is guaranteed to be structurally stable: small perturbations in the players' strategies do not generate additional equilibria, and the system converges back to the unique fixed point. This ensures that the equilibrium behavior is well-defined and robust from a game-theoretic perspective. For the formulation at hand, the BNE conditions can be derived by looking at the critical points of the analytical expressions, which implies the following conditions [285]

$$\frac{\partial u_\mathrm{M}(p^\star, q)}{\partial q} = \frac{\partial \Delta_\mathrm{F}(p^\star, q)}{\partial q} - K = 0$$
$$\frac{\partial \mathbb{E}_K[u_\mathrm{N}(p, q^\star_K)]}{\partial p} = -\mathbb{E}_K\left[\frac{\partial \Delta_\mathrm{F}(p, q^\star_K)}{\partial p} + K\right] = 0$$
$$\text{which implies: } \int_0^\infty \left[\frac{\partial \Delta_\mathrm{F}(p, q^\star_K)}{\partial p} + C\right] f_K(k)\, dK = 0. \tag{7.12}$$

Rearranging $\frac{\partial \Delta_\mathrm{F}(p^\star, q)}{\partial q} = K$ results in

$$(d + p^\star + q)^{-2} = 2K\,. \tag{7.13}$$

Solving (7.13) yields two formal solutions for q. However, one of them is negative for all admissible values of K and thus not feasible in the model. Therefore, the only relevant solution is

$$q^\star_K = -d - p^\star + \frac{1}{\sqrt{2K}}.$$

For the equilibrium to be economically meaningful, we require $q^\star_K \geq 0$. This condition fails to hold for sufficiently large K, implying the existence of a critical threshold K° such

that the optimal deviation vanishes beyond it. This threshold can be computed directly from (7.13) by setting $q = 0$.

$$K^\circ := \frac{1}{2(d+p^\star)^2}. \tag{7.14}$$

Thus, the equilibrium behavior can be fully characterized as follows:

$$q_K^\star = \begin{cases} -d - p^\star + (2K)^{-1/2}, & \text{if } K < K^\circ, \\ 0, & \text{if } K \geq K^\circ. \end{cases} \tag{7.15}$$

This result formalizes the intuitive economic insight that beyond a certain level of K — representing, for instance, cost intensity or regulatory pressure — the incentive to deviate vanishes. In compact form, we can write the necessary and sufficient condition as

$$K \geq K^\circ \quad \Rightarrow \quad q_K^\star = 0.$$

Substituting $q_K^\star$ into (7.12), we get $p^\star$ as the solution in p of

$$\begin{aligned} &\int_0^{K^\circ} \left(K - (2p^2)^{-1} + C\right) f_K(k)\ dK \\ &+ \int_{K^\circ}^{\infty} \left(-\frac{d(d+2p)}{2p^2(d+p)^2} + C\right) f_K(k)\ dK \ = \ 0. \end{aligned} \tag{7.16}$$

We can split the integral, obtaining:

$$\begin{aligned} &\left(C - \frac{1}{2p^2}\right) \underbrace{\int_0^{K^\circ} f_K(k)\ dK}_{\alpha} + \underbrace{\int_0^{K^\circ} K f_K(k)\ dK}_{K_0} \\ &+ \left(-\frac{d(d+2p)}{2p^2(d+p)^2} + C\right) \underbrace{\int_{K^\circ}^{\infty} f_K(k)\ dK}_{1-\alpha} \ = \ 0. \end{aligned} \tag{7.17}$$

where α is the probability that the attacker is active (i.e., $q{>}0$) and K_0 is its expected injection cost when active. Rearranging, we obtain $p^\star$ as the zero of polynomial:

$$v_4 p^4 + v_3 p^3 + v_2 p^2 + v_1 p + v_0 \tag{7.18}$$

whose coefficients are $v_4 = 2(C + K_0)$, $v_3 = 4d(C + K_0)$, $v_2 = -\alpha + 2d^2(C + K_0)$, $v_1 = -2d$ and $v_0 = -d^2$.

With some tedious mathematical manipulations, one can show that there is only one feasible solution, which is:

$$\begin{aligned} p^\star &= -\frac{v_3}{4v_4} - Q + \frac{1}{2}\sqrt{-4Q^2 - 2r + S/Q} \quad \text{, where:} \\ S &= \frac{8v_4^2 v_1 - 4v_4 v_3 c + v_3^3}{8v_4^3}, \quad r = \frac{8v_4 v_2 - 3v_3^2}{8v_4^2} \\ Q &= \frac{1}{2}\sqrt{-\frac{2}{3}r + \frac{1}{3v_4}\left(D + \frac{y}{D}\right)} \\ D &= \sqrt[3]{\frac{x + \sqrt{x^2 - 4y^3}}{2}}, \qquad y = 12v_4 v_0 - 3v_3 v_1 + v_2^2 \\ x &= 27v_4 v_1^2 - 72v_4 v_2 v_0 + 27v_3^2 v_0 - 9v_3 v_2 v_1 + 2v_2^3 . \end{aligned} \tag{7.19}$$

Combining (7.15) and (7.19), the solution is:

$$\begin{cases} p^\star = -\dfrac{v_3}{4v_4} - Q + \dfrac{1}{2}\sqrt{-4Q^2 - 2r + S/Q} \\ q_K^\star = \max\left\{0, -p^\star - d + \dfrac{1}{\sqrt{2K}}\right\}. \end{cases} \tag{7.20}$$

Table 7.1: PDF of considered distributions

	PDF $f_K(k)$	Variance
triangular	$\frac{\lvert k-K_0+\epsilon\rvert}{\epsilon}\,\mathrm{rect}\left(\frac{k-K_0}{2\epsilon}\right)$	$\epsilon^2/6$
uniform	$\frac{1}{2\epsilon}\mathrm{rect}\left(\frac{k-K_0}{2\epsilon}\right)$	$\epsilon^2/3$
Bernoulli	$0.5\delta(k-K_0-\epsilon)+0.5\delta(k-K_0+\epsilon)$	ϵ^2

This maximizes the utilities, as, beyond meeting first-order conditions, second-order conditions show that the utilities are concave in their decision variables. Thus, (7.20) is a BNE.

7.3.3 Numerical Results and Discussion

We consider a single-sensor scenario representative of slow-evolving yet safety-critical monitoring tasks, such as the continuous tracking of core body temperature in healthcare applications. In these settings, measurements change gradually, but occasional adversarial tampering can cause severe consequences [286]. The relatively high transmission cost models the constraints of wearable and battery-powered devices, which operate under tight energy budgets and limited computational capacity. This parameterization captures the intended trade-offs: a drift rate of d=0.001 implies that information ages slowly (time scale of tens of minutes), while a transmission cost C=4 ensures that update rates remain finite. Treating the attacker's cost K as a random variable with distribution $f_K(k)$ represents uncertainty in the intensity or frequency of tampering attempts, where typical magnitudes are $K \in [2, 10]$ and $d = \mathcal{O}(10^{-2})$–$\mathcal{O}(10^{-3})$.

Symmetric Attack Distributions. We first analyze the case in which the attacker's cost distribution $f_K(k)$ is symmetric on $[K_0-\epsilon, K_0+\epsilon]$, with ϵ denoting half the support width. Three representative families—triangular, uniform, and Bernoulli—are considered to explore different levels of dispersion while maintaining the same mean. Figure 7.7 shows the optimal transmission rate at the Bayesian Nash Equilibrium (BNE) for the three distributions, compared against two benchmarks: no attack ($\alpha = 0$) as in [287], and a fully active adversary ($\alpha = 1$) as in [285]. When the median cost K_0 is small, the attacker remains active, and the equilibrium coincides with the fully adversarial case. As K_0 increases, the transmitter perceives lower attack probability and reduces its transmission rate accordingly, until the attack becomes non-credible ($\alpha = 0$), recovering the non-adversarial regime.

To examine the representativeness of the expected injection cost, we generate 100 realizations of K for each $f_K(k)$ and compute the corresponding equilibrium injection rates $q_K^\star$ (Fig. 7.8). The comparison between $\mathbb{E}_K[q_K^\star]$ and $q_{K_0}^\star$ reveals that the attacker's behavior is not perfectly ergodic, yet deviations from the mean are moderate, especially for low-variance distributions. Thus, $q_{K_0}^\star$ provides a reliable lower bound for the actual injection rate and offers a practical approximation for the transmitter's belief.

Uncertainty in K directly affects the AoII. Figure 7.9 plots the ratio between $\Delta_{\mathsf{F}}(p^\star, \mathbb{E}_K[q_K^\star])$ and $\mathbb{E}_K[\Delta_{\mathsf{F}}(p^\star, q_K^\star)]$ as a function of K_0. Ratios below one indicate that variability in K amplifies the expected AoII. For highly dispersed Bernoulli distributions, the increase reaches approximately 12%, whereas for uniform and triangular cases the effect is negligible. Hence, the Bayesian game closely approximates the complete-information model of [285], except under extreme uncertainty.

The transmitter's expected utility, plotted in Fig. 7.10, confirms that optimal performance is achieved when its belief matches the mean attacker type K_0. Assuming overly pessimistic or optimistic beliefs ($K_0 \pm \epsilon$) leads to a measurable loss in expected utility, validating the Bayesian consistency of the equilibrium.

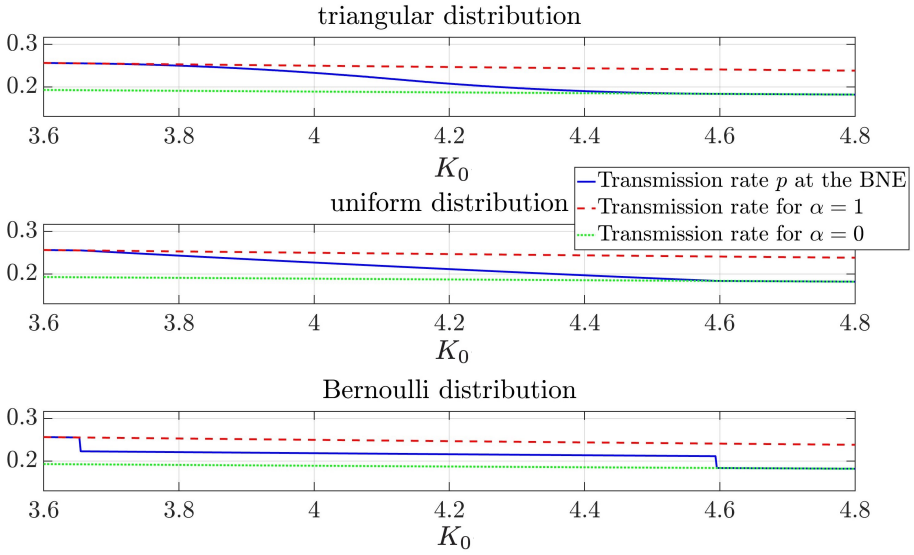


Figure 7.7: Comparison between transmission rates at BNE and benchmarks $\alpha = 1$ and $\alpha = 0$. $f_K(k)$ is symmetric with $\epsilon = 3$ and varying median values. Drift $d = 0.1$, transmission cost $C = 4$.

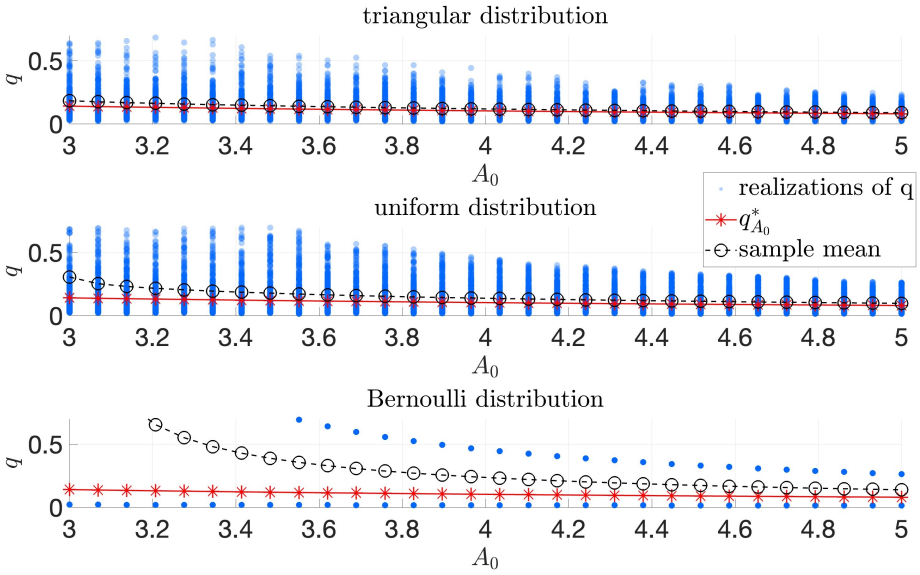


Figure 7.8: Injection rates for triangular, uniform, and Bernoulli $f_K(k)$. $C = 4$, $d = 0.001$. Solid red: prediction from K_0; black dots: realizations; dashed black: sample mean.

Figures 7.11 and 7.12 illustrate parameter sensitivity. Increasing the drift d reduces the attacker's optimal rate $q^\star$, as rapid natural aging diminishes the benefit of injecting false data. Conversely, higher transmission costs C depress $p^\star$ and indirectly favor the adversary by lowering the update frequency. These results emphasize how physical and economic factors jointly determine equilibrium behavior.

Asymmetric Attack Distributions. We now consider asymmetric priors, namely: (i) a triangular distribution on $[1, 8]$ with varying median, and (ii) a Bernoulli distribution with unequal probabilities over $\{3, 5\}$. Both introduce skewness in $f_K(k)$, modifying the strategic balance between transmitter and attacker. Figure 7.13 shows that the equilibrium transmission rate p decreases almost linearly with the expected attack level $\mathbb{E}_K[q^\star_K]$, regardless of the underlying distribution, while the empirical injection rates

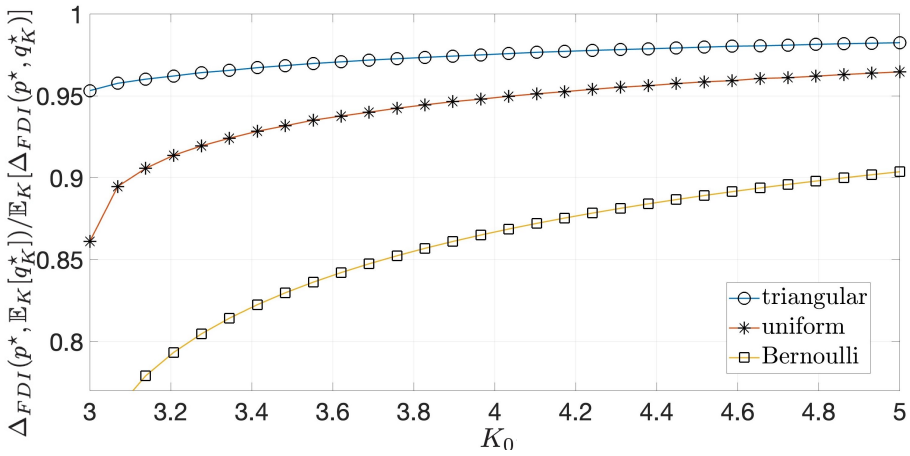


Figure 7.9: Ratio $\frac{\Delta_{\mathsf{F}}(p^\star, \mathbb{E}_K[q_K^\star])}{\mathbb{E}_K[\Delta_{\mathsf{F}}(p^\star, q_K^\star)]}$ under different $f_K(k)$. $C = 4$, $d = 0.001$.

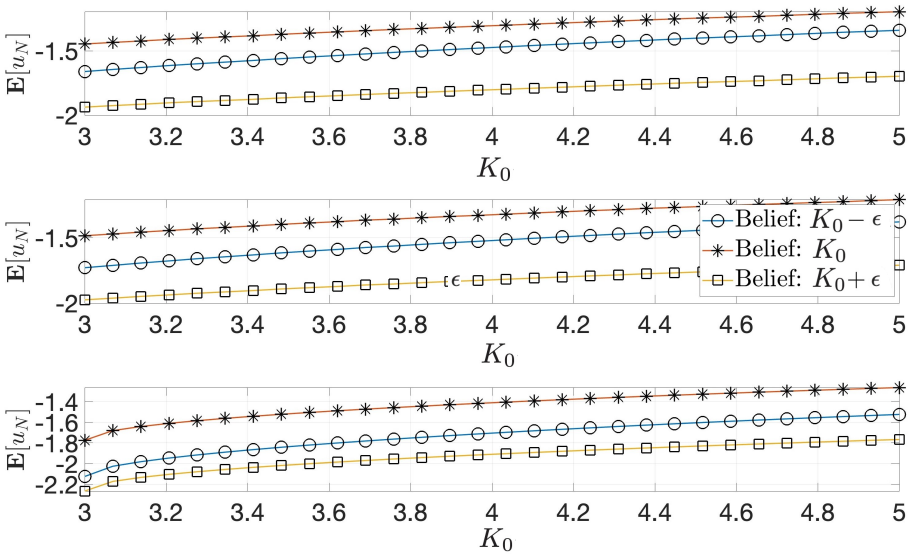


Figure 7.10: Expected utility of the transmitter under beliefs $K_0 - \epsilon$, $K_0 + \epsilon$, and K_0. $C = 4$, $d = 0.001$.

match closely with theoretical predictions. As asymmetry grows, the attacker allocates more effort toward the most probable type, leading to marginally higher AoII values. The nonlinear effect of uncertainty is illustrated in Fig. 7.14, showing the ratio between $\Delta_{\mathsf{F}}(p^\star, \mathbb{E}_K[q_K^\star])$ and $\mathbb{E}_K[\Delta_{\mathsf{F}}(p^\star, q_K^\star)]$. Both asymmetric distributions yield ratios slightly above one, indicating mild convexity: uncertainty marginally increases the expected AoII, peaking at intermediate injection costs and vanishing at the extremes. Overall, despite the small degradation, the Bayesian equilibrium remains stable and robust to distributional skewness.

Discussion. The combined analytical and numerical results reveal several consistent insights about FDI dynamics under uncertainty. First, the expected attack cost K_0 provides a reliable predictor of adversarial behavior when $f_K(k)$ has limited variance, allowing the transmitter to adopt near-optimal strategies with minimal information. Second, uncertainty in K systematically benefits the attacker, as variability increases the expected AoII, though the effect remains moderate for realistic parameter ranges. Third, equilibrium strategies depend strongly on the interplay between drift and cost: slower processes or higher transmission costs amplify vulnerability, whereas rapid natural aging

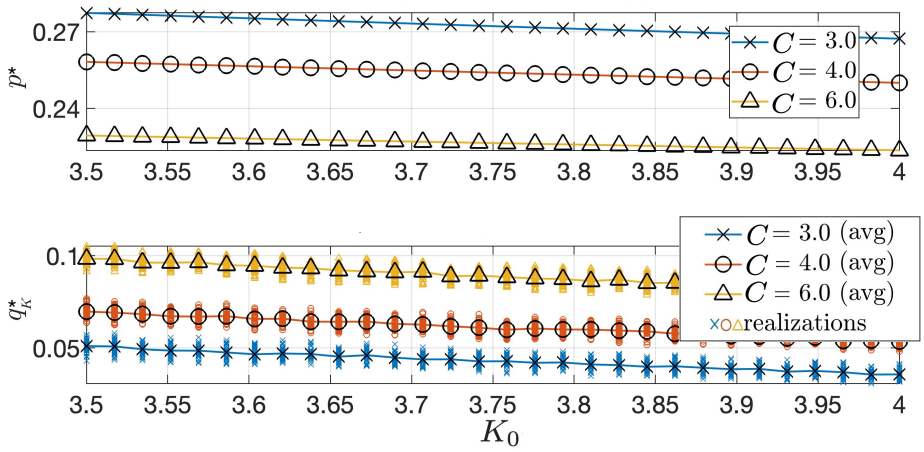


Figure 7.11: Bayesian Nash Equilibrium for $f_K(k)$ following a triangular distribution with $\epsilon = 0.3$ and drift $d = 0.05$. Solid lines indicate the sample mean, while markers represent individual realizations for varying median values.

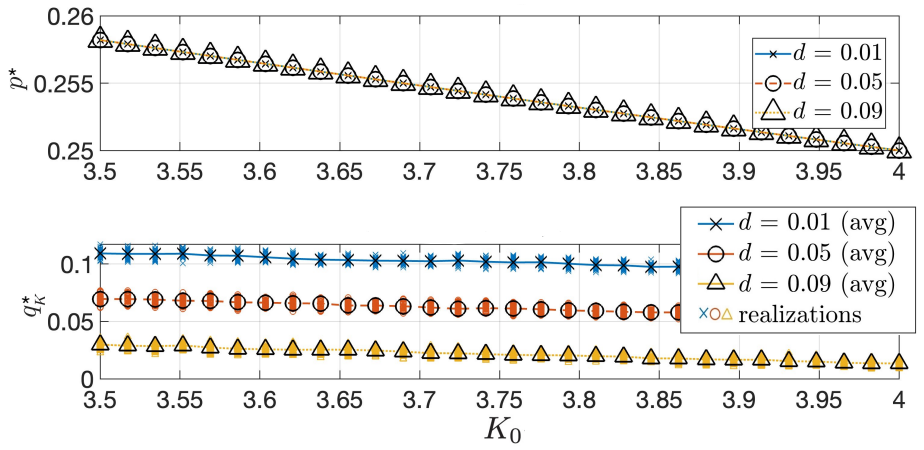


Figure 7.12: Bayesian Nash Equilibrium for $f_K(k)$ following a triangular distribution with $\epsilon = 0.3$ and transmission cost $C = 4$. Solid lines indicate the sample mean, while markers represent individual realizations for varying median values.

or inexpensive updates improve resilience. Finally, asymmetric priors and skewed attack distributions cause only marginal deviations in equilibrium, confirming the robustness of the Bayesian Nash framework.

7.4 AoII in Multi-Sensor Systems

In the initial part of the analysis, we considered a *single-sensor* setup in order to isolate and understand the fundamental effects of adversarial interference on information correctness and freshness. However, in realistic scenarios, monitoring and control systems typically rely on multiple sensors that observe different physical quantities of the same process. Each sensor thus provides a distinct, yet complementary, stream of information to the supervisory unit.

When multiple sensing units operate concurrently, their communication channels often share limited network or processing resources. As a consequence, the aggregate transmission activity must respect a global constraint determined by the receiver's processing capacity or by the available network bandwidth. This implies that the

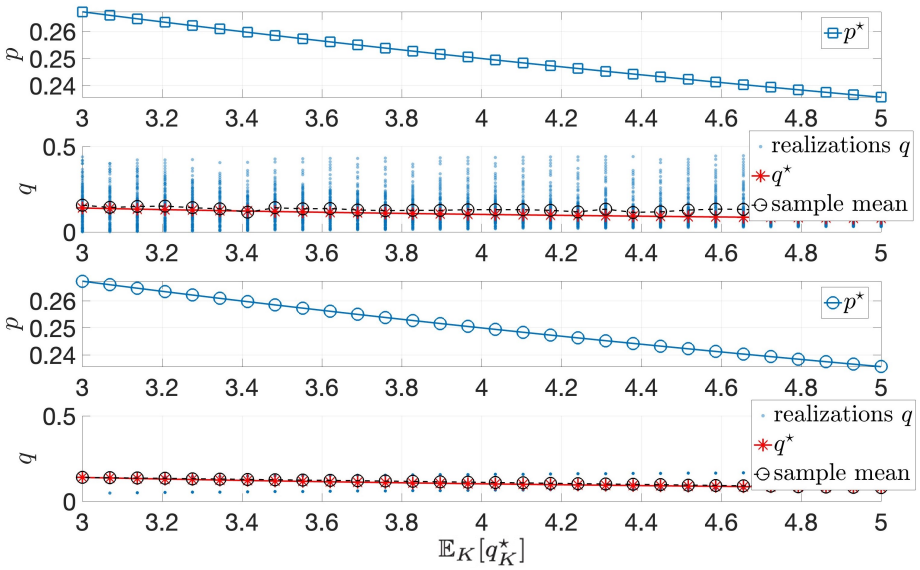


Figure 7.13: Equilibrium transmission rate p and injection rate q as functions of the expected attack level $\mathbb{E}_K[q_K^\star]$, evaluated under asymmetric prior distributions.

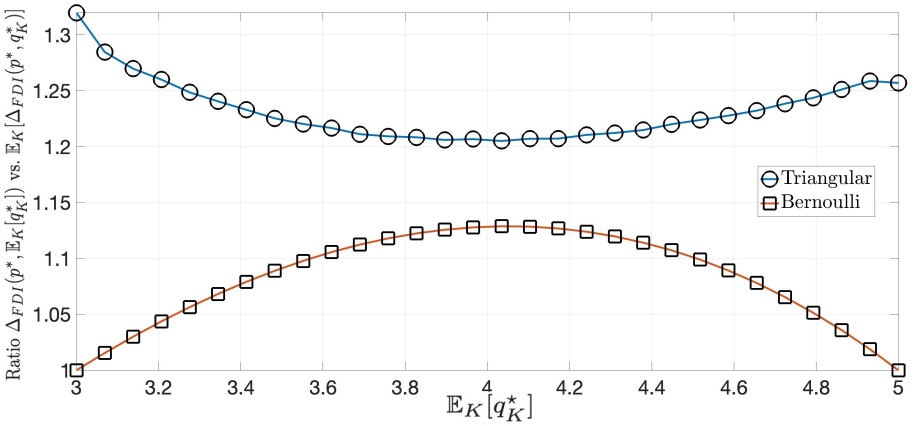


Figure 7.14: Ratio $\Delta_{\mathsf{F}}(p^\star, \mathbb{E}_K[q_K^\star])/\mathbb{E}_K[\Delta_{\mathsf{F}}(p^\star, q_K^\star)]$ computed for triangular and Bernoulli prior distributions.

overall transmission rate of the system is bounded, and individual sensor rates must be allocated accordingly, leading to coupling among sensors even when their measurements are independent.

To capture this more realistic scenario, we extend the previous single-sensor formulation to a multi-sensor framework in which each sensor maintains its own correctness dynamics but competes for shared communication and attack resources. Each sensing unit $i \in \{1, \ldots, m\}$ is characterized by a drift rate d_i, a transmission rate p_i, and an attack rate q_i, which govern its local evolution of correctness according to the same two–state continuous-time Markov model introduced earlier.

In this section, we focus on a *complete information* setting, where both the transmitter and the adversary are assumed to have full knowledge of the system parameters—namely, the drift rates d_i, the cost coefficients C_i and K_i, and the global resource constraints. This assumption allows us to clearly characterize the strategic coupling among sensors and to derive the equilibrium allocation of resources under perfect knowledge of the environment. In the following, we formalize this multi-sensor extension, define the aggregate *Age of Incorrect Information* , and analyze the resulting strategic interaction under shared resource constraints.

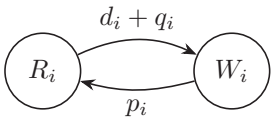


Figure 7.15: Continuous-time Markov process representing the correctness dynamics of sensor i, where state transitions capture the stochastic evolution between correct and incorrect information conditions.

7.4.1 Modeling Approach and Motivation

In the previous sections, we analyzed the interaction between a single transmitter and a single attacker, focusing on one sensor whose measured process was described by a scalar signal $y(t) \in \mathbb{R}$. This simplified setting allowed us to isolate the fundamental effects of adversarial interference on information correctness and freshness. However, realistic monitoring and control systems are composed of multiple heterogeneous sensors, each observing a different physical quantity of the same process. The *Age of Incorrect Information* naturally extends to this case, as each sensor provides its own data stream whose correctness may degrade over time or be affected by targeted attacks.

The discrepancy-based definition of AoII introduced in Definition 6.3.1 applies individually to each sensor. Accordingly, for sensor $i \in \{1, \dots, m\}$, we define:

$$\delta_i(t) = \ell_i(t) \cdot g_i\big(y_i(t), y_i(t_i^u)\big), \tag{7.21}$$

where $\ell_i(t)$ denotes the elapsed time since the last correct update of sensor i, and $g_i(\cdot)$ is an indicator function that equals 1 whenever the information stored at the receiver is incorrect (either outdated or corrupted) and 0 otherwise. Thus, $\delta_i(t)$ represents the instantaneous AoII associated with sensor i.

Each sensing unit is modeled as an independent two–state continuous-time Markov chain describing the local correctness dynamics. The sensor transitions from a correct state R_i (accurate) to an incorrect state W_i (outdated or falsified) at a total rate $d_i + q_i$, where d_i is the natural drift rate and q_i is the false data injection rate exerted by the attacker. Conversely, it transitions back to the correct state R_i at rate p_i, determined by the legitimate transmitter's update policy. This local process is illustrated in Figure 7.15.

The overall system state is given by the vector

$$S(t) = (S_1(t), \dots, S_m(t))$$

, where each component $S_i(t) \in \{R_i, W_i\}$. Assuming asynchronous state transitions—i.e., only one sensor changes state at any given time—the system evolves as a CTMC with 2^m possible states. The ideal configuration

$$R \triangleq (R_1, R_2, \dots, R_m),$$

corresponds to the situation where all sensors information available to the receiver is simultaneously correct.

The instantaneous AoII of the entire system, denoted $\delta(t)$, is defined as the arithmetic mean of the individual AoIIs:

$$\delta^{\text{tot}}(t) = \frac{1}{m} \sum_{i=1}^{m} \delta_i(t). \tag{7.22}$$

This aggregate measure captures the global freshness and correctness of information available to the receiver, providing a unified performance metric for multi-sensor systems under attack.

Both the transmitter and the adversary operate under shared resource constraints. The transmitter is subject to a total transmission budget $P_{\max}$, enforcing $\sum_i p_i \le P_{\max}$, while the adversary is limited by a total injection capacity $Q_{\max}$, ensuring $\sum_i q_i \le Q_{\max}$. These

constraints model practical limitations such as bandwidth availability, computational capacity, and energy expenditure, and introduce strategic trade-offs in how each player allocates resources across sensors. This setup forms the foundation for the subsequent game-theoretic analysis.

7.4.2 Return Times and System-Level AoII

Consider a single sensor i evolving as in Fig. 12.2a, whose transitions are:

$$R_i \xrightarrow{d_i+q_i} W_i, \quad W_i \xrightarrow{p_i} R_i.$$

A renewal cycle consists of a period in state R_i, of random duration $\tau_{R,i} \sim \mathrm{Exp}(d_i + q_i)$, followed by a period in state W_i, of duration $\tau_{W,i} \sim \mathrm{Exp}(p_i)$. The expected cycle length is

$$\mathbb{E}[T_i] \;=\; \frac{1}{d_i + q_i} + \frac{1}{p_i}. \tag{7.23}$$

During the wrong period, the local AoII $\delta_i(t)$ increases linearly from zero. The area accumulated under the AoII curve in one such episode is

$$A_i = \int_0^{\tau_{W,i}} s \, \mathrm{d}s = \frac{\tau_{W,i}^2}{2}.$$

By taking expectation and using the second moment of the exponential distribution,

$$\mathbb{E}[A_i] \;=\; \frac{1}{2}\,\mathbb{E}[\tau_{W,i}^2] = \frac{1}{2p_i^2}. \tag{7.24}$$

Applying the renewal-reward theorem, the long-run time average AoII of sensor i is the ratio

$$\Delta_{\mathsf{F}}^{i}(p_i, q_i) \;=\; \frac{\mathbb{E}[A_i]}{\mathbb{E}[T_i]} = \frac{d_i + q_i}{2p_i\,(d_i + q_i + p_i)}. \tag{7.25}$$

Aggregating across all sensors, the system-wide performance metric is the arithmetic mean

$$\Delta_{\mathsf{F}}^{tot}(p, q) \;=\; \frac{1}{m} \sum_{i=1}^{m} \frac{d_i + q_i}{2\,p_i\,(d_i + q_i + p_i)}. \tag{7.26}$$

The closed-form expression (7.26) shows that the system AoII increases with larger drift or attack rates $(d_i + q_i)$ and decreases with higher update rates p_i. This trade-off encapsulates the fundamental tension between natural obsolescence, adversarial corruption, and transmitter refresh strategies.

7.4.3 Game-Theoretic Analysis

This section formalizes the strategic interaction between the legitimate transmitter and the adversary as a two–player non-cooperative game. The transmitter allocates refresh rates across sensors under a finite budget, while the adversary injects false data under its own budget. The quality of information available at the monitor is captured by the *expected AoII*, whose closed form was derived in Sec. 7.4.2. We proceed by (1) defining the players utility functions, (2) analyzing a non-adversarial benchmark that clarifies the transmitter's incentives, (3) defining the game formally, and (iv) establishing existence, uniqueness, and structure of the NE, together with an interpretation of the resulting allocation rules.

Strategic Game and Equilibrium Concept We now couple the two decision makers. For each sensor i we define the single-sensor utilities

$$\begin{aligned} u^i_{\mathrm{T}}(p_i, q_i) &= -\Delta^i_{tot}(p_i, q_i) - C_i p_i, \\ u^i_{\mathrm{M}}(p_i, q_i) &= \quad \Delta^i_{tot} p_i, q_i) - K_i q_i, \end{aligned} \tag{7.27}$$

and aggregate them as

$$\begin{aligned} u_{\mathrm{T}}(p, q) &= -\Delta_{tot}(p, q) - C^T p, \\ u_{\mathrm{M}}(p, q) &= \quad \Delta_{tot}(p, q) - K^T q, \end{aligned} \tag{7.28}$$

with $C = (C_1, \ldots, C_m)$ and $K = (K_1, \ldots, K_m)$. The transmitter minimizes AoII while paying for refresh; the adversary maximizes AoII while paying for injections.

The interaction is modeled as a simultaneous-move game

$$\mathcal{G}^M_{\mathsf{F}} = (\mathcal{P}, \mathcal{A}, \mathcal{U}), \tag{7.29}$$

with players $\mathcal{P} = \{\mathrm{T}, \mathrm{M}\}$, action space $\mathcal{A} = \mathbb{R}^m_+ \times \mathbb{R}^m_+$ (subject to the two budget constraints), and utilities (7.28). The players' best-response problems are

$$\max_{p \geq 0} \; -\Delta_{tot}(p, q) - C^T p \quad \text{s.t.} \; \sum_i p_i \leq P_{\max}, \tag{7.30}$$

$$\max_{q \geq 0} \quad \Delta_{tot}(p, q) - K^T q \quad \text{s.t.} \; \sum_i q_i \leq Q_{\max}. \tag{7.31}$$

This formulation is justified by our memoryless arrival/aging assumptions: the Markovian dynamics induce stationary payoffs that are fully summarized by the rate vectors (p, q), so a one-shot game captures the steady-state strategic tension.

Non-Adversarial Benchmark We set $q_i = 0$ for all i to model the absence of the adversary. This isolates the transmitter's incentives in a regular scenario without any adversary. Thus,

$$\Delta^i_{tot}(p_i, 0) = \frac{d_i}{2\, p_i\, (d_i + p_i)}.$$

With the budget P_{max}, the transmitter solves

$$\max_{p \geq 0} \; -\sum_{i=1}^m \Delta_i(p_i, 0) \; - \; C^\top p \quad \text{s.t.} \quad \sum_{i=1}^m p_i \leq P_{\max}.$$

Let $\lambda_T \geq 0$ be the multiplier of the budget. For any active component ($p_i > 0$), KKT stationarity gives

$$-\frac{\partial \Delta^i_{tot}}{\partial p_i}(p_i, 0) = C_i + \lambda_T \iff \frac{d_i(d_i + 2p_i)}{2\, p_i^2\, (d_i + p_i)^2} = C_i + \lambda_T. \tag{7.32}$$

Define $\sigma_i \triangleq C_i + \lambda_T$, $x_i \triangleq p_i/d_i$ (for $d_i > 0$), and $\alpha_i \triangleq 2\, \sigma_i\, d_i^2$. Then (7.32) is equivalent to the quartic equation

$$\alpha_i x_i^4 + 2\alpha_i x_i^3 + \alpha_i x_i^2 - 2x_i - 1 = 0.$$

With the shift $x_i = y_i - \frac{1}{2}$, this becomes

$$y_i^4 + e\, y_i^2 + f_i\, y_i + r = 0, \quad e = -\tfrac{1}{2}, \;\; f_i = -\tfrac{2}{\alpha_i}, \;\; r = \tfrac{1}{16}.$$

Let η_i be the only real positive root of the resolvent cubic

$$\eta^3 + \tfrac{1}{4}\eta^2 - \tfrac{1}{16}\eta - \left(\tfrac{1}{64} + \tfrac{1}{2\alpha_i^2}\right) = 0.$$

Using Cardano's rule, write $\eta_i = \zeta_i - \frac{1}{12}$ with

$$\beta_i = -\tfrac{1}{12}, \quad \gamma_i = -\tfrac{1}{108} - \tfrac{1}{2\alpha_i^2}, \quad \Delta_i = \left(\tfrac{\gamma_i}{2}\right)^2 + \left(\tfrac{\beta_i}{3}\right)^3,$$

$$\zeta_i = \sqrt[3]{-\tfrac{\gamma_i}{2} + \sqrt{\Delta_i}} + \sqrt[3]{-\tfrac{\gamma_i}{2} - \sqrt{\Delta_i}}.$$

Set $A_i = \sqrt{2\eta_i + \frac{1}{2}}$. A positive root of the quartic is

$$y_i^* = \tfrac{1}{2}\left(A_i + \sqrt{\tfrac{1}{2} - 2\eta_i + \tfrac{4}{\alpha_i A_i}}\right),$$

and therefore

$$p_i^{\mathrm{NA}}(\lambda_T) = d_i\big(y_i^* - \tfrac{1}{2}\big)\,. \tag{7.33}$$

Among the algebraic roots, (7.33) is the unique positive one that satisfies (7.32). If $d_i = 0$, then $\Delta_i \equiv 0$ and $p_i^* = 0$.

Let $P(\lambda_T) = \sum_{i=1}^m p_i^{\mathrm{NA}}(\lambda_T)$ with p_i^{NA} given by (7.33). Since each p_i^{NA} is continuous and strictly decreasing in λ_T, $P(\lambda_T)$ is continuous and strictly decreasing. The multiplier is uniquely determined by complementarity:

$$P(\lambda_T^*) \le P_{\max}, \qquad \lambda_T^*\big(P_{\max} - P(\lambda_T^*)\big) = 0.$$

If $P(0) \le P_{\max}$ then $\lambda_T^* = 0$; otherwise there is a unique $\lambda_T^* > 0$ such that $P(\lambda_T^*) = P_{\max}$.

Equation (7.33) makes the economics transparent: larger drift d_i pushes p_i^{NA} up (fast signals merit more refresh), while larger cost C_i or a tighter budget (larger λ_T) pull it down. All effects are monotone, and the scalar budget equation is bisection-friendly.

7.4.4 Main Results

We now aim to prove the existence and uniqueness of the NE for game $\mathcal{G}$ defined in (7.29). The feasible sets for the transmitter and the attacker are the simplexes $\mathcal{P} = \{p \ge 0 : \sum_i p_i \le P_{\max}\}$ and $\mathcal{Q} = \{q \ge 0 : \sum_i q_i \le Q_{\max}\}$, respectively. These sets are nonempty, compact, and convex, which are key properties for a well-posed game.

Theorem 7.4.1 (*Existence and uniqueness of the NE*)

The game is concave and satisfies Rosen's diagonal strict concavity (DSC) [288, Thm. 2]. Therefore, it admits a unique *Nash equilibrium.* □

Proof. For a fixed strategy q, the objective function $\Delta_i(\cdot, q_i)$ is strictly convex in p_i. Consequently, the transmitter's objective $u_{\mathrm{T}}(\cdot, q)$ is strictly concave in p. Similarly, for a fixed strategy p, the attacker's objective $u_{\mathrm{M}}(p, \cdot)$ is strictly concave in q, as each $\Delta_i(p_i, \cdot)$ is strictly concave in q_i. The continuity of these functions on the compact and convex set $\mathcal{P} \times \mathcal{Q}$ is straightforward.

Let the equally-weighted pseudo-gradient be defined as

$$\Phi(p, q) = \begin{bmatrix} \nabla_p u_{\mathrm{T}}(p, q) \\ \nabla_q u_{\mathrm{M}}(p, q) \end{bmatrix}.$$

Since both u_{T} and u_{M} depend on (p, q) primarily through Δ, the Jacobian of the pseudo-gradient, J, takes on a block structure: $J = \begin{bmatrix} -\nabla_{pp}^2 \Delta & -\nabla_{pq}^2 \Delta \\ \nabla_{qp}^2 \Delta & \nabla_{qq}^2 \Delta \end{bmatrix}$. The symmetric part of this Jacobian is

$$\operatorname{sym} J \;=\; \begin{bmatrix} -\nabla_{pp}^2 \Delta & 0 \\ 0 & \nabla_{qq}^2 \Delta \end{bmatrix},$$

because the off-diagonal blocks, which are transposes of each other, cancel out. Based on the second derivatives for a single sensor, we know that $\nabla_{pp}^2 \Delta \succ 0$ (due to strict convexity in p) and $\nabla_{qq}^2 \Delta \prec 0$ (due to strict concavity in q). As a result, $\operatorname{sym} J \prec 0$ over

the interior of $\mathcal{P} \times \mathcal{Q}$. This proves that Φ is strictly monotone satisfying the Diagonal Strict Concavity condition. Rosen's theorem [288] then guarantees both the existence and, crucially, the *uniqueness* of the NE on the feasible sets $\mathcal{P} \times \mathcal{Q}$.

□

The above result, guarantees that for any given set of parameters, there is a single equilibrium point (p^*, q^*). This ensures that comparative statics analysis is well-posed and that numerical methods will converge to the same solution regardless of their initial state.

Let $(\lambda_T^*, \lambda_M^*) \geq 0$ be the Lagrange multipliers associated with the transmitter and attacker budgets, respectively. The KKT conditions provide the following equilibrium strategies for each sensor i the following injection rate:

$$q_i^* \;=\; \max\left\{0,\; -d_i - p_i^* + \frac{1}{\sqrt{2\,(K_i + \lambda_M^*)}}\right\}. \tag{7.34}$$

This indicates that attacks occur only if the sensor's natural protection $(d_i + p_i^*)$ is sufficiently low relative to the effective inverse cost for the attacker. When an attack is active $(q_i^* > 0)$, the transmitter's strategy is given by the interior rule:

$$p_i^* \;=\; \frac{1}{\sqrt{2\,(C_i + \lambda_T^* + K_i + \lambda_M^*)}}. \tag{7.35}$$

This resembles a water-filling law, where the transmitter's price $(C_i + \lambda_T^*)$ is augmented by the attacker's effective price $(K_i + \lambda_M^*)$. If the attack is inactive $(q_i^* = 0)$, the transmitter reverts to the optimization without an adversary, giving $p_i^{\mathrm{NA}}(\lambda_T^*)$, which is defined in (7.33).

Thus, sensors at equilibrium are either contested, where both players invest according to the square-root laws above, or uncontested, where the transmitter refreshes them as if there were no attacker.

The multipliers $(\lambda_T^*, \lambda_M^*)$ are uniquely determined by the complementarity slackness conditions:

$$\sum_i p_i^* \leq P_{\max},\;\; \lambda_T^*\big(P_{\max} - \sum_i p_i^*\big) = 0,$$
$$\sum_i q_i^* \leq Q_{\max},\;\; \lambda_M^*\big(Q_{\max} - \sum_i q_i^*\big) = 0.$$

These multipliers act as system-wide shadow prices [289]. For example, increasing $P_{\max}$ weakly decreases λ_T^*, which in turn drives all p_i^* up (leading to more frequent updates and a lower AoI). Symmetrically, an increase in $Q_{\max}$ weakly decreases λ_M^*, which makes attacks more profitable and tends to increase q_i^* (unless local costs K_i are already too high). The coupling between p_i^* and q_i^* through equations (7.34) and (7.35) means that relaxing one budget triggers a predictable rebalancing of efforts on both sides.

At the NE, each sensor falls into one of two regimes:

- Contested: $q_i^* > 0$. Both players invest resources according to the square-root rules. Increasing the local costs (K_i or C_i) shifts effort away from sensor i.
- Uncontested: $q_i^* = 0$. The attacker does not find it profitable to invest, and the transmitter's rate is exactly the no-attack optimum.

Transitions between these regimes are threshold-like. Once $d_i + p_i^*$ is sufficiently large or K_i is high enough, the attacker receives no return on investment for sensor i. A conservative sufficient condition can be used to guarantee that a sensor will not be attacked at the NE.

Lemma 7.4.1 (*Conservative hardening threshold*)

Let $\bar{p}_i \triangleq p_i^{\mathrm{NA}}(0)$ be the optimum in the absence of an adversary when the budget

constraint is inactive ($\lambda_T = 0$). If

$$K_i \;\geq\; \frac{1}{2\left(d_i + \bar{p}_i\right)^2}, \tag{7.36}$$

then $q_i^* = 0$ for all feasible $(\lambda_T^*, \lambda_M^*) \geq 0$. □

Proof. If we assume $q_i^* > 0$, the KKT conditions for the attacker imply $d_i + p_i^* + q_i^* = 1/\sqrt{2(K_i + \lambda_M^*)}$. Since $p_i^* \leq \bar{p}_i$ and $\lambda_M^* \geq 0$, the inequality in (7.36) leads to a contradiction. Therefore, the assumption $q_i^* > 0$ must be false. □

Inequality (7.36) provides a practical rule-of-thumb: if the local hardening cost K_i exceeds this bound, sensor i will never be attacked at equilibrium, regardless of the global budgets. Thus, it can be safely treated using the no-attack branch for the transmitter's strategy p_i^*. This gives a conservative criterion for when an adversary will never target a sensor, regardless of its budget: a practically useful design rule.

Remark 7.4.1

The drift rate d_i is a modeling parameter *tied to the monitored quantity and operating regime; it need not be universal nor fixed across all conditions. In practice, one may calibrate d_i differently across operating phases (e.g., start–up, steady state, disturbance handling). Our analysis should thus be read* pointwise in the parameter*: for any fixed specification of (d_i, C_i, K_i) the game is concave with Rosen DSC and admits a* unique *NE. If the regime changes, one simply resolves the same one–shot game with the updated d_i, thereby tracking the equilibrium across regimes. This interpretation naturally covers cases where, near a nominal setpoint, delayed data have limited impact (smaller effective d_i), while during transients or excursions timely updates are critical (larger effective d_i). For clarity and tractability, in the main development we keep d_i fixed within each analysis window; allowing piecewise–constant or slowly varying specifications amounts to recomputing the (unique) instantaneous equilibrium as parameters change.* □

Theorem 7.4.2 (*Fast global convergence to the NE*)

Under the assumptions of Thm. 7.4.1 (Rosen DSC) and smoothness of $u_{\mathrm{T}}, u_{\mathrm{M}}$, let $\Phi(p,q) = \left[\nabla_p u_{\mathrm{T}}(p,q)^\top,\ \nabla_q u_{\mathrm{M}}(p,q)^\top\right]^\top$, and consider the iteration

$$p^{k+1} = \Pi_{\mathcal{P}}\big(p^k + \eta\, \nabla_p u_{\mathrm{T}}(p^k, q^k)\big),$$
$$q^{k+1} = \Pi_{\mathcal{Q}}\big(q^k + \eta\, \nabla_q u_{\mathrm{M}}(p^k, q^k)\big).$$

Then there exist constants $\mu > 0$ and $L < \infty$ such that Φ is μ–strongly monotone and L–Lipschitz on $\mathcal{P} \times \mathcal{Q}$. For any fixed stepsize $\eta \in (0,\, 2\mu/L^2)$, the iterates converge globally and linearly *to the unique Nash equilibrium (p^*, q^*):*

$$\|(p^k, q^k) - (p^*, q^*)\| \;\leq\; \rho^k\, \|(p^0, q^0) - (p^*, q^*)\|,$$

where $\rho = \sqrt{1 - 2\eta\mu + \eta^2 L^2}\ < 1$. □

Proof. We rewrite the problem as a single variational inequality over X to streamline the analysis. Since $u_{\mathrm{T}}(\cdot, q)$ and $u_{\mathrm{M}}(p, \cdot)$ are concave with convex feasible sets, a point $z^* = (p^*, q^*) \in X$ is a Nash equilibrium iff each player's problem satisfies the first-order optimality condition:

$$\langle \nabla_p u_{\mathrm{T}}(p^*, q^*),\, p - p^* \rangle \;\leq\; 0 \quad \forall p \in \mathcal{P},$$

$$\langle \nabla_q u_{\mathrm{M}}(p^*, q^*),\, q - q^* \rangle \;\leq\; 0 \quad \forall q \in \mathcal{Q}.$$

Stacking these two inequalities gives the compact form

$$\langle \Phi(z^*),\, z - z^* \rangle \;\le\; 0 \qquad \forall z = (p, q) \in X. \tag{7.37}$$

Define $T(z) \triangleq \Pi_X\big(z + \eta\, \Phi(z)\big)$. By the optimality condition of Euclidean projection, $z^* = \Pi_X\,(z^* + \eta\Phi(z^*))$ is equivalent to

$$\langle z^* + \eta\Phi(z^*) - z^*,\, z - z^* \rangle \;\le\; 0 \quad \forall z \in X,$$

which is exactly (7.37). Thus the fixed points of T are precisely the Nash equilibria. By strong monotonicity of Φ, the equilibrium is unique.

By nonexpansiveness of projection,

$$\begin{aligned}\|T(z) - T(w)\| \le \|\, z + \eta\Phi(z) - w - \eta\Phi(w)\| = \\ = \|(I + \eta\Phi)(z) - (I + \eta\Phi)(w)\|.\end{aligned}$$

Squaring and expanding, and using μ–strong monotonicity and L–Lipschitzness of Φ,

$$\|T(z) - T(w)\|^2 \le \big(1 - 2\eta\mu + \eta^2 L^2\big)\, \|z - w\|^2.$$

Thus T is a strict contraction whenever $0 < \eta < 2\mu/L^2$, with contraction factor

$$\rho = \sqrt{1 - 2\eta\mu + \eta^2 L^2} \;< 1.$$

By Banach's fixed-point theorem, T has a unique fixed point z^* and $z^{k+1} = T(z^k)$ converges to z^* for any $z^0 \in X$ at the linear rate stated. By fixed-point interaction, this fixed point satisfies the players' first-order conditions, hence it is the unique Nash equilibrium. □

Remark 7.4.2

A robust choice is $\eta = \dfrac{\mu}{L^2}$ *(which does not require fine tuning). In practice, a short backtracking line search quickly finds a near-optimal* η*. The stepsize bound* $\eta \le 2\mu/L^2$ *ensures that the projected gradient–play map is a strict contraction, so the learning dynamics of the players remain stable. Intuitively,* μ *measures how strongly the game pulls strategies back toward the Nash equilibrium (stability margin), while* L *measures how fast the pseudo-gradient field* Φ *can change (sensitivity of the system). A larger* μ *allows more aggressive updates, while a larger* L *forces smaller updates to avoid overshooting.* □

In the context of dynamical systems, one can interpret the gradient–play iteration as a discrete-time feedback mechanism by which the system adapts its strategies (or control variables) online. The condition $\eta \le 2\mu/L^2$ plays the role of a "gain margin": it guarantees that the feedback loop is stable, while still allowing sufficiently large updates to track changes in the underlying dynamics. Choosing η close to the upper bound leads to fast convergence, enabling the system to react quickly to environmental or parameter changes (i.e., the drift d_i), whereas choosing η too large would cause oscillations or even divergence.

7.4.5 Numerical diagnostics for the multipliers

It is convenient to visualize the two scalar equations that determine the multipliers:

$$\begin{aligned}\mathcal{L}_{\lambda_T}(\lambda_T; \lambda_M) \;&\triangleq\; P_{\max} - \sum_{i=1}^{m} p_i^*(\lambda_T, \lambda_M), \\ \mathcal{L}_{\lambda_M}(\lambda_M) \;&\triangleq\; Q_{\max} - \sum_{i=1}^{m} q_i^*\big(\lambda_T^*(\lambda_M), \lambda_M\big).\end{aligned}$$

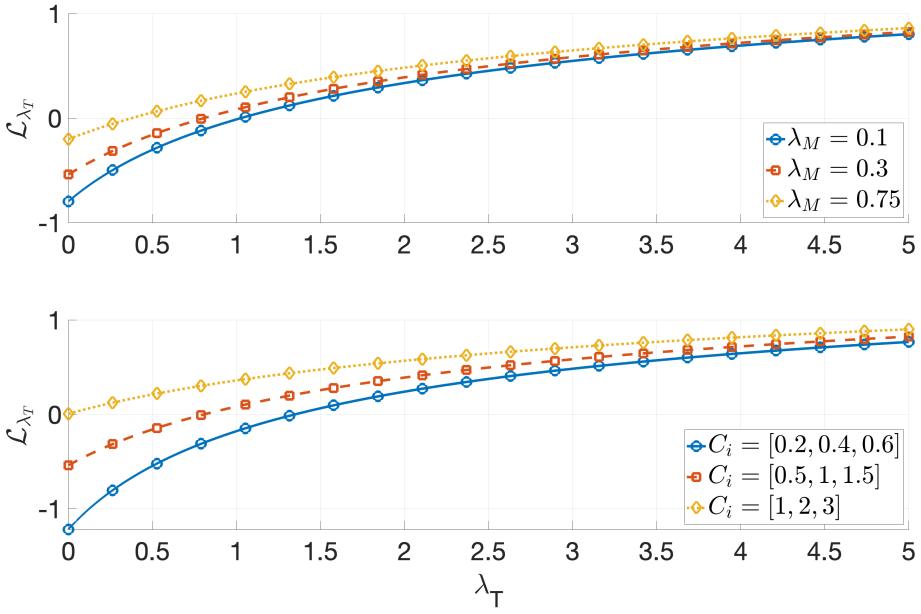


Figure 7.16: $\mathcal{L}_{\lambda_T}(\lambda_T; \lambda_M)$ for different (λ_M, C_i).

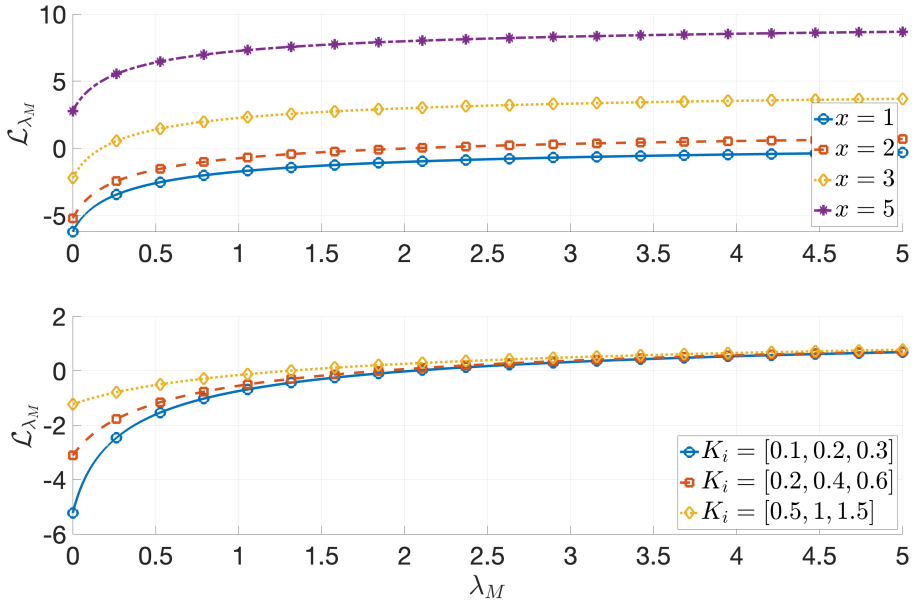


Figure 7.17: $\mathcal{L}_{\lambda_M}(\lambda_M)$ evaluated under different total values $x = P_{\max} + Q_{\max} + \sum_i d_i$ and attack costs K_i.

For any fixed $\lambda_M \geq 0$, $\mathcal{L}_{\lambda_T}(\cdot; \lambda_M)$ is *continuous and strictly increasing* because each p_i^* is continuous and strictly decreasing in λ_T. Hence it has at most one zero, and bisection finds the unique $\lambda_T^*(\lambda_M)$ when the budget binds. Likewise, $\mathcal{L}_{\lambda_M}(\cdot)$ is *continuous and strictly increasing* since each q_i^* decreases with λ_M (on the interior branch) and drops to zero at a finite threshold; thus λ_M^* is also a unique 1-D root when the attacker budget binds.

Figure 7.16 depicts the residual transmitter budget $\mathcal{L}_{\lambda_T}$ as λ_T varies, for several choices of λ_M and cost vectors C. In all panels the curve rises monotonically: when it crosses 0 there is exactly one solution λ_T^*; if it stays strictly positive, the budget is slack and $\lambda_T^* = 0$. Figure 7.17 shows the same logic for the attacker: the monotonically increasing $\mathcal{L}_{\lambda_M}$ either crosses 0 once (budget binds $\Rightarrow \lambda_M^* > 0$) or stays > 0 (slack

$\Rightarrow \lambda_M^* = 0$). These diagnostics make the fixed-point search transparent and numerically robust (simple bisection on an increasing continuous scalar function).

Two conservative tests decide at a glance whether a multiplier is forced to zero:

$$(M)\ \sum_{i=1}^{m} \max\left\{0, \frac{1}{\sqrt{2K_i}} - d_i\right\} \le Q_{\max} \implies \lambda_M^* = 0. \tag{7.38}$$

$$(T)\ \sum_{i=1}^{m} \frac{1}{\sqrt{2\,(C_i + K_i)}} \le P_{\max} \implies \lambda_T^* = 0. \tag{7.39}$$

In (7.38), the bound ignores transmitter refresh ($p_i^* \ge 0$), so it overestimates the total attack effort at $\lambda_M = 0$; if even that upper bound fits the attacker budget, the budget cannot bind. In (7.39), the bound takes the largest interior-branch refresh at $(\lambda_T, \lambda_M) = (0, 0)$; any positive multiplier can only *reduce* p_i^*, so if even this sum is below $P_{\max}$, the transmitter budget is slack. When either inequality is strict, the corresponding multiplier is necessarily zero and the closed forms in (7.34)–(7.35) specialize with that multiplier set to 0.

- *Sweep λ_M at several C vectors* and plot $\mathcal{L}_{\lambda_T}(\lambda_T; \lambda_M)$ as in Figure 7.16. Expect a unique crossing when the budget binds; for larger C the curve shifts up and crossings disappear (slack budget).
- *Sweep λ_T at several K sets* and plot $\mathcal{L}_{\lambda_M}(\lambda_M)$ as in Figure 7.17. Larger K (harder-to-attack sensors) shift the curve upward, making λ_M^* vanish over a wider parameter range.
- *Vary totals $P_{\max}, Q_{\max}$* to visualize transitions between binding and slack regimes. As $P_{\max}$ increases, the zero of $\mathcal{L}_{\lambda_T}$ moves left and eventually disappears ($\lambda_T^* \to 0$). As $Q_{\max}$ increases, the zero of $\mathcal{L}_{\lambda_M}$ moves left and then appears ($\lambda_M^* \to 0$).
- *Cross-check the slack tests.* Overlay vertical lines at the parameters where (7.38) or (7.39) start to hold, and verify that the corresponding curves are positive at zero and never cross the axis.

Remark 7.4.3

If $\mathcal{L}_{\lambda_M}(0) \ge 0$ then $\lambda_M^* = 0$ and the attacker budget is slack; otherwise there is a unique *$\lambda_M^* > 0$ with $\mathcal{L}_{\lambda_M}(\lambda_M^*) = 0$. The exact same dichotomy holds for $\mathcal{L}_{\lambda_T}$. This "single crossing or no crossing" behavior is a direct consequence of the continuity and strict monotonicity induced by the best-response structure in* (7.34)–(7.35). □

7.4.6 Numerical Results

This section presents numerical simulations that validate the theoretical findings and provide insight into the equilibrium behavior of the transmitter–adversary interaction in a multi-sensor environment. The objective is to characterize how equilibrium strategies evolve with varying system parameters—drift rates d_i, transmission costs C_i, and injection costs K_i—and how both agents allocate their constrained resources across heterogeneous sensors under asymmetric conditions.

We consider a system composed of $m = 3$ sensors, each characterized by distinct parameter values. The Nash Equilibrium (NE) is obtained by numerically solving the coupled system of Karush–Kuhn–Tucker (KKT) equations derived in Section 7.4.3, which define optimal strategies through the associated Lagrange multipliers.

We begin by examining the effect of the available transmission budget $P_{\max}$ on equilibrium strategies. The injection budget is fixed to $Q_{\max} = 0.5$, while $P_{\max}$ varies in the range $[0.2, 2]$. The system parameters are set to $d = [0.1, 0.2, 0.3]$, $C = [1, 2, 3]$, and $K = [0.1, 1, 1.5]$. As shown in Figure 7.18, when $P_{\max}$ is small, the transmitter's budget constraint is active and all resources are fully utilized. The transmitter allocates most

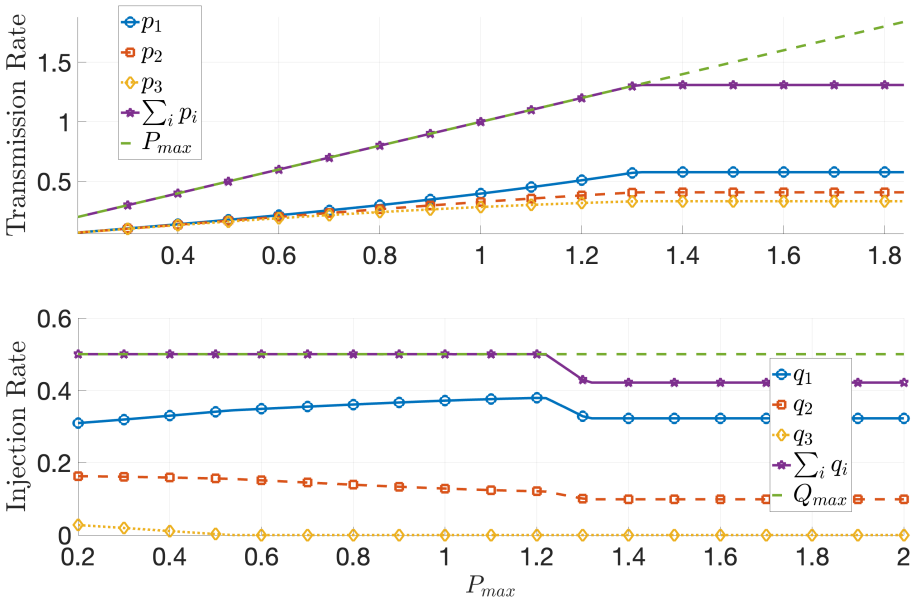


Figure 7.18: Equilibrium transmission p_i and injection rates q_i as a function of transmission budget $P_{\max}$, with fixed $Q_{\max} = 0.5$.

of its effort to sensors that are both cost-efficient and vulnerable—those with lower C_i and K_i. As $P_{\max}$ increases, these sensors receive proportionally more updates, while less critical sensors remain under-provisioned. For the adversary, the injection resources remain saturated up to approximately $P_{\max} \approx 1.2$, beyond which additional transmitter power reduces the effectiveness of attacks. The adversary thus gradually decreases its overall effort, prioritizing Sensor 1 (low d_1, low K_1) and eventually abandoning Sensor 3, which becomes unprofitable under higher transmission activity.

We next fix the transmission budget to $P_{\max} = 2$ and vary the injection budget $Q_{\max} \in [0.1, 2]$. Figure 7.19 shows that the adversary increases its injection rates until its constraint becomes inactive at $Q_{\max} \approx 1.85$. Before this threshold, the attacker fully exploits its available resources, focusing on sensors with low drift and injection costs. The transmitter, whose constraint remains inactive, reacts strategically to adversarial escalation: as attacks intensify, transmission rates increase, reaching a steady-state level once the marginal benefit of additional effort vanishes. This behavior highlights a saturation effect in which both players' optimal actions stabilize as their respective Lagrange multipliers approach zero.

We then investigate the sensitivity of equilibrium strategies to the drift rate of individual sensors. Keeping $d_1 = 0.1$, $d_3 = 0.3$, and varying $d_2 \in [0.1, 2]$, with $C = [1, 2, 3]$, $K = [0.1, 1, 1.5]$, $P_{\max} = 2$, and $Q_{\max} = 0.5$, the results in Figure 7.20 show that as d_2 increases, the attacker reallocates resources toward Sensors 1 and 3, which now provide greater payoff due to slower natural degradation. Meanwhile, transmission rates exhibit limited adaptation, constrained by the adversary's active resource limit ($\lambda_M > 0$). This asymmetry demonstrates that when one player operates under an active budget constraint, its ability to respond to parameter variations dominates the system's equilibrium adjustments.

We next vary the injection cost K_1 to evaluate how economic deterrence affects adversarial targeting. As shown in Figure 7.21, as K_1 increases from 0.1 to 8, the adversary gradually reduces its attack intensity on Sensor 1 and eventually ceases all injections for $K_1 > 6$. This withdrawal triggers a redistribution of effort toward Sensors 2 and 3. The transmitter, observing reduced threat on Sensor 1, correspondingly lowers p_1 and reallocates its budget to reinforce more exposed nodes. The turning point, where λ_M vanishes, marks the

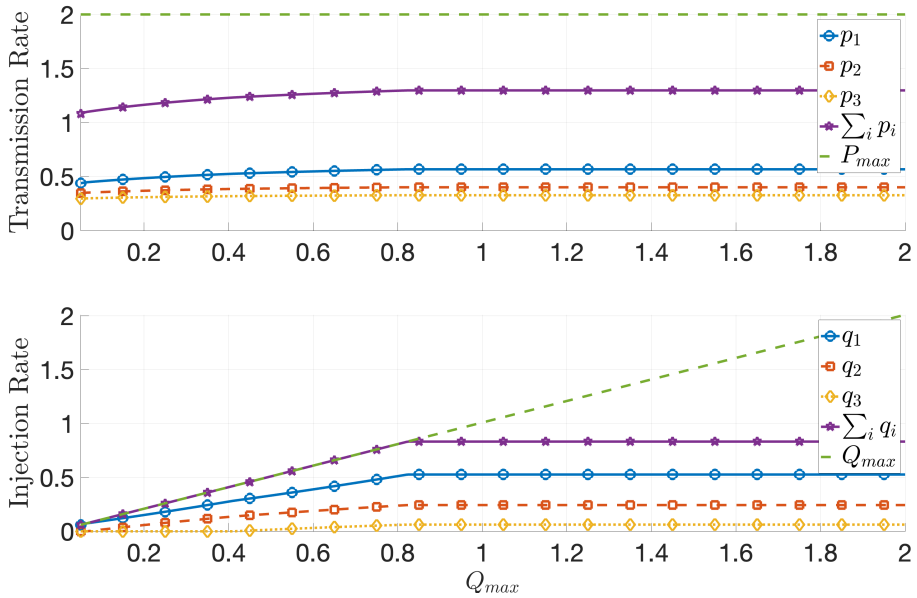


Figure 7.19: Equilibrium transmission p_i and injection rates q_i as a function of injection budget $Q_{\max}$, with fixed $P_{\max} = 2$.

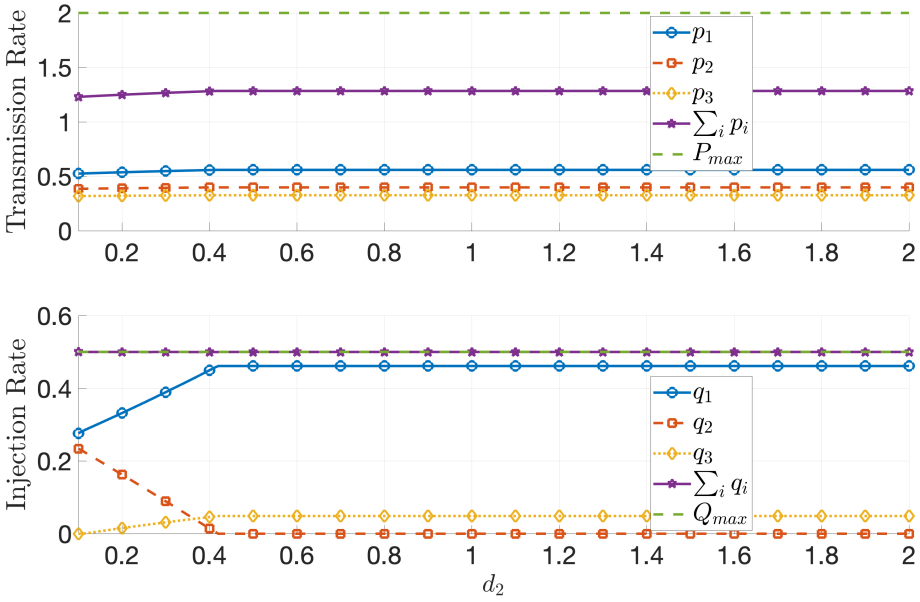


Figure 7.20: Equilibrium transmission p_i and injection rates q_i as a function of sensor drift d_2.

transition from resource-limited to unconstrained adversarial behavior—an indicator of strategic saturation.

A complementary experiment explores the impact of transmission cost C_3. As depicted in Figure 7.22, increasing C_3 progressively reduces p_3, prompting the transmitter to favor less expensive sensors. This reduced update frequency makes Sensor 3 increasingly appealing to the adversary, who raises q_3 correspondingly. For sufficiently large C_3, the attack intensity on Sensor 3 surpasses that on Sensor 2, reflecting a shift in adversarial targeting toward the weakest communication link. These results demonstrate that uneven transmission costs directly shape system vulnerability and that adversaries naturally exploit underprotected channels.

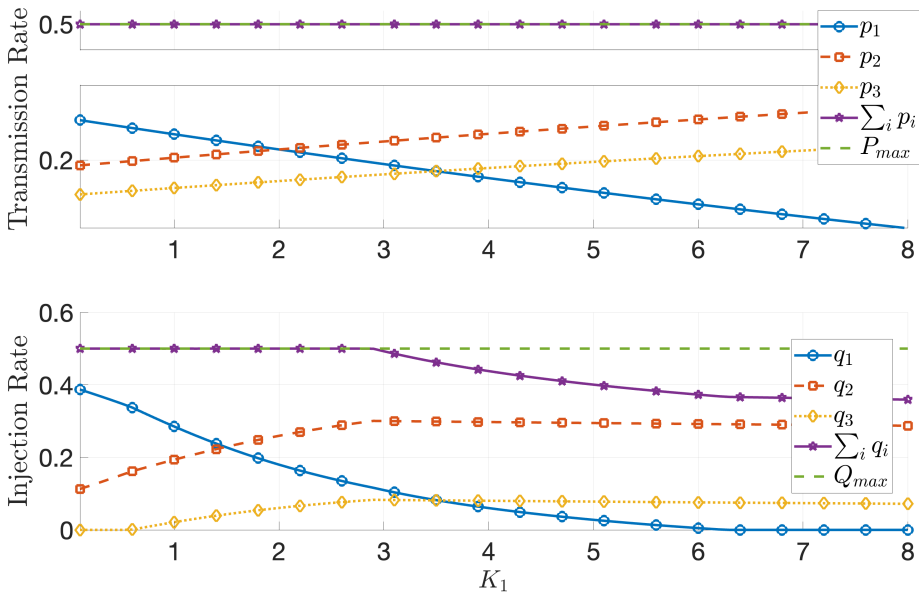


Figure 7.21: Equilibrium transmission p_i and injection rates q_i as a function of injection cost K_1, with $P_{\max} = Q_{\max} = 0.5$.

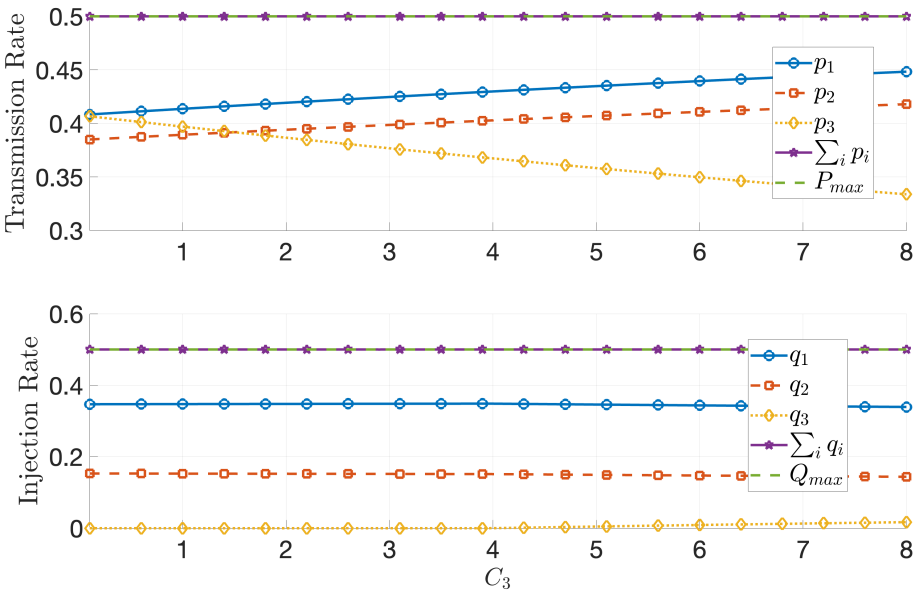


Figure 7.22: Equilibrium transmission p_i and injection rates q_i as a function of transmission cost C_3, with $P_{\max} = Q_{\max} = 0.5$.

Finally, we assess the effect of a uniform scaling of all transmission costs, $C_i = C_b \cdot k$, with $C_b = [1, 2, 3]$ and k varying continuously. Figure 7.23 shows that increasing k leads to a near-uniform decrease in transmission activity across sensors, as higher costs uniformly suppress update incentives. In response, the adversary intensifies its attacks, particularly focusing on Sensor 1, which remains the most profitable due to its low drift. This behavior exposes a systemic vulnerability: uniform cost increases without targeted reinforcement degrade global resilience and amplify adversarial leverage. Effective protection therefore requires asymmetric defense strategies that prioritize high-risk sensors rather than applying uniform policies across the network.

Overall, the numerical results reveal several consistent patterns. (i) Sensors with low drift,

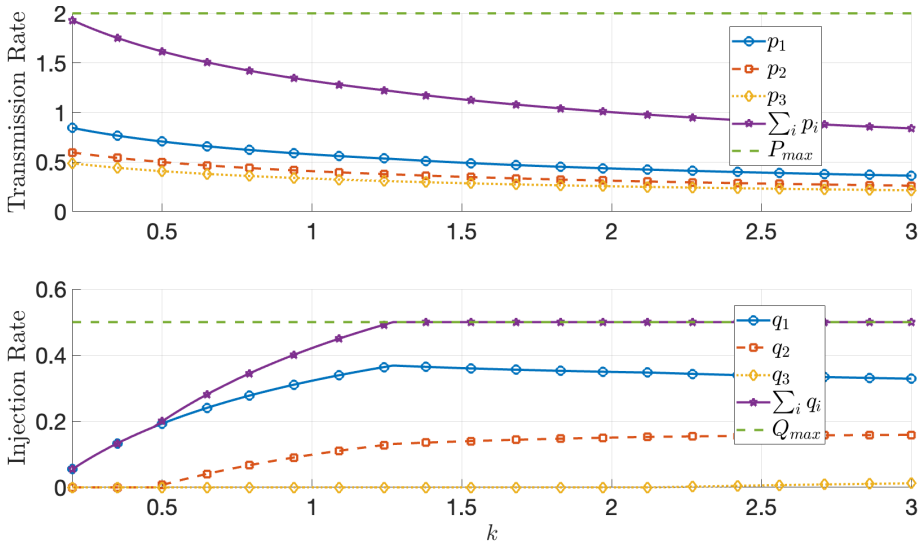


Figure 7.23: Equilibrium transmission p_i and injection rates q_i as a function of global cost scaling $C_i = C_b \cdot k$, with $P_{\max} = 2$, $Q_{\max} = 0.5$.

low transmission cost, and low injection cost attract the highest concentration of both transmission and attack activity, identifying them as strategic bottlenecks in the network. (ii) Active budget constraints strongly determine equilibrium sensitivity: once inactive, further budget increases yield negligible improvement. (iii) Increasing a sensor's injection cost or maintaining high transmission frequency effectively deters attacks, whereas higher transmission costs without compensatory measures exacerbate vulnerability. (iv) Uniform cost escalation weakens the entire network, underscoring the importance of asymmetric protection policies.

7.5 Discussion

The analysis presented throughout this chapter establishes a comprehensive game-theoretic framework for understanding the impact of false data injection attacks on information freshness and correctness in networked cyber–physical systems. By coupling stochastic models of information dynamics with strategic interaction, the study reveals how both the transmitter and the adversary adapt their behaviors in response to environmental parameters, cost asymmetries, and uncertainty.

A central outcome of the analysis is the formalization of the *Age of Incorrect Information* as a correctness-aware extension of the classical Age of Information. This metric provides a unified view of temporal and semantic degradation, explicitly distinguishing between natural aging due to system drift and malicious degradation due to data corruption. The introduction of AoII allows for the quantitative comparison of competing effects: while frequent updates mitigate temporal staleness, they may simultaneously increase exposure to injection attempts, thereby illustrating the fundamental trade-off between freshness and integrity.

The static game formulation between transmitter and attacker captures the essence of this trade-off. The existence of a unique Nash Equilibrium in the single-sensor setting ensures that both agents' strategies are predictable and analytically tractable, despite their conflicting objectives. From the transmitter's perspective, frequent updates serve as a deterrent mechanism, effectively reducing the window of opportunity for successful injections. Conversely, the adversary's optimal strategy balances the cost of injection against the potential benefit of misleading the receiver—resulting in selective, cost-efficient attack patterns that maximize disruption under resource constraints.

Extending the model to incomplete information enriches this strategic landscape. In the Bayesian setting, the transmitter no longer has full knowledge of the attacker's cost structure and must instead reason about a probability distribution over possible adversarial types. Numerical results demonstrate that the equilibrium remains stable even under significant uncertainty: when the variance of the attacker's cost distribution is moderate, the transmitter's best response is well approximated by assuming the expected attack cost. Higher uncertainty, however, consistently favors the adversary, as it allows sporadic but potent attacks that inflate the expected AoII. These findings underline the importance of belief modeling and adaptive estimation in security-aware communication design.

The multi-sensor generalization reveals an additional structural property: the equilibrium allocation of transmission and injection resources is inherently asymmetric. Both agents concentrate their efforts on a limited subset of sensors—typically those with low drift, low cost, or high informational value—leading to spatially localized conflict zones. Such concentration effects highlight that uniform or symmetric protection strategies are inefficient. Instead, resilience should be pursued through adaptive prioritization, directing energy and transmission capacity toward sensors that exhibit high vulnerability or high system relevance.

Another key insight concerns the interplay between physical dynamics and communication economics. The equilibrium depends jointly on the drift rates of the underlying processes and the relative transmission and injection costs. Fast-evolving processes naturally limit the attacker's incentive to corrupt data, as the underlying information becomes obsolete more rapidly, while higher transmission costs weaken the transmitter's ability to defend freshness. This interdependence emphasizes that resilience cannot be achieved solely through protocol design but must also account for the physical characteristics of the monitored system.

Beyond its immediate application to FDI scenarios, the framework developed in this chapter has broader methodological relevance. It bridges queuing and control-theoretic notions of information timeliness with game-theoretic reasoning, producing a tractable yet expressive model for adversarial communication. By combining closed-form equilibrium analysis with simulation-based validation, the results demonstrate how rational agents—both legitimate and malicious—shape the temporal reliability of distributed information systems.

Overall, this chapter shows that FDI attacks introduce not only additive errors but strategic distortions in how information evolves and is maintained across a network. Optimal defense therefore requires understanding the attacker's incentives and integrating correctness-aware metrics such as AoII into the design of update and scheduling policies. The next chapter extends these insights to more complex adversarial models, where incomplete information, resource coupling, and inter-sensor dependencies further influence the stability and robustness of equilibrium strategies.

Chapter 8

Denial-of-Service Attacks

This chapter focuses on the impact of *Denial-of-Service* (DoS) attacks on the timeliness and reliability of information updates in networked cyber–physical systems. Unlike False Data Injection attacks, which compromise information *correctness* by altering transmitted content, DoS attacks degrade *freshness* by obstructing the communication channel and preventing valid updates from being successfully delivered. The resulting disruption prolongs the persistence of outdated or invalid data at the receiver, thereby indirectly increasing the *Age of Incorrect Information* .

To capture this effect, the stochastic framework developed in the previous chapter is adapted to model communication blocking rather than data corruption. The DoS mechanism is represented as a modification of the transition structure in the two-state Markov chain governing information correctness: while natural drift still drives the system toward obsolescence, the jammer suppresses the restoring transitions by reducing the effective rate of successful updates. This leads to a compact analytical expression for the AoII as a function of both the transmitter's update rate and the jammer's activity level.

The transmitter and the jammer are then modeled as rational decision-makers pursuing opposing objectives: the former seeks to minimize the AoII while accounting for its communication cost, whereas the latter aims to maximize it subject to its own operational constraint. This interaction gives rise to a static, non-cooperative game whose equilibrium captures the balance between proactive updating and adversarial interference. The analysis reveals a critical cost threshold that determines whether jamming remains economically viable, providing direct design insight into how system parameters influence resilience.

Organization of the chapter. The structure of this chapter is as follows:

- **Section 8.0.1** presents the system model under DoS interference and derives the closed-form expression of the AoII as a function of the transmitter's update rate, the jammer's activity, and the natural drift of the monitored process.
- **Section 8.1** introduces the static game formulation between transmitter and jammer, proves the existence and uniqueness of the Nash Equilibrium, and identifies the threshold conditions separating active and inactive attack regimes.
- **Section 8.1.1** illustrates the equilibrium behavior through numerical simulations, highlighting how communication cost, drift dynamics, and attack efficiency shape information freshness under DoS conditions.
- **Section 8.1.2** concludes the chapter by discussing the main findings and their implications for the design of cost-aware, resilient communication strategies in adversarial environments.

8.0.1 System Modeling under DoS

At a specific time, an unauthorized entity gains entry to the communication channel, aiming to disrupt the connection through a DoS attack. Specifically, the malicious agent, termed the jammer, operates by causing each packet sent by the legitimate transmitter to face a probability f of failing to reach its destination. While the jammer's actions cannot be directly blocked, the transmitter can detect its presence and strategically increase its data transmission rate to counteract the malicious interference. We assume no propagation delay between the transmitter and receiver, allowing time calculations to be conducted from either end of the communication channel.

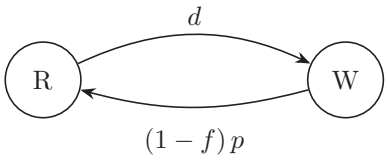


Figure 8.1: Continuous-time Markov chain for information correctness under Denial-of-Service attack. The transition R → W occurs at the drift rate d, while the W → R transition (restoration of correctness) is reduced by the jamming attack's failure rate f.

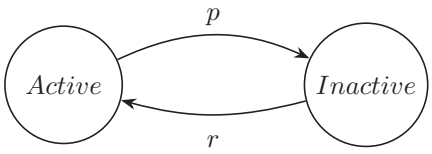


Figure 8.2: Continuous-time Markov chain determining the malicious agent's transition between active and inactive states. The rate r signifies the attacker's activity level, increasing its presence in the active (jamming) state.

As depicted in Figure 8.1, the system transitions from the R state to the W state due to the natural drift of the physical process, occurring at rate d. The jammer's activity directly affects the return transition from W to R. If the transmitter sends updates at rate p, and the jamming attack causes a failure probability of f, then the effective rate of successful updates, which restore correctness, becomes $(1-f)p$. This effectively reduces the frequency of transitions to the R state.

The relationship between the jammer's activity rate and the overall failure probability is crucial. The jammer's activity is quantified by its rate $r \in (0, \infty)$, where a higher r implies a more active malicious agent. This jamming activity r determines the fraction of blocked transmissions, f, through the following relationship:

$$f = \frac{r}{p+r}\,. \tag{8.1}$$

This formulation highlights that as r tends to zero, the failure rate f approaches zero, indicating no interference. Conversely, as r tends to infinity, f approaches 1, signifying a complete disruption of communication. This definition ensures that $f \in [0, 1)$, aligning with the properties required for a probability.

To further characterize the malicious agent's behavior and justify the parameter r, we can envision its operational status as another continuous-time Markov chain between an *Active* and an *Inactive* state, as illustrated in Figure 8.2. The jammer, when in the Active state, is capable of disrupting transmissions. It transitions from Active to Inactive at a rate determined by its internal strategy or resource availability. Conversely, it transitions from Inactive to Active at a rate r. This rate r effectively represents the jammer's activity rate or intensity: a higher r means the jammer is more frequently in, or quickly returns to, its active jamming state, leading to a higher overall probability of blocking transmissions as expressed in Equation (8.1).

As detailed in Section 6.2, we use the Age of Incorrect Information, $\Delta_{\mathsf{D}}(t)$, as our performance metric. For the purpose of this DoS analysis, recall that $\ell(t)$ is a linear time-increasing penalty function and $g(\cdot)$ is a binary function indicating correctness. The expected AoII, $\Delta_{\mathsf{D}} = \mathbb{E}_t[\Delta_{\mathsf{D}}(t)]$, is given by:

$$\Delta_{\mathsf{D}} = \frac{1/(2((1-f)p)^2)}{\frac{1}{(1-f)p} + \frac{1}{d}} = \frac{d}{2(1-f)p(d+(1-f)p)}\,. \tag{8.2}$$

Substituting the expression for f from Equation (8.1) into Equation (8.2), the expected

AoII for the DoS scenario can be explicitly written as:

$$\Delta_{\mathsf{D}} = \frac{d(p+q)^2}{2p^2\big(p^2 + d(p+q)\big)} \tag{8.3}$$

This equation quantifies the system's information freshness as a function of both the transmitter's and jammer's chosen rates.

8.1 Game-Theoretic Formulation and Nash Equilibrium

In the case of a Denial-of-Service attack, the adversary operates by intermittently blocking the transmission channel, preventing a fraction of packets from reaching the receiver. By (8.1) the corresponding *activity rate* of the attacker is defined by the general framework as:

$$r = \gamma_{\mathsf{D}}(q, f, p) = p\,\frac{f}{1-f}, \tag{8.4}$$

where p is the transmitter's update rate, f is the probability of packet failure, and r quantifies the jammer's activity level. This relation expresses the effective effort required by the attacker to sustain a failure probability f, which increases nonlinearly with both p and f. Eq. (8.4) quantifies the enemy's required activity rate to induce a failure probability f, given the legitimate agent's transmission rate p. The rationale behind this choice for the malicious agent activity rate r follows [290].

We now model the strategic interaction between the legitimate transmitter (T) and the malicious jammer (M) as a static, non-cooperative game of complete information, denoted by

$$\mathcal{G}_{\mathsf{D}} = (\mathcal{P}, \mathcal{A}, \mathcal{U}).$$

The set of players is $\mathcal{P} = \{\mathrm{T}, \mathrm{M}\}$. Player T chooses a transmission rate $p \in [0, \infty)$, while player M selects a jamming activity rate $q \in [0, \infty)$. The game is *static*, meaning both players act simultaneously and independently, without knowledge of the opponent's choice in the current round.

Following the general modeling approach in Section 6.2, both agents incur proportional costs. The transmitter pays a unit cost $C > 0$ per transmission rate, while the jammer bears a unit cost $K > 0$ per activity rate. Accordingly, the utility functions are defined as:

$$u_{\mathrm{T}}(p, r) = -\Delta_{\mathrm{D}}(p, r) - C\,p, \qquad u_{\mathrm{M}}(p, r) = \Delta_{\mathrm{D}}(p, r) - K\,r, \tag{8.5}$$

where $\Delta_{\mathrm{D}}(p, r)$ is the steady-state Age of Incorrect Information under DoS interference, given by Equation (8.3). The transmitter aims to *minimize* its AoII and cost, while the jammer aims to *maximize* AoII while minimizing its resource expenditure.

A Nash Equilibrium (NE) is a strategy pair $(p^\star, r^\star)$ such that neither player can unilaterally improve its utility, given the other's choice.

Theorem 8.1.1 (*Existence and Uniqueness of the Nash Equilibrium*)

The game $\mathcal{G}_{\mathsf{D}}$ admits a unique Nash Equilibrium $(p^\star, r^\star)$ in pure strategies. □

Proof. Both utility functions in (8.5) are continuous and strictly concave in each player's decision variable, while the corresponding action spaces are convex and compact. By Glicksberg's extension of Nash's theorem, at least one pure-strategy NE exists. The strict concavity of each utility ensures that this equilibrium is unique. □

For continuous strategy spaces, the NE can be characterized by the first-order optimality conditions:

$$\frac{\partial u_{\mathrm{M}}(p, r)}{\partial r} = 0, \qquad \frac{\partial u_{\mathrm{T}}(p, r)}{\partial p} = 0. \tag{8.6}$$

Substituting from Equation (8.5) gives:

$$\frac{\partial \Delta_{\mathrm{D}}}{\partial r} = K, \qquad \frac{\partial \Delta_{\mathrm{D}}}{\partial p} = -C. \tag{8.7}$$

Using the explicit expression of $\Delta_{\mathrm{D}}(p, r)$ from Equation (8.3), we obtain the following coupled equilibrium conditions:

Rearranging the terms, it results that the NE can be obtained by solving the following system of equations:

$$\begin{cases} p = & \dfrac{2K \cdot r}{C - K} \\ r = & \dfrac{p}{d}\left(\dfrac{p\sqrt{1-2Kp^2}}{1-2Kp^2} - (p+d)\right) \end{cases}, \tag{8.8}$$

To ensure physical consistency, the equilibrium values $p^\star$ and $r^\star$ must satisfy $p^\star > 0$ and $r^\star > 0$; any negative values are truncated to zero, as negative rates lack physical meaning.

A special case arises when $r^\star = 0$, indicating that the jammer remains inactive. The interaction then collapses to a single-agent optimization for the legitimate transmitter, equivalent to the nominal case analyzed in the FDI framework (see Chapter 7). The transmitter's optimal rate in this scenario, denoted p_{N}, coincides with the no-attack solution.

Combining the active and inactive attack regimes, the Nash Equilibrium can be expressed as:

$$p^\star = \begin{cases} \dfrac{2K \cdot r^\star}{C - K}, & \text{if } K < K_{\mathrm{D}}^\star, \\ p_{\mathrm{N}}, & \text{otherwise,} \end{cases} \tag{8.9}$$

$$r^\star = \begin{cases} \dfrac{p^\star}{d}\left(\dfrac{p^\star\sqrt{1-2K(p^\star)^2}}{1-2K(p^\star)^2} - (p^2+d)\right), & \text{if } K < K_{\mathrm{D}}^\star, \\ 0, & \text{otherwise.} \end{cases} \tag{8.10}$$

The attacker's activation condition is then given by:

$$\frac{1}{2}\left(\frac{1}{(p^\star)^2} - \frac{1}{(p^\star+d)^2}\right) < K < \frac{1}{2(p^\star)^2}, \tag{8.11}$$

which defines the feasible interval of jamming costs for which the attacker finds it profitable to maintain a nonzero activity rate ($r^\star > 0$).

Theorem 8.1.2 (*Existence and Uniqueness of the Critical Cost* $K_{\mathrm{D}}^\star$)

For all $C > 0$ and $d > 0$, there exists a unique threshold $K_{\mathrm{D}}^\star$ such that:

$$\begin{cases} p^\star > 0,\ r^\star > 0, & \text{if } K < K_{\mathrm{D}}^\star, \\ p^\star = p_{\mathrm{N}},\ r^\star = 0, & \text{if } K \geq K_{\mathrm{D}}^\star. \end{cases}$$

□

Proof. For all feasible $C, K, d > 0$, the following equality holds:

$$\frac{(C-K)p}{2K} = \frac{p}{d}\left(\frac{p\sqrt{1-2Kp^2}}{1-2Kp^2} - (p+d)\right).$$

This algebraic relation admits a unique positive solution $p^\star$, representing the equilibrium transmission rate. The corresponding jamming rate is $r^\star = \frac{C-K}{2K} p^\star$, which is positive only if $C > K$. Therefore, the jammer remains active if and only if its cost is lower than the transmitter's, yielding:

$$K_{\mathrm{D}}^\star = C.$$

□

Theorem 8.1.2 reveals that a DoS attack is economically viable only when the attacker's cost K is lower than the transmitter's cost C. When $K \geq C$, maintaining an active jamming strategy becomes unprofitable, and the optimal behavior for the attacker is to remain inactive ($r^\star = 0$). This finding underscores a fundamental resilience property of the system: by increasing the attacker's operational cost—through enhanced communication security, redundancy, or frequency hopping—the equilibrium naturally shifts toward the defender's advantage, neutralizing the attacker's incentive to act.

8.1.1 Numerical Results

We now present numerical simulations to illustrate the strategic interaction between the legitimate transmitter and the malicious jammer under a DoS attack. From Theorem 8.1.2 we know that the jammer is active at equilibrium only when its cost is lower than the transmitter's, i.e., $K < C$. To analyze this regime systematically *and* to study how the relative costs shape equilibrium strategies, we parameterize the jammer's cost as

$$K \;=\; \alpha\, C,$$

with $\alpha > 0$ a dimensionless scaling factor that captures the attacker's *relative cost efficiency*. By construction, $\alpha < 1$ corresponds to the economically viable attack region $K < C$ (jamming cheaper than transmitting), whereas $\alpha > 1$ maps to $K > C$, where the jammer is deterred and remains inactive.

For the simulations presented in this section, we fix $\alpha = 0.3$, corresponding to a relatively cost-efficient attacker, and vary the transmission cost C to investigate how both agents' equilibrium strategies and the resulting AoII evolve under different operational conditions. The drift rate d is set to two representative values, $d = [0.1, 0.5]$, to capture systems with slow and fast intrinsic dynamics, respectively.

Figure 8.3 compares the transmitter's activity rate p as a function of increasing transmission cost C under both nominal (dashed lines) and DoS attack (solid lines) conditions. As depicted, the jammer's intervention generally leads to an increase in the transmitter's rate p compared to normal operation. This is because the transmitter tries to compensate for the blocked packets. As the drift rate d increases, signifying faster information obsolescence, the equilibrium transmission rate p also increases for a given cost, shifting the curves to the right. Moreover, the difference between p with and without attack diminishes as C escalates, indicating that beyond a certain point, the high transmission costs outweigh the benefits of increasing the rate further to mitigate the attack.

Figure 8.4 illustrates the ratio of the malicious activity rate r to the transmitter rate p as a function of increasing transmission cost C. When C is low, the attacker's activity r is relatively low compared to p, resulting in a small r/p ratio. However, as C increases, the attacker finds it more advantageous to intensify its jamming efforts, leading to a rise in this ratio. This trend suggests that as the cost for the legitimate transmitter rises, the attacker becomes proportionally more active, implying a shift in the strategic landscape where jamming becomes more impactful relative to the transmitter's efforts.

Figure 8.5 shows the impact of the DoS attack on the average AoII as a function of the transmission cost C. The solid lines represent AoII under attack, while dashed lines show AoII under normal operation. As expected, AoII generally increases with rising transmission costs, as the transmitter is incentivized to transmit less frequently.

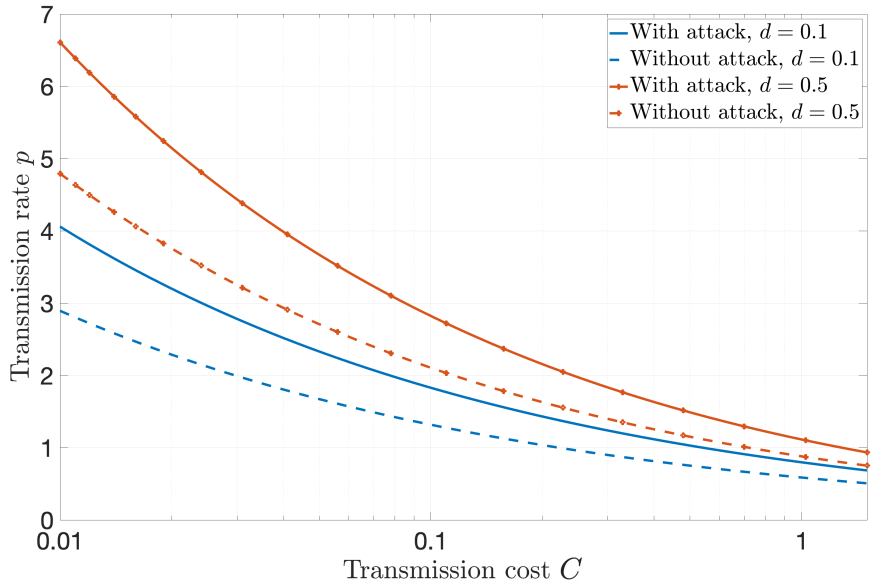


Figure 8.3: Comparison of transmitter activity rate p as a function of increasing transmission price C with fixed value of $\alpha = 0.3$, when $d = [0.1, 0.5]$.

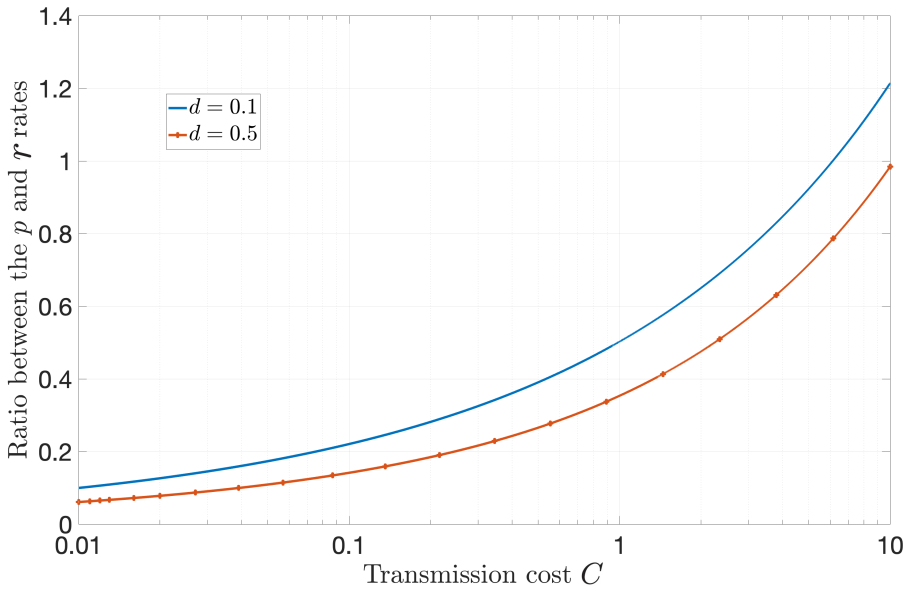


Figure 8.4: Ratio between the attacker rate r and the transmitter rate p as a function of increasing transmission price C with fixed value of $\alpha = 0.3$ when $d = [0.1, 0.5]$.

Furthermore, higher drift rates d necessitate more frequent transmissions to maintain freshness, leading to higher AoII values across the board. Importantly, despite the transmitter's efforts to mitigate the attack by increasing p, AoII consistently degrades under attack conditions compared to the nominal case, highlighting the adversary's detrimental impact.

Finally, Figure 8.6 illustrates the transmitter's equilibrium activity rate p as a function of the jammer's activity cost $K = \alpha C$, with C fixed at 0.5. As predicted by the

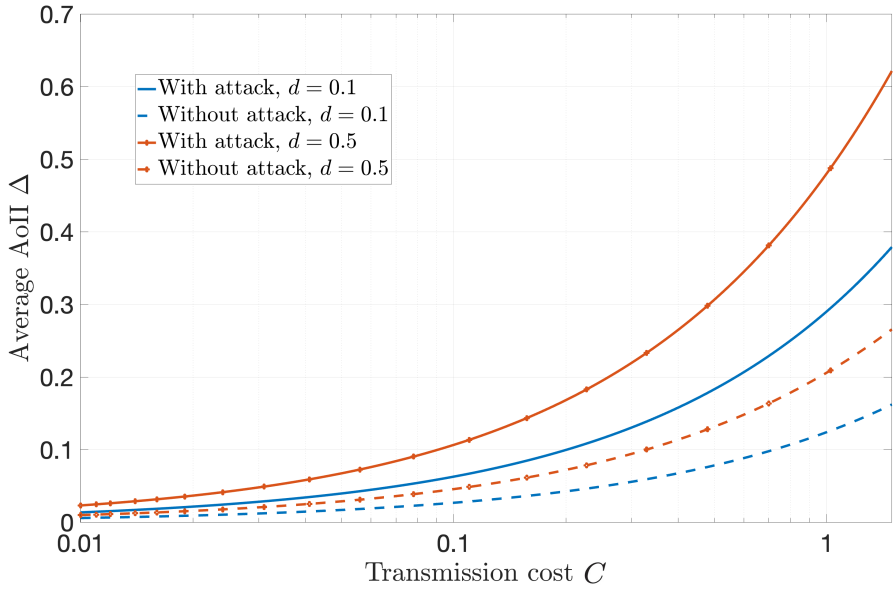


Figure 8.5: Comparison of AoII as a function of increasing transmission price C with fixed $\alpha = 0.3$ when $d = [0.1, 0.5]$.

analytical results, when the attacker's cost increases, sustaining a jamming activity becomes progressively less convenient. In particular, when $K = C = 0.5$ (i.e., $\alpha = 1$), the system reaches the critical threshold identified in Theorem 8.1.2: beyond this point, the attacker's cost equals the transmitter's, and jamming ceases to be economically viable. Consequently, the malicious agent remains inactive ($r^\star = 0$), and the transmitter's optimal rate p smoothly converges toward its nominal no-attack value. This confirms that increasing the attacker's operational cost effectively deters malicious activity and restores normal transmission behavior for the legitimate node.

8.1.2 Discussion

The analysis of Denial-of-Service attacks provides a complementary perspective to that of False Data Injection, highlighting how the *absence* of valid updates can be as detrimental as their *corruption*. Here, the attacker's goal is not to falsify information but to delay its renewal, thereby amplifying the system's exposure to stale data. By obstructing the communication channel, the jammer effectively throttles the corrective transition in the information dynamics, causing the receiver to remain in an outdated state for longer intervals. This mechanism underscores that even purely temporal interference can lead to significant informational degradation, as captured by the increase in the Age of Incorrect Information.

The equilibrium analysis clarifies the conditions under which jamming remains a viable strategy. The existence of a unique Nash Equilibrium ensures that both the transmitter and the jammer adopt predictable, steady-state behaviors. The critical cost threshold $K_{\mathrm{D}}^\star = C$ demarcates two qualitatively distinct regimes: for $K < C$, the attacker operates actively, as the cost of sustaining interference is lower than that of legitimate transmission; for $K \geq C$, the attack becomes economically unsustainable, and the jammer rationally withdraws. This property reveals an intrinsic form of *economic resilience*, whereby system security can be strengthened not only by technical means but also by manipulating the relative cost structure of the interaction—for example, through energy hardening, spectrum randomization, or channel diversity.

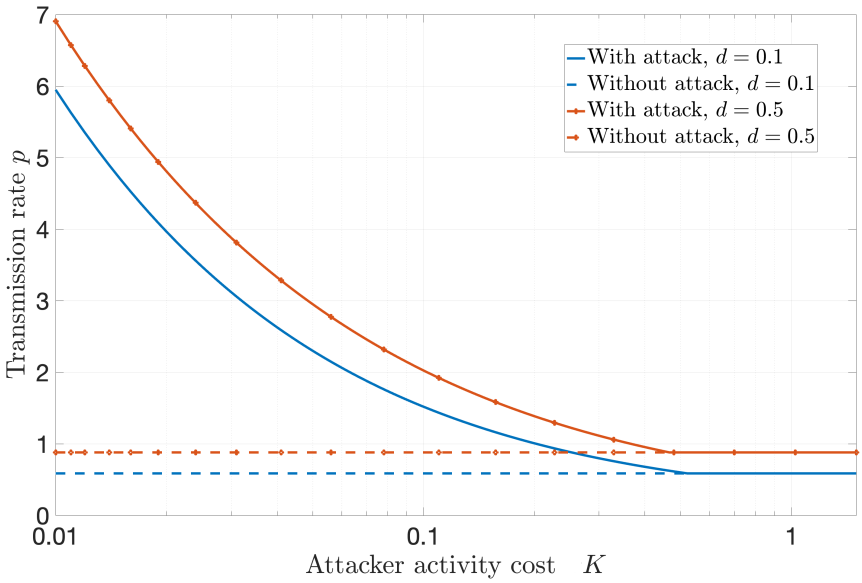


Figure 8.6: Comparison of transmitter activity rate p as a function of increasing attacker cost $K = \alpha C$ with fixed value of $C = 0.5$, when $d = [0.1, 0.5]$.

The numerical results further substantiate these theoretical findings. When transmission costs are low, the legitimate node compensates for packet loss by increasing its update rate, thereby mitigating the AoII degradation induced by jamming. As costs rise, however, this compensatory behavior becomes prohibitively expensive, allowing the jammer to dominate the interaction and drive the AoII upward. The sensitivity analysis with respect to the drift rate d shows that fast-evolving processes are intrinsically more demanding to protect: higher drift accelerates information obsolescence, forcing the transmitter to operate at higher rates to preserve freshness. Conversely, slow dynamics offer greater tolerance to jamming, as information remains relevant for longer even under partial packet loss.

A key insight emerging from the model is that DoS efficiency depends jointly on three dimensions—*dynamics*, *cost*, and *activity*. High transmission cost or high process volatility magnify the attack's impact, whereas increasing the adversary's operational cost or introducing communication redundancy shifts the equilibrium in favor of the defender. This observation emphasizes the importance of designing systems where the marginal cost of legitimate communication is minimized and that of sustained interference is maximized, a principle that can be realized through lightweight redundancy, opportunistic retransmission, or power-adaptive strategies.

Overall, the DoS analysis illustrates how the temporal layer of communication can be weaponized to compromise information freshness without direct data manipulation. The proposed framework quantifies this effect in analytical terms and reveals the structural levers—particularly cost asymmetry and drift dynamics—that govern resilience. In essence, DoS attacks exploit inefficiencies in the time domain; robust system design should therefore focus on restoring temporal agility through adaptive update policies, channel diversity, and resource-aware defense mechanisms. These insights complete the characterization of adversarial impacts on information freshness, setting the stage for the multi-sensor and distributed extensions developed in the following chapter.

Chapter 9

Data Substitution Attacks

This chapter examines the class of *Data Substitution* (DS) attacks, a particularly insidious form of cyber interference in which the adversary not only blocks legitimate transmissions but also replaces them with falsified data. While Denial-of-Service attacks disrupt timeliness and False Data Injection attacks distort semantic correctness, Data Substitution combines both effects: it simultaneously prevents genuine updates from reaching the receiver and injects misleading information that appears valid. This dual mechanism leads to a compounded degradation of information freshness and accuracy, making DS a critical threat to the integrity of networked sensing and control systems.

In the stochastic framework introduced previously, we adapt the two-state Markov model of information correctness to capture this hybrid disruption. Here, the transition from the correct to the incorrect state is accelerated by falsified updates, while the reverse transition is inhibited due to the interception of authentic packets. As a result, both the rate at which errors occur and the duration for which they persist are amplified, producing a stronger deterioration in the *Age of Incorrect Information* compared to either FDI or DoS in isolation. This representation allows for a unified quantification of how substitution affects the temporal and semantic dimensions of informational reliability.

The interaction between the legitimate transmitter and the attacker is then formulated as a static, non-cooperative game of complete information. The transmitter aims to minimize the AoII and its communication expenditure by choosing an optimal transmission rate, while the attacker seeks to maximize informational disruption by tuning its substitution activity under a cost constraint. The equilibrium analysis yields closed-form expressions for both agents' strategies and reveals the existence of a critical cost threshold above which substitution becomes economically infeasible. This threshold encapsulates the system's intrinsic resilience and provides explicit design guidance for mitigating hybrid attacks.

Organization of the chapter. The remainder of this chapter is organized as follows:

- **Section 9.1** introduces the system model for Data Substitution and derives the corresponding AoII expression using a continuous-time Markov representation.
- **Section 9.2** formulates the strategic interaction between the transmitter and the attacker as a static non-cooperative game, defining the utility functions and strategic variables.
- **Section 9.3** derives analytical expressions for the Nash Equilibrium, identifying the optimal transmission and substitution rates as functions of system and cost parameters.
- **Section 9.4** discusses the main findings, emphasizing the compounded nature of Data Substitution attacks and their implications for resilient communication architectures.

9.1 System Modeling under Data Substitution

At a specific time, an adversary gains access to the communication channel, and aims to *replace* legitimate transmissions with falsified ones. In this type of *Data Substitution* attack, the adversary intercepts each packet sent by the legitimate transmitter and substitutes a fraction of them with corrupted data before they reach the receiver. This makes the attack particularly insidious: unlike DoS, communication appears to proceed normally, yet the information content becomes progressively unreliable.

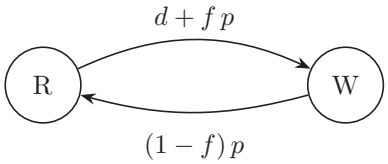


Figure 9.1: Continuous-time Markov chain for information correctness under Data Substitution attack. The transition R $\to$ W occurs due to both natural drift (d) and falsified updates (fp), while W $\to$ R occurs only when a legitimate update is successfully received.

We denote by $f \in (0, 1)$ the *substitution probability*, representing the fraction of packets intercepted and replaced by the attacker. If the legitimate transmitter sends updates at rate p, the rate of successful (authentic) transmissions that actually reach the receiver is reduced to $(1 - f)p$. At the same time, the falsified packets generated by the attacker effectively act as false data injections occurring at rate fp. We again assume negligible propagation delay between transmitter and receiver, so that timing considerations can be made equivalently at either end of the communication link.

In this setup, transitions between these states are governed by:

- R $\to$ W at rate $d + fp$, due to either natural information drift or substitution of legitimate packets;
- W $\to$ R at rate $(1 - f)p$, corresponding to successful legitimate updates.

Compared to the DoS case, the Data Substitution attack not only reduces the correction rate but also actively increases the rate of incorrect transitions, resulting in a more severe degradation of information accuracy at the receiver.

The relationship between the attacker's operational activity and the substitution probability is again fundamental. Let $r \in (0, \infty)$ denote the *attacker's activity rate*, quantifying how intensively the adversary interferes with the communication channel. A higher r implies more frequent or more effective packet replacements. The substitution probability f is related to r and p through:

$$f = \frac{r}{p + r}\,, \tag{9.1}$$

which satisfies $f \in [0, 1)$. As $r \to 0$, the substitution probability tends to zero, representing no interference; conversely, as $q \to \infty$, the attacker replaces nearly all transmissions ($f \to 1$). Similarly to the DoS case, the attacker's *activity rate* can be expressed in terms of the substitution parameters as:

$$r = \gamma_{\mathsf{S}}(q, f, p) = p\,\frac{f}{1 - f}. \tag{9.2}$$

This formulation links the adversary's operational effort r to the substitution probability f, highlighting that maintaining a high fraction of replaced packets requires increasingly large activity rates as $f \to 1$.

The attacker's internal behavior can be modeled by a simple two-state Markov process, alternating between *Active* and *Inactive* phases, as shown in Figure 9.2. During the Active phase, the attacker intercepts and substitutes legitimate packets. The transition from Inactive to Active occurs at rate r, representing the adversary's persistence or aggressiveness. This abstraction captures the stochastic availability of the attacker and its cumulative impact on the communication channel.

Following the same reasoning adopted in Section 6.2, we use the *Age of Incorrect Information* as the principal performance metric to quantify the degradation of information

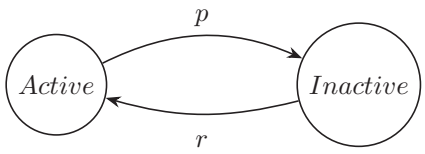


Figure 9.2: Continuous-time Markov chain describing the attacker's operational states. The rate r represents the frequency or intensity of activation for the substitution mechanism.

accuracy caused by substitution attacks. Given the transition rates in Figure 9.1, the expected AoII can be expressed as:

$$\Delta_{\mathsf{S}} = \frac{1/(2((1-f)p)^2)}{\frac{1}{(1-f)p} + \frac{1}{d+fp}} = \frac{d+fp}{2(1-f)p(d+p(1-f))}. \tag{9.3}$$

Substituting f from Equation (9.1), the AoII becomes an explicit function of p, r, and d:

$$\Delta_{\mathsf{S}}(p,r) = \frac{(dp + r(d+p))(p+r)^2}{2p^2[p^2 + d(p+r)]}. \tag{9.4}$$

This expression quantifies how the substitution activity r and the transmitter's response p jointly influence the long-term information accuracy at the receiver. Unlike DoS attacks, where the adversary merely blocks updates, Data Substitution both obstructs and corrupts the communication flow—combining the effects of disruption and deception into a single, unified threat.

9.2 Game-Theoretic Formulation

We model the strategic interaction between the legitimate transmitter (T) and the malicious attacker (M) as a two-player non-cooperative static game of complete information:

$$\mathcal{G}_{\mathrm{S}} = (\mathcal{P}, \mathcal{A}, \mathcal{U}),$$

where the set of players is $\mathcal{P} = \{\mathrm{T}, \mathrm{M}\}$, the action spaces are $\mathcal{A}_{\mathrm{T}} = \{p \in [0, \infty)\}$ for the transmitter and $\mathcal{A}_{\mathrm{M}} = \{f \in [0, 1)\}$ for the attacker, and $\mathcal{U} = \{u_{\mathrm{T}}, u_{\mathrm{M}}\}$ denotes the set of utility functions.

The transmitter seeks to maintain information correctness while minimizing its communication cost. Accordingly, its utility function is defined as:

$$u_{\mathrm{T}}(p,r) = -\Delta_{\mathsf{S}}(p,r) - C\,p, \tag{9.5}$$

where $C > 0$ represents the unit cost of transmission and $\Delta_{\mathsf{S}}(p,f)$ is the average Age of Incorrect Information as derived in Equation (9.4). The transmitter therefore aims to *minimize* the joint impact of the AoII and its transmission effort.

Conversely, the attacker's goal is to maximize the overall degradation of information accuracy while minimizing its own resource expenditure. Its utility function is thus expressed as:

$$u_{\mathrm{M}}(p,r) = \Delta_{\mathsf{S}}(p,r) - K\,r, \qquad r = \frac{pf}{1-f}, \tag{9.6}$$

where $K > 0$ denotes the unit cost of attack activity and r is the equivalent activity rate corresponding to a substitution probability f, as defined in Equation (9.1).

The game proceeds as follows: the transmitter and the attacker choose their strategies p and f simultaneously, each optimizing its own utility while anticipating the opponent's rational behavior.

Theorem 9.2.1 (*Existence of Nash Equilibrium*)

For all $C > 0$, $K > 0$, and $d > 0$, the Data Substitution game $\mathcal{G}_{\mathrm{S}}$ admits at least one Nash Equilibrium in pure strategies. □

Proof. The strategy sets $p \in [0, \infty)$ and $r \in [0, \infty)$ are convex and compact. The utilities $u_{\mathrm{T}}(p, r)$ and $u_{\mathrm{M}}(p, r)$ are continuously differentiable and strictly concave in their respective variables. By Glicksberg's generalization of Nash's theorem, the game admits at least one pure-strategy Nash Equilibrium. □

Corollary 9.2.1 (*Uniqueness of the Nash Equilibrium*)

The Nash Equilibrium of $\mathcal{G}_{\mathrm{S}}$ is unique. □

Proof. Strict concavity of u_{T} in p and u_{M} in r implies that both best-response functions are single-valued and strictly monotonic. Therefore, they intersect at a single point, which uniquely defines the equilibrium pair $(p^\star, r^\star)$. □

9.3 Analytical Derivation of the Nash Equilibrium

From the definitions of the utilities in Equations (9.5)–(9.6), and using the expression for the Age of Incorrect Information in Equation (9.4) together with the relationship between f and r given by Equation (9.1), we can rewrite:

$$u_{\mathrm{T}}(p, r) = -\Delta_{\mathsf{S}}(p, r) - C\, p, \qquad u_{\mathrm{M}}(p, r) = \Delta_{\mathsf{S}}(p, r) - Kr.$$

The Nash Equilibrium $(p^\star, r^\star)$ is obtained by simultaneously solving the first-order optimality conditions for both players:

$$\frac{\partial u_{\mathrm{T}}}{\partial p} = -\frac{\partial \Delta_{\mathsf{S}}}{\partial p} - C = 0, \tag{9.7}$$

$$\frac{\partial u_{\mathrm{M}}}{\partial r} = \frac{\partial \Delta_{\mathsf{S}}}{\partial r} - K = 0. \tag{9.8}$$

Equation (9.7) captures the transmitter's optimal balance between information freshness and transmission cost, while Equation (9.8) characterizes the attacker's trade-off between the gain in informational disruption and the corresponding cost of substitution activity.

By substituting $\Delta_{\mathsf{S}}(p, f)$ from Equation (9.4) into (9.7)–(9.8) and performing algebraic simplifications, we obtain closed-form expressions for the equilibrium strategies:

$$p^\star = \frac{1}{\sqrt{2K}}, \tag{9.9}$$

and the corresponding equilibrium attacker activity rate:

$$q^\star = \frac{C\big(1 + d^2 + 2d\sqrt{2K}\big) - 2d(\sqrt{2} + d\sqrt{K})K^{3/2}}{\sqrt{2K}(1 + 2dK)^2}. \tag{9.10}$$

These expressions describe the steady-state strategies of both agents. The equilibrium transmission rate $p^\star$ depends only on the attack cost K, whereas $q^\star$ also depends on the transmitter cost C and the system drift d. As expected, larger values of K (more expensive attacks) result in lower equilibrium activity levels for the attacker and, consequently, a lower effective substitution rate.

A special case occurs when $q^\star = 0$, meaning that the attacker remains inactive. In this case, the problem reduces to a single-agent optimization for the legitimate transmitter, equivalent to the nominal case analyzed in the FDI framework (see Chapter 7). The transmitter's optimal update rate $p^\star$ is then given by the no-attack solution, denoted p_{N}.

Combining the active and inactive attack regimes, the Nash Equilibrium can be expressed as:

$$p^\star = \begin{cases} \dfrac{1}{\sqrt{2K}}, & \text{if } K < K_{\mathrm{S}}^\star, \\ p_{\mathrm{N}}, & \text{otherwise,} \end{cases} \tag{9.11}$$

$$r^\star = \begin{cases} \dfrac{C\big(1+d^2+2d\sqrt{2K}\big) - 2d(\sqrt{2}+d\sqrt{K})K^{3/2}}{\sqrt{2K}(1+2dK)^2} K < K_{\mathrm{S}}^\star, \\ 0, \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{otherwise.} \end{cases} \tag{9.12}$$

The attacker's activation condition can be written as:

$$\frac{1}{2}\left(\frac{1}{p^2} - \frac{1}{(p+d)^2}\right) < K < \frac{1}{2p^2}. \tag{9.13}$$

This inequality defines the feasible range of jamming cost values for which the attacker finds it profitable to remain active ($r^\star > 0$).

The attacker remains active at equilibrium only when the equilibrium activity rate $q^\star$ in Equation (9.10) is positive. The condition $q^\star > 0$ yields the following inequality:

$$C > K\left(1 + \sqrt{\frac{d^2K}{2}}\right), \tag{9.14}$$

which establishes the range of cost parameters for which the attacker finds it profitable to engage in data substitution.

Theorem 9.3.1 (*Existence of the Substitution Activity Threshold*)

There exists a unique $K_{\mathrm{S}}^\star > 0$ satisfying

$$C = K_{\mathrm{S}}^\star\left(1 + \sqrt{\frac{d^2K_{\mathrm{S}}^\star}{2}}\right), \tag{9.15}$$

such that

$$\begin{cases} K < K_{\mathrm{S}}^\star & \Rightarrow\ r^\star > 0, \\ K \geq K_{\mathrm{S}}^\star & \Rightarrow\ r^\star = 0. \end{cases}$$

□

Proof. From (9.11), $p_{\mathsf{S}} > 0$ for all feasible values of game parameters. Analyzing (9.12) reveals that $r > 0$ requires

$$\frac{C(1+d^2+2\sqrt{2}d\sqrt{K}) - 2d(\sqrt{2}+d\sqrt{K})K^{3/2}}{(\sqrt{2K}(1+2dK)^2)} > 0.$$

whose denominator is always positive, thus it is sufficient that

$$2Cd - 2K - \sqrt{2}d^2K^{3/2} > 0,$$

$$\text{which is satisfied for} \qquad C > K\left(1 + \sqrt{d^2K/2}\right). \tag{9.16}$$

Now, recall that $C > 0$, and note that $K\left(1 + \sqrt{d^2K/2}\right)$ is an increasing monotone function that is zero-valued when $K = 0$, and tends to $+\infty$ as K goes to $+\infty$. This implies that there is only one K_{S}^* that satisfies

$$C = K_{\mathsf{S}}^*\left(1 + \sqrt{d^2K_{\mathsf{S}}^*/2}\right),$$

$$\text{that is} \qquad K_{\mathsf{S}}^{*} = \frac{2}{3d^2} - \frac{12Cd^2 - 4}{3d^2 u} + \frac{u}{3d^2} \tag{9.17}$$

where $u = (8{+}9Cd^2(3Cd^2{-}4) + 3\sqrt{3}C^{3/2}d^3(27Cd^2 - 8)^{1/6}$.
Thus, for $K < K_{\mathsf{S}}^{*}$, (9.16) holds, implying $r_{\mathsf{S}} > 0$. □

The threshold $K_{\mathsf{S}}^{\star}$ thus provides a quantitative benchmark for resilience: if the attacker's operational cost exceeds this critical value, the equilibrium outcome collapses to the nominal (attack-free) case, and the system maintains correct operation despite potential substitution attempts.

9.4 Discussion

Among the three adversarial models analyzed in this work, the *Data Substitution* (DS) attack emerges as the most disruptive form of interference, as it jointly compromises both the temporal and semantic dimensions of information integrity. Unlike DoS attacks, which merely delay the delivery of updates, or FDI attacks, which corrupt data content while preserving timing, DS simultaneously accelerates the degradation of correctness and inhibits its restoration. The Markov representation makes this dual impact explicit: the transition to the incorrect state is intensified by the injection of falsified packets, while the recovery transition is throttled by the interception of genuine updates. This dual manipulation of the state dynamics yields a nonlinear amplification of the *Age of Incorrect Information* (AoII), producing a degradation that cannot be attributed to either mechanism alone.

The game-theoretic analysis provides a quantitative interpretation of this phenomenon. At equilibrium, the transmitter's optimal update rate depends solely on the attacker's cost K, reflecting that legitimate behavior is primarily constrained by the adversary's economic efficiency rather than by the system dynamics. In contrast, the attacker's optimal activity rate also depends on the transmitter's communication cost C and on the natural drift rate d, capturing how the payoff structure and the process evolution jointly determine the attack's profitability. The existence of a critical threshold $K_{\mathsf{S}}^{\star}$ delineates the boundary between active and inactive attack regimes: when $K < K_{\mathsf{S}}^{\star}$, substitution remains viable and the attacker exerts continuous interference; when $K \geq K_{\mathsf{S}}^{\star}$, the equilibrium collapses to the nominal case, and the system naturally regains stability. This property reveals an intrinsic *self-stabilizing mechanism*, where economic asymmetry can neutralize even complex hybrid threats.

From a design perspective, these findings highlight the importance of cost engineering as a defense principle. Increasing the effective cost of substitution—through mechanisms such as cryptographic authentication, watermarking of sensor data, or multi-path transmission—can push the attacker's operating point beyond the threshold $K_{\mathsf{S}}^{\star}$, rendering active interference economically irrational. Complementarily, reducing the legitimate transmission cost, for instance via adaptive rate control or resource-aware scheduling, improves the defender's capacity to react to substitution without excessive expenditure. In this sense, resilience emerges not from eliminating the threat entirely, but from shaping the cost landscape such that rational adversaries are disincentivized from acting.

Overall, the DS analysis exposes the systemic vulnerability arising from the coupling of temporal and semantic degradation in information exchange. By quantifying this coupling and embedding it within a strategic equilibrium framework, the model provides both theoretical insight and actionable criteria for safeguarding distributed sensing and control architectures against hybrid adversarial behaviors. These results complete the characterization of single-sensor attack dynamics and establish the foundation for the multi-sensor and distributed formulations explored in the following chapter.

Chapter 10

Comparative Analysis

The preceding chapters investigated three distinct classes of adversarial interference affecting the reliability of networked cyber–physical systems: *False Data Injection* (FDI, Chapter 7), *Denial-of-Service* (DoS, Chapter 8), and *Data Substitution* (DS, Chapter 9). Each attack was analyzed in isolation, yielding specific insights into how deception, disruption, or their combination alter the temporal correctness of information. Yet, while those studies established the individual dynamics and equilibria of each adversarial policy, they leave open a broader and practically motivated question: *how do these threats compare when evaluated under the same structural, cost, and dynamical conditions?*

This chapter addresses that question by developing a unified analytical framework that enables a direct, consistent comparison across the three adversarial strategies. Using the continuous-time Markov formulation of information correctness and the *Age of Incorrect Information* as a common performance metric, we quantify how FDI, DoS, and DS differ in both their operational mechanisms and strategic outcomes. To ensure comparability, we focus on the canonical single-sensor configuration under complete information—where both transmitter and attacker possess full knowledge of system and cost parameters—allowing equilibrium strategies to be derived in closed form and interpreted on equal footing. This abstraction captures the essential dynamics of strategic interaction without the confounding effects of uncertainty or network coupling.

Through this perspective, we highlight the fundamental contrasts between stealthy and disruptive adversaries, between semantic and temporal degradation, and between low-cost and resource-intensive interference. By analyzing the resulting Nash equilibria, we establish how each attack's feasibility depends on the relative cost structure and on the intrinsic dynamics of the monitored process. The comparative framework thereby elucidates both the *severity*—the attainable degradation in information freshness—and the *efficiency*—the cost-effectiveness of achieving it—of different adversarial policies. It also provides explicit analytical thresholds that delineate the transition between active and inactive attack regimes, offering a principled basis for cost-aware system hardening.

Organization of the chapter. The remainder of this chapter is organized as follows:

- **Section 10.1** contrasts the structural mechanisms underlying FDI, DoS, and DS, outlining their operational logic and their distinct modes of affecting information correctness.
- **Section 10.2** compares the corresponding equilibrium properties and cost-based feasibility thresholds, identifying the conditions under which each attack remains rationally viable.
- **Section 10.3** presents numerical simulations illustrating the comparative evolution of AoII and equilibrium transmission rates across different cost regimes and system parameters.
- **Section 10.4** synthesizes the analytical and numerical findings, drawing general conclusions on the relative effectiveness and deterrence of adversarial strategies, and outlining implications for resilient CPS design.

10.1 Adversarial Strategies in Perspective

To comprehensively analyze how each type of attack affects the system's performance, we delve into the characteristics of three main threat models individually.

False Data Injection : This insidious attack involves introducing misleading data into the system, crafted to appear entirely legitimate. Its primary goal is to compromise the receiver's decision-making process without causing any visible disruption to communication channels. The attacker's objective here is to maximize the Age of Incorrect Information while simultaneously minimizing the inherent cost associated with injecting these false updates. In response, the defender must meticulously adjust the transmission rate to diligently mitigate prolonged periods of misinformation. This model inherently assumes continuous data visibility by the receiver, making stealthiness a critical and defining advantage for the attacker.

DoS: In a stark contrast to FDI, a Denial of Service attack aims to overtly prevent the successful delivery of legitimate updates. This is typically achieved by blocking or jamming communication transmissions. As a direct consequence, the receiver experiences significant delays or outright missing updates, which in turn inevitably increases the AoII as the true state of the system ages unnoticed. The attacker's effort in this scenario is effectively modeled as a blocking rate, compelling the defender to dramatically increase transmission redundancy in a desperate attempt to maintain information freshness and system integrity.

Data Substitution : Representing a hybrid and arguably the most sophisticated threat, Data Substitution involves the adversary replacing correct data with corrupted versions, critically without necessarily preventing their reception. This model effectively captures both timing and semantic corruption, rendering it the most complex and potentially damaging adversarial policy. The defender, unfortunately, cannot readily distinguish between truthful and altered updates, and is therefore forced to aggressively increase the transmission rate to counter the profound uncertainty and induced misinformation.

10.2 Equilibrium Properties and Attack Feasibility

For each distinct adversarial strategy, we have derived a pure-strategy Nash Equilibrium, characterized by the optimal transmission rate p^* chosen by the legitimate transmitter and the corresponding optimal attacker effort r^*. A consistent structural feature across all models is the existence of a critical threshold K^*, such that the attacker remains active if, and only if, its cost parameter K lies below this threshold.

It is important to note that the variable r assumes a slightly different operational meaning depending on the specific attack type. In the case of *False Data Injection* , the attacker's effort directly coincides with the injection rate, i.e., $r = q$, since every adversarial action corresponds to the generation of falsified packets. Conversely, for both *Denial-of-Service* and *Data Substitution* attacks, r quantifies the overall activity rate required to induce, respectively, the *failure probability* f (in DoS) or the *substitution probability* f (in DS). In these cases, the relation between r and f follows the general form

$$r = p\,\frac{f}{1-f},$$

indicating that sustaining a high level of interference (large f) demands a disproportionately higher operational effort from the attacker. This unified representation allows for a consistent comparison across attack types while preserving the specific physical interpretation of each adversarial mechanism. Mathematically, this relationship is expressed as:

$$K > K^* \quad \Rightarrow \quad r^* = 0.$$

These specific thresholds are not universal but are intricately dependent on the system's natural drift rate d, the defender's transmission cost C, and the resulting average AoII under each attack model. In particular, we derived the thresholds as:

K_{F}^* : for False Data Injection,

K_{D}^* : for Denial of Service,

K_{S}^* : for Data Substitutiob.

An insight gained from this comparative analysis is the differential feasibility of each attack type. FDI, for instance, remarkably remains feasible even in scenarios where the attacker's cost K is greater than the defender's transmission cost C (i.e., $K > C$). This resilience is attributed to its inherently lower operational burden compared to the more disruptive DoS or DS attacks, which often require continuous and resource-intensive engagement. Conversely, DoS is generally viable only when the defender's transmission costs C are substantially higher than the attacker's costs K (i.e., $C > K$), as high transmission costs significantly amplify the impact of successful suppression. DS , being a hybrid and more complex threat, demands a delicate balance: it requires both a relatively low attacker cost and at least moderate transmission costs for the defender to remain truly efficient as a strategic option.

This clear differentiation among attack strategies, based on their feasibility and cost parameters, allows for several key strategic applications. It enables robust *risk evaluation*, by providing a framework to accurately determine whether a system is vulnerable based on the interplay of attacker's cost K and defender's cost C. It facilitates precise *defense tuning*, by offering insights into how to adjust C or other system parameters to strategically push the attacker beyond their critical operational threshold, thereby deterring attacks by design. Ultimately, it supports sophisticated *strategic classification*, allowing for threats to be ranked not only by their potential severity but also by their inherent feasibility under specific resource constraints for both the attacker and the defender.

10.3 Simulation-Based Comparative Assessment

Moving from theoretical insights to tangible outcomes, we now quantify the distinct impact of each attack strategy through comprehensive numerical simulations. These results graphically illustrate the equilibrium behaviors exhibited by the system when subjected to FDI, DoS, and DS policies.

Figures 10.1 and 10.2 display the system's transmission rate at equilibrium as the defender's transmission cost C and the attacker's cost K vary, respectively. When the defender's transmission cost C is low, attackers find it remarkably ineffective to launch sustained assaults, consequently, the defender maintains high-frequency updates, ensuring optimal information freshness. Among all the analyzed strategies, DS consistently provokes the steepest and most aggressive increase in the transmission rate. This behavior is directly attributable to the compounded uncertainty that DS introduces, forcing the defender to overcompensate by transmitting more frequently to restore data integrity. In stark contrast, FDI causes the least disturbance to the transmission rate, yet it persists as a viable and persistent threat across a remarkably wide range of cost parameters, making it a subtle but enduring challenge.

Figures 10.3 and 10.4 report the target metric of average AoII under varying C and K. As anticipated by the theoretical analysis, DS attacks consistently induce the highest average AoII. This significant degradation is due to their dual impact, simultaneously corrupting both the timing and the correctness of the information. FDI generally outperforms DoS in terms of its impact on AoII when the attacker's cost K is lower than the defender's transmission cost C (i.e., $K < C$), demonstrating its efficiency under favorable cost conditions. However, FDI's potency diminishes considerably when the attacker becomes significantly resource-constrained, highlighting its reliance on a certain level of operational viability.

The ratio r/p, representing the adversary's activity rate relative to the legitimate agent's transmission rate, as depicted in Figures 10.5 and 10.6, offers a deeper insight into the cost-effectiveness of each attack. This ratio strikingly reveals that DS demands considerably less effort from the attacker to maintain significant pressure on the system and degrade AoII. This confirms DS 's inherently high efficiency and compelling cost-effectiveness, particularly evident when the transmission cost C is moderate or high, indicating that even a small DS effort can force a large defensive response. DoS, conversely, necessitates a sustained and substantial effort from the attacker to achieve its objectives

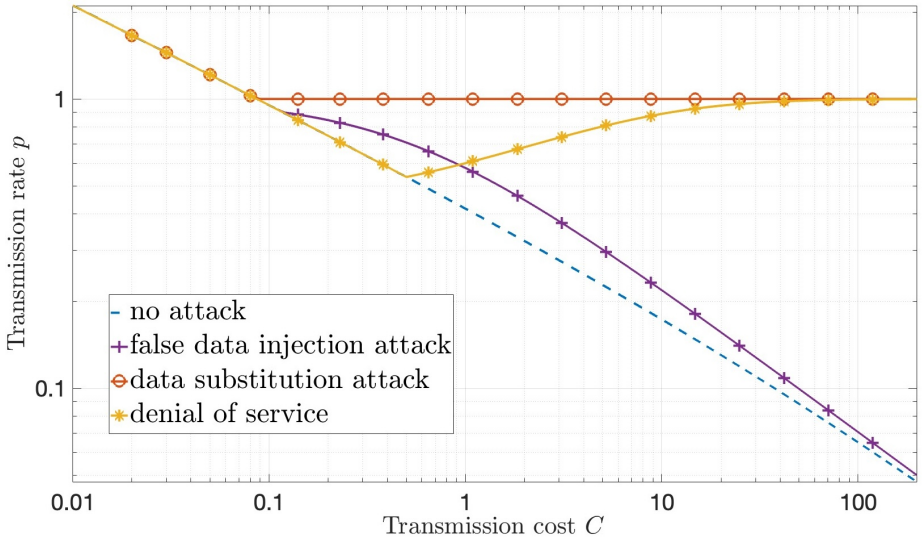


Figure 10.1: Transmission rate at NE, varying the transmitter's cost of action. Drift rate is $d = 0.1$, enemy activity cost is $K = 0.5$.

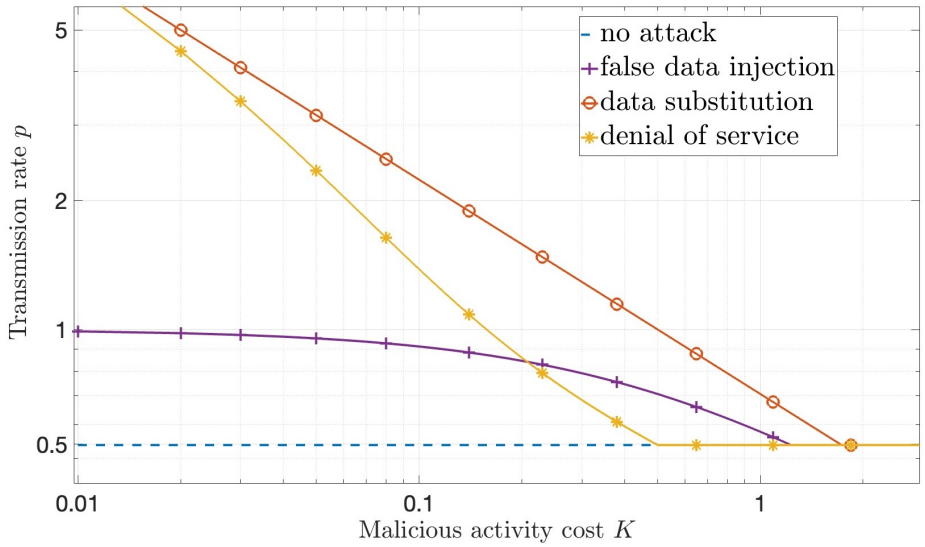


Figure 10.2: Transmission rate at NE, varying the opponent's cost of action. Drift rate is $d = 0.1$, transmission cost is $C = 0.5$.

and is, quite notably, easily deterred by the defender's ability to engage in low-cost transmissions. This quantitative comparison underscores the asymmetric nature of these threats and provides a basis for prioritizing defensive measures.

10.4 Discussion

This comprehensive comparative study, integrating rigorous theoretical modeling with empirical numerical simulations, profoundly highlights the distinctive nature and strategic implications of each adversarial strategy within the context of information freshness in sensor systems.

The FDI attack, while often subtle and operating under conditions of incomplete information for the receiver, demonstrates an efficiency that allows it to remain a persistent threat. Its impact, however, is bounded, suggesting that while it can introduce misleading data, its overall capacity to degrade information freshness might be less severe compared

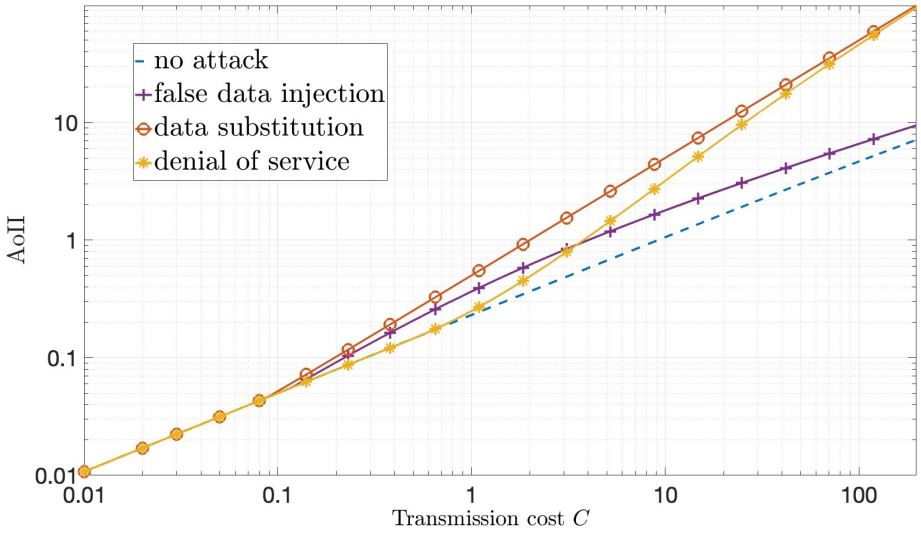


Figure 10.3: Average AoII. Drift rate is $d = 0.1$, enemy activity cost is $K = 0.5$.

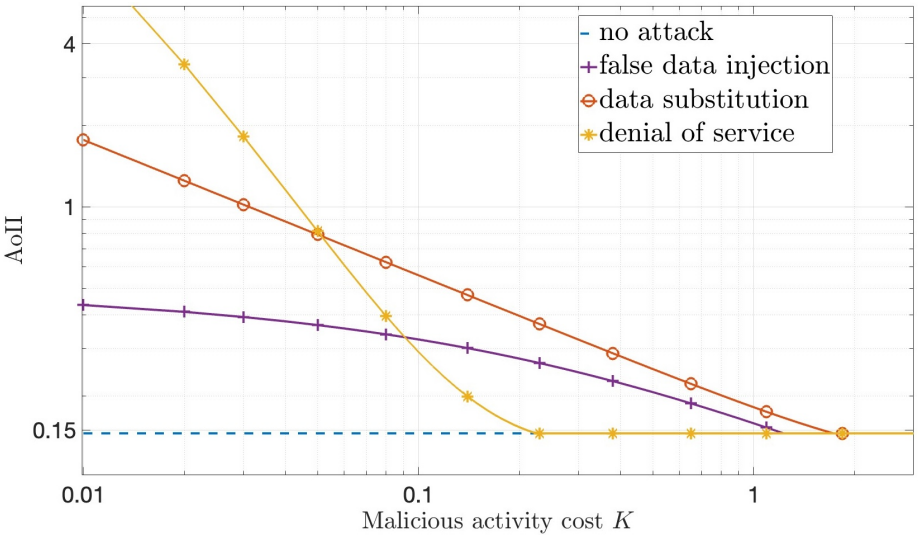


Figure 10.4: Average AoII. Drift rate is $d = 0.1$, transmission cost is $C = 0.5$.

to more disruptive attacks if not sustained. The inherent stealthiness makes it difficult to detect, but its effectiveness is contingent on the defender's inability to distinguish corrupted from legitimate updates.

In contrast, DoS attacks are straightforward in their mechanism: they simply block communication. Our analysis shows that DoS is generally effective only in scenarios where the defender's transmission costs are high. In such 'high-cost regimes', the attacker can achieve significant impact with relatively less effort, making it a viable strategy. However, if the defender can maintain low transmission costs or easily increase redundancy, DoS attacks are readily deterred, becoming a less attractive option for the adversary. This points to direct communication robustness as a strong countermeasure.

The DS strategy emerges as the most sophisticated and damaging of the three. It essentially combines the disruptive potential of DoS (by delaying correct information due to the defender's need to retransmit) with the deceptive nature of FDI (by injecting false information). This dual impact on both timing and correctness renders DS consistently superior in terms of AoII degradation and offers the attacker significant strategic flexibility. The high ratio of attacker effort to defender response, as illustrated by r/p, confirms its remarkable efficiency and cost-effectiveness for the adversary, particularly when the

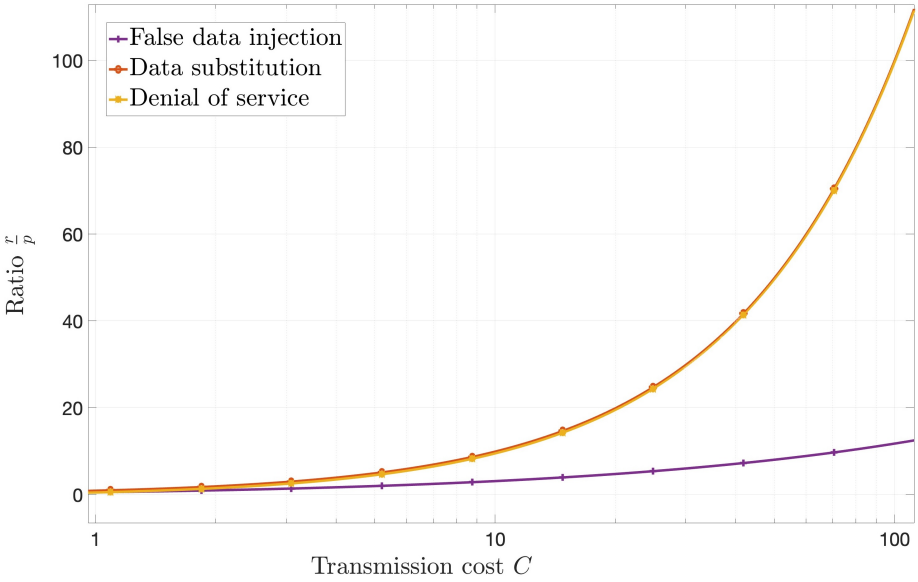


Figure 10.5: Ratio between enemy activity rate r and legitimate agent transmission rate p, varying C. Drift rate $d = 0.1$, enemy cost $K = 0.5$.

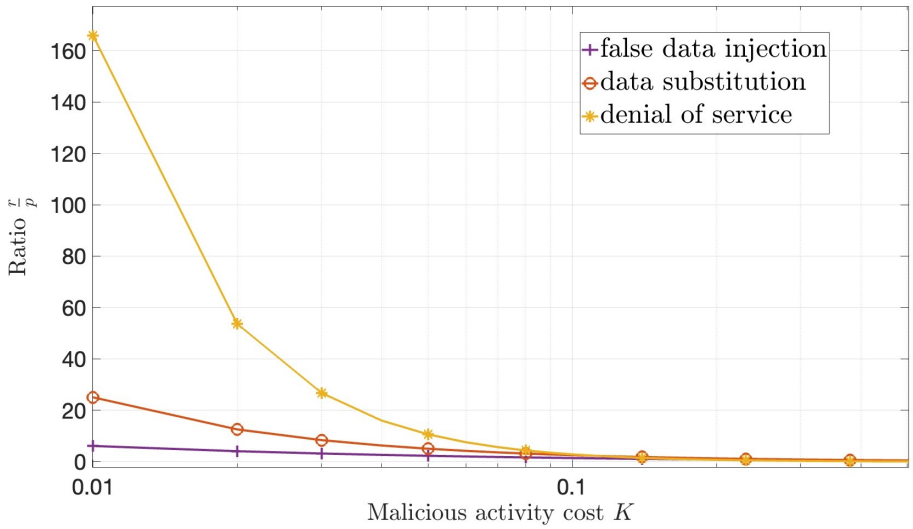


Figure 10.6: Ratio between enemy activity rate r and legitimate agent transmission rate p, varying K. Drift rate $d = 0.1$, transmission cost $C = 0.5$.

defender faces moderate to high transmission costs. This makes DS the most challenging threat to mitigate, as it forces the defender to aggressively overcompensate, often at high cost.

Crucially, a unifying principle across all analyzed attack models is that their feasibility and effectiveness are intrinsically governed by structural thresholds based on the cost parameters (K^* for the attacker's cost relative to the defender's C and d). This fundamental insight is profoundly valuable for system designers. It empowers them to develop proactive, cost-based mitigation strategies that can, by design, render certain attacks economically infeasible for the adversary. For instance, raising the cost of injection for specific sensors (e.g., through enhanced cryptographic authentication or more robust physical security) can push the adversary beyond their K^* threshold, effectively eliminating that attack vector. Conversely, lowering transmission costs for the defender can make DoS attacks unsustainable for the adversary.

Moreover, the equilibrium analysis, which meticulously details how adversaries and defenders adapt to each other's strategies under varying conditions, provides truly actionable insights. It allows for predictive system design, where engineers can anticipate an adversary's likely response to changes in system parameters or defensive investments. This enables the design of resilient systems that can maintain acceptable levels of information freshness even in the presence of intelligent, resource-constrained adversaries. Ultimately, understanding these strategic interactions and the underlying cost-benefit dynamics is paramount for building secure and reliable sensor networks in an increasingly hostile operational environment.

References — Information Freshness in Adversarial Multi-Agent Systems

[52] Banerjee, S. and Ulukus, S., "The freshness game: Timely communications in the presence of an adversary," *IEEE/ACM Transactions on Networking*, vol. 32, no. 5, pp. 4067–4084, 2024. DOI: 10.1109/TNET.2024.3410228.

[54] Yates, R. D., Sun, Y., Brown, D. R., Kaul, S. K., Modiano, E., and Ulukus, S., "Age of information: An introduction and survey," *IEEE Journal on Selected Areas in Communications*, vol. 39, no. 5, pp. 1183–1210, May 2021, Publisher: Institute of Electrical and Electronics Engineers (IEEE). DOI: 10.1109/jsac.2021.3065072. Accessed: Jul. 22, 2025.

[238] Liang, S. et al., "UAV-enabled secure data collection and energy transfer in IoT via diffusion model-enhanced deep reinforcement learning," *IEEE Internet of Things Journal*, 2024. DOI: 10.1109/JIOT.2024.3521496.

[239] Chu, K., Cheng, S., and Zhu, L., "A robust routing strategy based on deep reinforcement learning for mega satellite constellations," *Electronics Letters*, vol. 59, no. 11, 2023. DOI: 10.1049/ell2.12820.

[240] Wang, Q., Chen, H., Mohapatra, P., and Pappas, N., "Secure status updates under eavesdropping: Age of information-based secrecy metrics," presented at the IEEE INFOCOM 2024 - IEEE Conference on Computer Communications Workshops, INFOCOM WKSHPS 2024, 2024. DOI: 10.1109/INFOCOMWKSHPS61880.2024.10620867.

[241] Zhang, M., Arafa, A., Wei, E., and Berry, R., "Optimal and Quantized Mechanism Design for Fresh Data Acquisition," *IEEE Journal on Selected Areas in Communications*, vol. 39, no. 5, pp. 1226–1239, 2021.

[242] Tripathi, V. and Modiano, E., "An online learning approach to optimizing time-varying costs of AoI," presented at the Proceedings of the International Symposium on Mobile Ad Hoc Networking and Computing (MobiHoc), 2021, pp. 241–250. DOI: 10.1145/3466772.3467053.

[243] Zhuang, Z., Wang, J., Qi, Q., Liao, J., and Han, Z., "Adaptive and robust routing with lyapunov-based deep RL in MEC networks enabled by blockchains," *IEEE Internet of Things Journal*, vol. 8, no. 4, pp. 2208–2225, 2021. DOI: 10.1109/JIOT.2020.3034601.

[244] Li, J., Li, J., Lv, L., Ju, Z., and Gong, F., "Multiobjective optimization of AoI and rate for secure dual-functional RIS-assisted networks: A cross-layer design," *IEEE Internet of Things Journal*, vol. 12, no. 8, pp. 10 349–10 363, 2025. DOI: 10.1109/JIOT.2024.3510560.

[245] Bonagura, V., Panzieri, S., Pascucci, F., and Badia, L., "Strategic interaction over age of incorrect information for false data injection in cyber-physical systems," *IEEE Transactions on Control of Network Systems*, vol. 12, no. 1, pp. 872–881, 2025. DOI: 10.1109/TCNS.2024.3431389.

[246] Javani, A. and Wang, Z., "Load Balancing in Federated Learning," 2024, pp. 3950–3955.

[247] Han, Z. et al., "Smart optimization solution for channel access attack defense under UAV-aided heterogeneous network," *IEEE Internet of Things Journal*, vol. 10, no. 21, pp. 18 890–18 897, 2023. DOI: 10.1109/JIOT.2023.3281942.

[248] Banerjee, S., Kaswan, P., and Ulukus, S., "Age-based cache updating under timestomping," presented at the Proceedings of the International Symposium on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks, WiOpt, 2023, pp. 635–642. DOI: 10.23919/WiOpt58741.2023.10349889.

[249] Cao, Y., Guo, Z., Miao, Q., Sun, R., He, J., and Li, X., "Minimization of age of information for satellite-terrestrial covert communication with a full-duplex receiver," presented at the Proceedings - 2024 International Conference on Networking and Network Applications, NaNA 2024, 2024, pp. 242–246. DOI: 10.1109/NaNA63151.2024.00046.

[250] Buratto, A., Mora, A., Bujari, A., and Badia, L., "Game theoretic analysis of AoI efficiency for participatory and federated data ecosystems," in *Proc. IEEE ICC Wkshps*, 2023, pp. 1301–1306.

[251] Zhou, F., Wei, Y., and Leng, S., "Minimizing Age of Information in Multi-party UAV-Assisted Data Collection with Overlapping Areas," in *2024 IEEE 24th International Conference on Communication Technology (ICCT)*, Oct. 2024, pp. 1277–1282. Accessed: Apr. 30, 2025.

[252] Fang, Z., Wang, J., Jiang, C., Zhang, Q., and Ren, Y., "AoI-Inspired Collaborative Information Collection for AUV-Assisted Internet of Underwater Things," *IEEE Internet of Things Journal*, vol. 8, no. 19, pp. 14 559–14 571, Oct. 2021. Accessed: Apr. 30, 2025.

[253] Cao, Z., "Optimization Design of Multi-UAV Communication Network Based on Reinforcement Learning," *Wireless Communications and Mobile Computing*, vol. 2022, 2022.

[254] Yang, Y., Wei, X., Xu, R., Peng, L., and Liu, L., "Game-based channel access for AoI-oriented data transmission under dynamic attack," *IEEE Internet of Things Journal*, vol. 9, no. 11, pp. 8820–8837, 2022. DOI: 10.1109/JIOT.2021.3116142.

[255] Xu, K., Xu, Y., Wang, X., and Wang, X., "A Hybrid Collaborative Learning for Age of Information Minimization in Massive Access," *IEEE Transactions on Vehicular Technology*, vol. 74, no. 2, pp. 2739–2752, Feb. 2025. Accessed: Apr. 30, 2025.

[256] Wang, X., Zhang, J., Chen, C., He, J., Ma, Y., and Guan, X., "Trust-AoI-aware codesign of scheduling and control for edge-enabled IIoT systems," *IEEE Transactions on Industrial Informatics*, vol. 20, no. 2, pp. 2833–2842, 2024. DOI: 10.1109/TII.2023.3299040.

[257] Sathyavageeswaran, N., Yates, R., Sarwate, A., and Mandayam, N., "Privacy leakage in discrete-time updating systems," presented at the IEEE International Symposium on Information Theory - Proceedings, vol. 2022-June, 2022, pp. 2076–2081. DOI: 10.1109/ISIT50566.2022.9834673.

[258] Al-Habob, A. A., Dobre, O. A., and Poor, H. V., "Age- and Correlation-Aware Information Gathering," *IEEE Wireless Communications Letters*, vol. 11, no. 2, pp. 273–277, Feb. 2022. Accessed: Apr. 30, 2025.

[259] Wang, L., Sun, J., Sun, Y., Zhou, S., and Niu, Z., "Age of Information Guaranteed Scheduling for Asynchronous Status Updates in Collaborative Perception," in *2023 21st International Symposium on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks (WiOpt)*, Aug. 2023, pp. 1–8. Accessed: Apr. 30, 2025.

[260] Liang, J., Chan, T.-T., and Pan, H., "Minimizing Age of Collection for Multiple Access in Wireless Industrial Internet of Things," *IEEE Internet of Things Journal*, vol. 11, no. 2, pp. 2753–2766, Jan. 2024. Accessed: Apr. 30, 2025.

[261] Kim, D., Hwang, G., Jo, O., and Shin, K., "Q-learning Based Medium Access Technology for Minimizing AoI in LoRa Wireless Relay Networks," *IEEE Access*, 2024.

[262] Tang, Z., Sun, Z., Yang, N., and Zhou, X., "Age of Information of Multi-User Mobile-Edge Computing Systems," *IEEE Open Journal of the Communications Society*, vol. 4, pp. 1600–1614, 2023. Accessed: Apr. 30, 2025.

[263] Chiariotti, F., "Age of Information Analysis for a Shared Edge Computing Server," *IEEE Transactions on Communications*, vol. 72, no. 12, pp. 7826–7841, Dec. 2024. Accessed: Apr. 30, 2025.

[264] Abedi, M. R., Mokari, N., Javan, M. R., Saeedi, H., Jorswieck, E. A., and Yanikomeroglu, H., "Safety-Aware Age of Information (S-AoI) for Collision Risk Minimization in Cell-Free mMIMO Platooning Networks," *IEEE Transactions on Network and Service Management*, vol. 21, no. 3, pp. 3035–3053, Jun. 2024. Accessed: Apr. 30, 2025.

[265] Banerjee, S., Ulukus, S., and Ephremides, A., "Age of information games between power constrained schedulers and adversaries," *Journal of Communications and Networks*, vol. 25, no. 5, pp. 631–642, 2023. DOI: 10.23919/JCN.2023.000046.

[266] Banerjee, S., Ulukus, S., and Ephremides, A., "Age of information of a power constrained scheduler in the presence of a power constrained adversary," presented at the IEEE INFOCOM 2023 - Conference on Computer Communications Workshops, INFOCOM WKSHPS 2023, 2023. DOI: 10.1109/INFOCOMWKSHPS57453.2023.10226091.

[267] Bhattacharjee, R. and Sinha, A., "Competitive algorithms for minimizing the maximum age-of-information," *Performance Evaluation Review*, vol. 48, no. 2, pp. 6–8, 2020. DOI: 10.1145/3439602.3439606.

[268] Wang, W., Srivastava, G., Lin, J.-W., Yang, Y., Alazab, M., and Gadekallu, T., "Data freshness optimization under CAA in the UAV-aided MECN: A potential game perspective," presented at the IEEE Transactions on Intelligent Transportation Systems, Issue: 11, vol. 24, 2023, pp. 12 912–12 921. DOI: 10.1109/TITS.2022.3167485.

[269] Jia, X., Chen, G., and Sun, Y., "Physical layer security age of information optimization for UAV-jammer assisted relay cooperation system," presented at the IEEE Vehicular Technology Conference, 2024. DOI: 10.1109/VTC2024-Fall63153.2024.10757446.

[270] Zheng, L., Ren, J., Liu, Y., and Chen, Q., "Analysis and optimization of age of information and age of leaked information for IoT networks," *IEEE Wireless Communications Letters*, vol. 13, no. 12, pp. 3583–3587, 2024. DOI: 10.1109/LWC.2024.3481059.

[271] Wang, D. et al., "Minimizing AoI under covertness constraints in internet of things networks," presented at the 2021 IEEE 3rd International Conference on Frontiers Technology of Information and Computer, ICFTIC 2021, 2021, pp. 694–700. DOI: 10.1109/ICFTIC54370.2021.9647130.

[272] Ma, Y., Liu, K., Liu, Y., and Zhu, L., "Timeliness and secrecy-aware uplink data aggregation for large-scale UAV-IoT networks," *IEEE Internet of Things Journal*, vol. 11, no. 10, pp. 17 341–17 356, 2024. DOI: 10.1109/JIOT.2024.3357123.

[273] Mei, H., Zhang, H., Zhou, X., and Wang, J., "AoI minimization for air-ground integrated sensing and communication networks with jamming attack," *IEEE Transactions on Vehicular Technology*, 2025. DOI: 10.1109/TVT.2025.3558061.

[274] Wang, P., Liu, K., Ma, Y., and Gao, Q., "AoI and energy-aware data collection for IRS-assisted UAV-IoT networks under jamming," *IEEE Internet of Things Journal*, 2024. DOI: 10.1109/JIOT.2024.3519668.

[275] Fei, Z., Wang, Y., Qin, X., Zhao, J., and Wang, X., "Joint Wireless and Computational Resources Allocation for Optimizing the Age of Correlated Information in Fog Computing Networks," *IEEE Transactions on Network Science and Engineering*, vol. 11, no. 1, pp. 326–339, Jan. 2024. Accessed: Apr. 30, 2025.

[276] Zhong, A., Li, Z., Wu, D., Tang, T., and Wang, R., "Stochastic Peak Age of Information Guarantee for Cooperative Sensing in Internet of Everything," *IEEE Internet of Things Journal*, vol. 10, no. 17, pp. 15 186–15 196, Sep. 2023. Accessed: Apr. 30, 2025.

[277] Crosara, L., Laurenti, N., and Badia, L., "Strategic status updates in an eavesdropping game," presented at the 28th European Wireless Conference, EW 2023, 2023, pp. 290–295.

[278] Yang, Y. et al., "AoI optimization for UAV-aided MEC networks under channel access attacks: A game theoretic viewpoint," presented at the IEEE International Conference on Communications, vol. 2022-May, 2022, pp. 1–6. DOI: 10.1109/ICC45855.2022.9838975.

[279] Allipuram, S., Mohapatra, P., Pappas, N., Parmar, S., and Chakrabarti, S., "Performance analysis of a MIMO system with bursty traffic in the presence of energy harvesting jammer," *IEEE Transactions on Green Communications and Networking*, vol. 6, no. 2, pp. 1157–1172, 2022. DOI: 10.1109/TGCN.2021.3117470.

[280] Cassandras, C. G. and Lafortune, S., *Introduction to discrete event systems*. Springer, 2008.

[281] Bonagura, V., Panzieri, S., Pascucci, F., and Badia, L., "A game of age of incorrect information against an adversary injecting false data," in *Proc. IEEE CSR*, 2023.

[282] Tian, M., Dong, Z., and Wang, X., "Analysis of false data injection attacks in power systems: A dynamic bayesian game-theoretic approach," *ISA transactions*, vol. 115, pp. 108–123, 2021.

[283] Fudenberg, D. and Tirole, J., *Game theory*. MIT press, 1991.

[284] Başar, T. and Olsder, G. J., *Dynamic noncooperative game theory*. SIAM, 1998.

[285] Bonagura, V., Panzieri, S., Pascucci, F., and Badia, L., "Strategic interaction over age of incorrect information for false data injection in cyber-physical systems," *IEEE Trans. Control Netw. Syst.*, vol. 12, no. 1, pp. 872–881, 2025.

[286] Nakata, S., Arie, T., Akita, S., and Takei, K., "Wearable, flexible, and multifunctional healthcare device with an isfet chemical sensor for simultaneous sweat ph and skin temperature monitoring," *ACS sensors*, vol. 2, no. 3, pp. 443–448, 2017.

[287] Maatouk, A., Kriouile, S., Assaad, M., and Ephremides, A., "The age of incorrect information: A new performance metric for status updates," vol. 28, no. 5, pp. 2215–2228, 2020.

[288] Rosen, J. B., "Existence and uniqueness of equilibrium points for concave n-person games," *Econometrica*, pp. 520–534, 1965.

[289] Badia, L., "Age of information from two strategic sources analyzed via game theory," in *Proc. IEEE Int. Wkshp Comput. Aided Modeling Design Commun. Links Netw. (CAMAD)*, 2021, pp. 1–6.

[290] Bonagura, V., Panzieri, S., Pascucci, F., and Badia, L., "Strategic interaction over age of incorrect information for false data injection in cyber-physical systems," *IEEE Trans. Control Netw. Syst.*, vol. 12, no. 1, pp. 872–881, 2025.

Secure and Resilient Consensus in Networked Multi-Agent Systems

Chapter 11

Introduction to Secure and Resilient Consensus in Networked Multi-Agent Systems

In the previous parts of thus thesis, we addressed two complementary aspects of cyber–physical system protection. First, we introduced a *model-based anomaly detection and isolation* framework capable of identifying and reconstructing the effects of adversarial actions within distributed estimation schemes. Under suitable observability and boundedness assumptions, we showed that the proposed approach allows both the detection of stealthy attacks and the recovery of the nominal system state. However, even when the adversary is detected, complete removal or neutralization may not be feasible in practice.

To mitigate this limitation, the second part of the thesis investigated *information freshness under adversarial conditions*, focusing on how malicious interference and network imperfections affect the timeliness and correctness of the information available to supervisory systems. Through the *Age of Incorrect Information* (AoII) framework, we analyzed how different attack strategies impact the temporal validity and integrity of exchanged data, and we derived optimal defensive policies that balance communication effort and estimation accuracy. This provided a first step toward maintaining system awareness in the presence of persistent or unremovable adversaries.

Building upon these foundations, this part of the thesis extends the discussion from point-to-point estimation and supervision to *distributed coordination and decision-making*. In modern networked control systems and cyber–physical infrastructures, decision making and control are increasingly distributed across multiple agents. Each agent possesses partial information about the global state and must interact with others to reach agreement on shared quantities, such as average measurements, coordinated actions, or resource allocations. This cooperative mechanism, known as *consensus*, underpins a wide range of applications—from vehicle platooning and sensor fusion to distributed optimization and formation control.

Consider a network of n agents interconnected through a directed graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V} = \{1, \dots, n\}$ denotes the set of agents and $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ represents the set of communication links. Let $\mathcal{N}_i^{\text{in}} = \{j \in \mathcal{V} : (j, i) \in \mathcal{E}\}$ be the set of in-neighbors of agent i, i.e., those agents from which i can receive information. Each agent i holds a state variable $w_i(t) \in \mathbb{R}$ that evolves in continuous time according to the protocol

$$\dot{w}_i(t) = \sum_{j \in \mathcal{N}_i^{\text{in}}} a_{ij}\big(w_j(t) - w_i(t)\big), \tag{11.1}$$

where $a_{ij} > 0$ are communication weights associated with existing links, and $a_{ij} = 0$ otherwise. In compact form, the global network dynamics can be written as

$$\dot{\boldsymbol{w}}(t) = -L\boldsymbol{w}(t), \tag{11.2}$$

where $\boldsymbol{w}(t) = [w_1(t), \dots, w_n(t)]^\top$ collects all agent states and L is the (weighted) Laplacian matrix of the graph $\mathcal{G}$.

The consensus problem consists in designing local update laws—such as (11.1)—so that all agents asymptotically reach an agreement, that is,

$$\lim_{t \to \infty} |w_i(t) - w_j(t)| = 0, \qquad \forall\, i, j \in \mathcal{V}. \tag{11.3}$$

When this condition is satisfied, the agents converge to a common value, referred to as the *consensus value*. The specific value achieved depends on the initial conditions, the communication weights, and the topology of $\mathcal{G}$.

It is well known [13] that if $\mathcal{G}$ is directed and strongly connected, then the continuous-time dynamics (11.1) guarantee asymptotic agreement such that

$$\lim_{t\to\infty} \boldsymbol{w}(t) = \mathbf{1}_n \boldsymbol{\zeta}^\top \boldsymbol{w}(0), \tag{11.4}$$

where $\boldsymbol{\zeta}^\top$ is the (normalized) left eigenvector of L associated with the zero eigenvalue, satisfying $\mathbf{1}_n^\top \boldsymbol{\zeta} = 1$. Equation (11.4) implies that the final consensus value is a *weighted average* of the agents' initial states, where the weights correspond to the entries of $\boldsymbol{\zeta}$ and depend on the network topology.

In the particular case where $\mathcal{G}$ is *balanced*—i.e., each node's in-degree equals its out-degree—the Laplacian L satisfies $\mathbf{1}_n^\top L = L\mathbf{1}_n = \mathbf{0}_n$. Under this condition, all agents converge to the arithmetic mean of their initial conditions:

$$\lim_{t\to\infty} \boldsymbol{w}(t) = \frac{1}{n} \mathbf{1}_n \mathbf{1}_n^\top \boldsymbol{w}(0), \tag{11.5}$$

a property commonly referred to as *average consensus*.

In general, consensus dynamics such as (11.1) provide a foundational mechanism for distributed coordination, allowing agents to synchronize states, share information, or cooperatively estimate quantities of interest, even in the absence of centralized control. However, when the communication network is affected by faults, delays, or adversarial interference, convergence to the desired consensus value can no longer be guaranteed—an issue that motivates the development of resilient and secure consensus algorithms discussed in the following sections. This motivates the study of *secure and resilient consensus*, i.e., consensus protocols capable of withstanding faults, attacks, and privacy breaches while preserving convergence to a valid agreement. In this context:

- **Resilience** refers to the ability to maintain consensus accuracy despite a subset of misbehaving or malicious nodes;
- **Security** ensures authenticity and integrity of the exchanged information, preventing data manipulation or spoofing;
- **Privacy preservation** aims to hide sensitive local information while still achieving global agreement.

The secure consensus problem is therefore a natural continuation of the adversarial freshness problem discussed in the previous part of the thesis. While the *Age of Incorrect Information* framework captured how attacks affect the correctness of point-to-point telemetry, secure consensus addresses the distributed counterpart—how multiple agents can safely exchange and fuse information over a potentially compromised network. Both problems share a common challenge: designing communication and decision strategies that remain reliable under uncertainty and adversarial interference.

Relation to the Author's Publications. The research presented in this part of the thesis builds upon the author's contributions to secure and resilient consensus in networked multi-agent systems. The foundations of the *evidence-theoretic resilient consensus* approach were developed in [C1], where the integration of Dempster–Shafer Evidence Theory into the consensus process was proposed to enhance resilience against malicious or unreliable agents. This work, introduced a distributed mechanism for evaluating and updating the trustworthiness of neighboring nodes during consensus, allowing the system to mitigate adversarial influence while maintaining convergence. The theoretical analysis and simulation validation of this framework were further refined and extended in the journal publication [J1]. That study formalized the stability and synchronization properties of nonlinear dynamical systems used for secure information exchange, demonstrating how synchronization can enhance both robustness and data confidentiality.

Organization of the Chapter. The remainder of this chapter is organized as follows:

- **Section 11.1** reviews the main developments in resilient and privacy-preserving consensus for networked multi-agent systems. It outlines the principal mechanisms proposed in the literature to counteract adversarial nodes, ensure reliable coordination, and preserve data confidentiality, thereby positioning the contributions of this work within the broader research landscape.
- **Section 11.2** presents the original contributions of the thesis to secure and resilient consensus. It introduces the evidence-theoretic consensus framework for resilience against malicious agents and the nonlinear synchronization approach for privacy preservation, highlighting their conceptual foundations and analytical properties.

11.1 Literature review on Resilient Consensus

The problem of achieving consensus in distributed multi-agent systems has been extensively investigated over the past two decades. Classical consensus and distributed agreement algorithms [8], [13], [18], [119] provide a principled mechanism for merging locally available information at each node into a globally consistent estimate. These algorithms have found widespread application, ranging from distributed estimation [16] and cooperative control of mobile robots [14] to the modeling of collective dynamics in social and technological networks [291]. Because of this versatility, consensus protocols have been identified as key enablers for addressing complex socio-economic and technological challenges [292], [293], [294]. Yet, despite their robustness and scalability, they are not immune to severe limitations concerning security and privacy, which have become increasingly pressing in modern interconnected infrastructures.

On the one hand, consensus mechanisms are vulnerable to attacks that compromise the integrity of the computation. Malicious agents may join the network, or legitimate nodes may be hijacked by external adversaries, with the goal of biasing the final consensus or even preventing convergence altogether [295]. Such risks are exacerbated in safety-critical domains where sensitive data are involved, including smart grids [30], [105], industrial Internet of Things (IIoT) applications [29], [106], and critical infrastructures where false alarms or manipulated sensor readings could have dramatic consequences [296], [297]. On the other hand, even in the absence of overtly malicious nodes, privacy concerns emerge: the exchange of raw states between agents may reveal confidential information, a scenario particularly problematic when multiple organizations with potentially competing interests must cooperate.

To address resilience against adversarial nodes, a first line of research has explored local filtering strategies. A representative approach is the mean-subsequence-reduced (MSR) algorithm, in which each agent discards extreme values before updating its state [298], [299]. These methods can tolerate a bounded number of adversaries under specific graph-theoretic conditions, but they rely on prior knowledge of the maximum number of malicious nodes or the presence of trusted agents. Furthermore, they only mitigate the influence of adversaries rather than explicitly identifying compromised nodes. Alternative solutions exploit multi-hop or two-hop communication [300], [301], allowing agents to cross-check information from indirect neighbors and thereby detect inconsistencies. While such protocols are more powerful in principle, they typically rest on restrictive assumptions about network density, communication range, or adversary capabilities, which may limit their deployment in real-world settings.

In parallel, reputation-based methods have recently gained attention as a flexible mechanism to identify and isolate unreliable agents [302], [303], [304], [305], [306]. The intuition is to assign and update trust scores to each node, based on observed behavior, and then weigh these scores in subsequent consensus updates. Early designs include committee-based approaches [304], in which small groups of nodes evaluate the behavior of others and assign reputations collectively. While effective in blocking persistently malicious nodes, these schemes introduce significant overhead and may slow down the system evolution.

More recent frameworks compute reputations locally, either on individual links [305] or by aggregating the discrepancies observed across neighborhoods [306]. These methods improve scalability and adaptability, but they must be carefully designed to avoid false positives: legitimate nodes with divergent but valid initial conditions risk being incorrectly penalized. In this context, the Dempster–Shafer Evidence Theory [307], [308], [309] offers a principled way to fuse uncertain and possibly conflicting information from multiple neighbors, and distributed implementations have been explored in [310], [311], [312]. However, most of these rely on aggregating information from the whole network, whereas in many consensus scenarios, only local and neighbor-based assessments are feasible.

Orthogonal to resilience, another central challenge concerns privacy. Even when all agents are honest, the repeated exchange of states may allow curious participants or external observers to reconstruct sensitive data. Several techniques have been developed to mitigate this risk. One prominent family of methods relies on additive-noise obfuscation: agents perturb their transmissions by injecting random values, which are designed to cancel out over time so that the correct average is eventually recovered [20], [28], [313]. Variants extend these techniques to account for delays [314] and quantization effects [26]. Related approaches inject admissible perturbations into the local dynamics so that privacy is preserved without altering the final consensus value [112]. In other cases, differential privacy mechanisms are applied to ensure formal guarantees of data protection [21], [111]. Despite their simplicity, noise-based methods typically introduce a trade-off between accuracy and privacy and may require cumbersome bookkeeping to ensure eventual noise cancellation [315].

A second line of defense resorts to cryptographic primitives. In particular, semi-homomorphic encryption such as the Paillier scheme allows agents to operate on encrypted data, ensuring that only the final consensus value is decrypted [22], [23], [316]. While offering strong confidentiality guarantees, such approaches are computationally demanding and may be unsuitable for time-sensitive applications. To balance security and efficiency, lightweight variants based on quantization [25] or state decomposition [24] have been proposed. Complementary measures such as digital watermarking provide additional protection against unauthorized use or manipulation of exchanged signals [317]. Yet, the challenge remains to simultaneously guarantee strong privacy, computational efficiency, and exact convergence.

Beyond noise and cryptography, nonlinear dynamics and chaos have been investigated as alternative means of embedding privacy into consensus. Chaotic systems exhibit long-term unpredictability and sensitivity to initial conditions, properties that have been exploited in encryption and masking [110], [113], [114]. In particular, Chua oscillators have been used for key generation [115] and image encryption [116], [117], though most of these contributions focus on payload encryption rather than directly preserving privacy in consensus processes. This line of research opens up the possibility of integrating chaotic synchronization with distributed averaging, thereby obfuscating individual contributions while still allowing recovery of the correct aggregate.

Taken together, the literature shows that while several effective mechanisms have been developed, none is fully satisfactory in addressing both resilience and privacy simultaneously. MSR and multi-hop protocols are robust but require restrictive assumptions; reputation-based systems are adaptive but demand careful design to avoid false alarms; noise-injection and cryptographic techniques provide privacy but introduce overhead or accuracy loss; and chaos-based solutions are promising but underexplored in the context of distributed consensus. Recent work suggests that combining linear consensus with the synchronization of nonlinear oscillators, such as Chua systems, can bridge these gaps [27]. This approach avoids quantization errors typical of encryption, eliminates the need for noise bookkeeping, and can naturally integrate local reputation mechanisms based on Evidence Theory [307], [308], [309]. In this way, it points toward a new generation of consensus algorithms that are at once resilient to adversarial behavior, protective of sensitive information, and practical for deployment in real-world, large-scale distributed systems.

11.2 Thesis Contribution

This thesis advances the study of *secure and resilient consensus* in networked multi-agent systems by integrating concepts from evidence theory, nonlinear dynamics, and information privacy. The overarching goal is to ensure that a network of autonomous agents can reach agreement even in the presence of malicious nodes, unreliable communication, and confidentiality constraints. To this end, the thesis introduces two complementary contributions that jointly address resilience and privacy in distributed coordination.

1. Evidence-theoretic resilient consensus. The first contribution is the development of a novel consensus algorithm grounded in *evidence theory*, which enables each agent to evaluate the reliability of its neighbors based on behavioral evidence collected during the consensus process. By modeling trust as a dynamic belief variable, the algorithm can identify and isolate nodes that deviate from expected behavior—either intentionally or due to faults—and automatically discount their influence in the aggregation step. This mechanism allows the network to maintain convergence toward a common value even when a subset of agents behaves maliciously or transmits inconsistent data. Analytical results establish sufficient conditions for convergence and resilience under bounded adversarial activity.

2. Privacy-preserving consensus via nonlinear synchronization. The second contribution exploits the intrinsic synchronization properties of certain classes of nonlinear dynamical systems to design a consensus protocol that preserves the privacy of individual agents. Instead of exchanging explicit state values, agents share only partially transformed or encrypted information, allowing legitimate nodes to synchronize without revealing their internal states. This approach effectively integrates encryption within the consensus dynamics, ensuring that coordination can be achieved securely without relying on external cryptographic layers. Theoretical analysis and simulation results demonstrate that the proposed nonlinear exchange mechanism maintains convergence while guaranteeing state confidentiality.

Chapter 12

Privacy-Preserving Consensus

In this chapter, we analyze the effects of adversarial manipulation on networked consensus dynamics and discussed algorithmic strategies to ensure resilience against malicious agents. However, protection against external attacks is only one side of the problem. Even when all agents behave correctly, the continuous exchange of information required to achieve consensus may inadvertently disclose sensitive data—such as local measurements, private preferences, or operational states.

In many cyber–physical and networked decision systems, the states of individual agents encode private or commercially sensitive information. Examples include smart grids, where energy consumption reveals user behavior; distributed sensing, where local measurements carry confidential data; and financial or industrial networks, where local costs or risks must remain private. Traditional consensus protocols, which rely on explicit state exchange, inherently lack privacy protection: an adversary eavesdropping on communication links—or even a legitimate but curious participant—can often reconstruct other agents' initial conditions by observing the iterative evolution of the states.

Building on the formalism introduced in Chapter 11, this chapter focuses on the complementary issue of *privacy preservation* in distributed consensus. While the previous discussion emphasized robustness and resilience against malicious interference, here we address the challenge of safeguarding sensitive local information during cooperative decision making. The objective is to design mechanisms that guarantee convergence to a correct global agreement while preventing external observers—or other agents—from inferring private data.

Structure of the chapter. The remainder of the chapter is organized as follows.

- **Section 13.2** introduces the formal problem setting, including the system model, the definition of malicious agents, and the reputation-based consensus dynamics.
- **Section 12.2** details the proposed dynamic reputation assignment mechanism based on Evidence Theory, while Section 12.3 presents the weight correction algorithm that compensates transient distortions. Finally, simulation results are provided in
- **Section 12.4**, and concluding remarks are summarized in Section 13.5.

12.1 Problem Setting

Let us consider a scenario where a set of n agents is interconnected by an undirected graph $G = \{V, E\}$, and each of them holds an initial condition[1] $x_0^{(i)} \in \mathbb{R}$. The aim of each agent is to reach the average of the initial conditions, by interacting via a discrete-time distributed consensus dynamics

$$x_{k+1}^{(i)} = x_k^{(i)} + \tau \sum_{j \in \mathcal{N}_i} \left(x_k^{(j)} - x_k^{(i)} \right), \tag{12.1}$$

where $x^{(i)}, x^{(j)} \in \mathbb{R}$ and τ is a scalar such that $\tau \in (0, 1/\max^{(i)}\{|\mathcal{N}_i|\}$. Let us now consider the presence of corrupted nodes in the network, which aim to divert the algorithm and compromise the achievement of the average initial conditions. For the attack model, we consider the presence of *malicious* agents, defined as follows.

Definition 12.1.1 (*Malicious Agent*)

An agent v_i is malicious *if it is able to arbitrarily change the value of its state and*

[1] For the sake of simplicity, In this chapter we consider the scalar case.

transmit values that are different from the ones expected according to the nominal interaction scheme. □

Since, as is well known, the action of an agent in the network characterized by a undirected graph affects the final value by $1/n$, in order to have the greatest possible effect, a reasonable strategy for a malicious node should be to propose a value that is out of scale, and far from the initial conditions of the other agents in the network. In order to allow the malicious agent to make an effective attack and put us in the worst possible position to counter its effect, we consider the following assumption.

Assumption 12.1.1

Each malicious node has complete knowledge of the initial conditions of all other agents. □

Thanks to this assumption, the malicious node is able to select out-of-scale values to inject into the network. In order for the standard consensus algorithm to become resilient to this type of attack, one possible strategy is to build a system of reputations capable of identifying corrupted nodes.

Definition 12.1.2 (*Local Reputation*)

The local reputation $r_k^{(i,j)} \in [0,1]$ *is a numerical value assigned at each instant k to node a $u^{(j)} \in \mathcal{N}_i$ by node v_i, locally computed using its own evidence and the one provided by its neighbors. This value will be closer to* 1 *the more node $u^{(j)}$ proves to be correctly following the protocol according to v_i.* □

Let us use $\mathcal{M} \subset \{v_1, \dots, v_n\}$ to denote the set of malicious agents and let $\mathcal{R} = \{v_1, \dots, v_n\} \setminus \mathcal{M}$ denote the non-malicious ones. We consider the following dynamics for each agent $v_i \in \mathcal{R}$

$$x_{k+1}^{(i)} = x_k^{(i)} + \tau \sum_{j \in \mathcal{N}_i} r_k^{(i,j)} \left(x_k^{(j)} - x_k^{(i)} \right) + u_k^{(i)}, \tag{12.2}$$

where $u_k^{(i)}$ is an input chosen by each agent (to be discussed in Section 12.3) and the term $r_k^{(i,j)}$ represents the time-varying weight expressing the reputation of node $u^{(j)}$ according to node v_i, with $u^{(j)} \in \mathcal{N}_i$. As for the agents $v_i \in \mathcal{M}$, we assume

$$x_{k+1}^{(i)} = z_k^{(i)}, \tag{12.3}$$

where $z_k^{(i)}$ is an arbitrary value that could either be random or chosen ad hoc by the $i-$th agent[2].

Remark 12.1.1

The dynamics of all the agents $v_i \in \mathcal{M}$, expressed Eq. (12.3), evolves according to an arbitrary update and without regard to the value of reputations, as if $r_k^{(i,j)} = 0$ for all $u^{(j)} \in \mathcal{N}_i$. □

12.2 Dynamic Reputation Assignment

The proposed reputation algorithm is based on the Evidence Theory [307], [308], which makes it possible to construct belief functions that base the degree of confidence of a hypothesis on the subjective probability of the given hypothesis, and those related to it. This allows to embrace of a wide variety of rules that can be detailed as desired, since belief functions can be easily modulated. In this chapter, we propose a choice of such

[2]For instance, the malicious agent might choose to update its state so as to slowly diverge from the consensus value.

functions that allows us to encompass different scenarios, including cases in which we have outlier nodes, malicious nodes that voluntarily do not follow the protocol, or even nodes that are subject to faults or which are corrupted and controlled from a certain time on. To this end, we rely on the key concept of distance between the values of two nodes as an indicator of data reliability in the consensus algorithm. This choice, adopted by a wide variety of approaches in the literature [298], [299], is justified by the fact that, in a consensus-type algorithm in which all states tend to the same value, the decreasing distance between them indicates good reliability in following the protocol. In particular, the bigger the state distance between two nodes, the more the mutual reputation should drop.

Definition 12.2.1 (*State distance between two nodes*)

The state distance between two nodes $v_i \in V$ and $u^{(j)} \in V$ calculated respectively at the time-instant k_i and k_i is defined as $d^{(i,j)}(k_i, k_j) = |x_{k_i}^{(i)} - x_{k_j}^{(j)}|$. □

If $k_i = k_j$, the notation can be simplified referring to $d^{(i,j)}(k_i, k_j)$ as $d_k^{(i,j)}$. Some definitions are in order to properly characterize the proposed reputation algorithm.

Definition 12.2.2 (*Frame of Discernment*)

The frame of discernment $\Omega = \{\omega_1, \ldots, \omega_n\}$ *is a finite set of possible values of a variable ω, where the element ω_i are assumed to be mutually exclusive.* □

Therefore, the frame of discernment essentially amounts to all the hypotheses that it is possible to make about the problem at hand. In this chapter, we consider a particular frame of discernment, set as $\Omega = \{F, S\}$, where F stands for Fake (i.e., a malicious agent) and S stands for Sincere (i.e., a regular node).

Definition 12.2.3 (*Power Set*)

The power set *originated by the frame of discernment Ω is defined as $\Gamma(\Omega) = \{\gamma_1, \ldots, \gamma_{|\Gamma|}\}$, and contains every subset $\gamma_i \subseteq \Omega$. The cardinality of the power set is $2^{|\Omega|}$.* □

In other words, the power set is formed by all possible subsets of the frame of discernment. Based on Ω, for the problem at hand, we have that $\Gamma(\Omega) = \{F \cup S, F, S, \emptyset\}$.

Definition 12.2.4 (*Basic Probability Assignment-BPA*)

The basic probability assignment *(also known as* mass function *or* belief function*) is represented by a function $m : \Gamma(\Omega) = 2^{\Omega} \to [0, 1]$. It is constrained by $m_{\emptyset} = 0$ and $\sum_{\gamma_i \in \Gamma(\Omega)} m_{\gamma_i} = 1$.* □

In this framework, the focus is on quantifying the belief of propositions of the form: *the truth value ω is contained in γ_i*. The value of the BPA for the set γ_i, represented by $m^{(\gamma_i)}$, expresses the quantity of evidence that supports the claim that the true value belongs to the set γ_i, but to no particular subset of γ_i. Considering a BPA assignment, the elements of the power set with values greater than zero are called *focal sets*. Assuming that $b_k^{(i,j)} \in \Gamma(\Omega)$ is the time-varying belief distribution of the node $u^{(j)}$ according to node v_i and referring to BPA, we consider a belief assignment in the form:

$$b^{(i,j)}(k_i, k_j) = \left\{1 - f(d^{(i,j)}(k_i, k_j)), 0, f(d^{(i,j)}(k_i, k_j)), 0\right\}, \tag{12.4}$$

where f is a function of the distance $f(d^{(i,j)}(k_i, k_j))$ used to assign a belief to the fact $u^{(j)}$ is sincere. Notice that if $k_i = k_j$, the notation can be simplified to $b_k^{(i,j)}$.

In the case study presented In this chapter, each node on the basis of the information available to it, attributes a BPA to all its neighbors. Afterward, each node sends its own evidence to its neighbors and combines the received information to derive a common belief of the current situation. In this work, we denote by $b_k^{(i,j)}$ the belief function of the node v_i with respect to the node $u^{(j)}$ defined locally through the direct assignment of

for $v_j \in \mathcal{N}_i$ **do**
  $m_k^{(i,j)} = \texttt{Smets}\{b_k^{(i,j)}, b_{k,0}^{(i,j)}\}$;
  for $v_l \in \mathcal{N}_i$ **do**
    if $v_j \neq v_l$ **then**
      $m_k^{(i,j)} = \texttt{Smets}\{m_k^{(i,j)}, b_k^{(l,j)}\}$;
for $v_j \in \mathcal{N}_i$ **do**
  if $m_{S,k}^{(i,j)} \geq \alpha$ **then**
    $r_k^{(i,j)} = m_{S,k}^{(i,j)}$
  else
    $r_k^{(i,j)} = 0$

Algorithm 12.1: Agent v_i computes $r_k^{(i,j)}$ for all $u^{(j)} \in \mathcal{N}_i$

mass, and with $m_k^{(i,j)}$ the outcome of the combination of all the information received. Several combination rules have been proposed over the years, with different features and different application fields. Among them, we choose Smets' combination rule, since it uses the principle of *pignistic probability* [318] to combine multiple probability estimates into a single, overall probability estimate, and avoids the normalization while preserving commutativity and associativity properties. Then, $\forall \gamma_a, \gamma_b, \gamma_c \in \Gamma(\Omega)$ it holds

$$m_{\gamma_a,k}^{(i,j)} = \texttt{Smets}\{b_k^{(i,j)}, b_k^{(l,j)}\}_{\gamma_a} = \sum_{\gamma_b \cap \gamma_c = \gamma_a} b_{\gamma_b,k}^{(i,j)} b_{\gamma_c,k}^{(l,j)}.$$

We assume that $m_k^{(i,j)}$ is the time-varying mass function of the node $u^{(j)} \in \mathcal{N}_i$ according to node v_i, and it is obtained by merging evidences coming from v_i and $v_l \in \mathcal{N}_i, v_l \neq u^{(j)}$. The time-varying weight expressing the reliability of the node $u^{(j)}$ according to node v_i with $u^{(j)} \in \mathcal{N}_i$ is denoted by $r_k^{(i,j)}$ and is computed from $m_k^{(i,j)}$ with the novel Algorithm (12.1), which introduces the protocol from the point of view of the i-th agent.

Note that the calculation of the mass function at time-instant k does not take into account the mass function calculated at time-instant $k-1$, but it could be modified to include past evaluations to counteract different attackers models. For example it could be useful in all cases in which the attacker starts far from the mean value, but converge towards other nodes.

In practice, Algorithm (12.1) performs the calculation of $m_k^{(i,j)}$ with $u^{(j)} \in \mathcal{N}_i$, relying both on the belief function of the node $u^{(j)}$ according to the node v_i, and the believe function of the node $u^{(j)}$ according to the nodes v_l for each $v_l \in \mathcal{N}_i \setminus u^{(j)}$. Then, $r_k^{(i,j)}$ is computed exploiting the built up evidence that the node $u^{(j)}$ according to the node v_i is Sincere. The parameter $\alpha \in [0,1]$ is used to permanently exclude nodes with a too low reputation from the network, and can be chosen arbitrarily. Specifically, when the evidence "is Sincere" of the node $u^{(j)}$ according to the node v_i falls below α, the node $u^{(j)}$ is treated by node v_i as a malicious agent by placing the relative reliability weight equal to 0.

12.3 Weight Correction

In the previous section, we have developed a framework for letting each node v_i assess the reputation of a neighbor $u^{(j)}$ based on evidence theory. In this view, in the event that the reputations converge to a fixed value, it may be desirable to weigh the agents' states based on their final weights, as this would provide a more accurate representation of their current reputations. Yet, without any compensation in place, the consensus process relies on incorrect weights during the transient, distorting the overall result. In order to tackle this issue, in this section we propose to implement a distributed strategy to track down the last available weights and to enforce convergence of the state of non-malicious agents

to a weighted average of their initial conditions, as if the limit reputation values were used right from the beginning. Generally speaking, let us consider a situation where the reputation weights reach steady-state values $\overline{r}^{(i,j)}$ and the set of agents $\{v_1, \ldots, v_n\}$ is partitioned into two mutually exclusive sets $\mathcal{Q}$ and $\mathcal{J}$, where the first one contains agents v_i that are associated to nonzero steady-state reputation values $\overline{r}^{(i,j)}$ for at least one neighbor $u^{(j)}$. Moreover, denoting by $\widehat{G}$ the obtained from G by removing the nodes in $\mathcal{J}$ along with their incident edges, let us assume that $\widehat{G}$ is connected. With respect to this scenario, we want each agent in $\mathcal{Q}$ to converge to the same asymptotic state of the following ideal dynamics

$$\tilde{x}_{k+1}^{(i)} = \tilde{x}_k^{(i)} + \tau \sum_{j \in \mathcal{N}_i \bigcap \mathcal{Q}} \overline{r}^{(i,j)} \left(\tilde{x}_k^{(j)} - \tilde{x}_k^{(i)} \right), \tag{12.5}$$

with $\tilde{x}_0^{(i)} = x_0^{(i)}$. In this view, we assume that the aim of each agent $v_i \in \mathcal{Q}$ is to select the input $u_k^{(i)}$ guaranteeing that it holds $\lim_{k\to\infty} |x_k^{(i)} - \tilde{x}_k^{(i)}| = 0$. Let us define $\overline{L} = \lim_{k\to\infty} L_k$, where L_k is the weighted Laplacian matrix corresponding to the weights $r_k^{(i,j)}$. For the sake of simplicity, in the remainder of this chapter, we assume that the agents are ordered so that the first $|\mathcal{Q}|$ ones correspond to the agents in $\mathcal{Q}$, while the remaining $n - |\mathcal{Q}|$ ones belong to $\mathcal{J}$. Interestingly, for the scenario at hand, we have that

$$\overline{L} = \begin{bmatrix} \widehat{L} & 0_{|\mathcal{R}|\times|\mathcal{M}|} \\ * & * \end{bmatrix},$$

where $\widehat{L}$ is the limiting Laplacian matrix associated to the graph $\widehat{G}$. Notice that, by construction, the ideal dynamics for the agents in $\mathcal{Q}$ is such that $\lim_{k\to\infty} x^{(i)}{}_k = \overline{\boldsymbol{w}}^T \boldsymbol{x}_0$, where $\overline{\boldsymbol{w}} = [\widehat{\boldsymbol{w}}^T \ \mathbf{0}_{|\mathcal{J}|}^T]^T$ and $\widehat{\boldsymbol{w}} \in \mathbb{R}^{\mathcal{Q}}$ is the unit sum left eigenvector of the matrix $\widehat{L} \in \mathbb{R}^{|\mathcal{Q}|\times|\mathcal{Q}|}$ associated to the zero eigenvalue. Therefore, the agents' objective is equivalent to requiring that $\lim_{k\to\infty} x_k^{(i)} = \overline{\boldsymbol{w}}^T \boldsymbol{x}_0$, for all $v_i \in \mathcal{Q}$. Our strategy to accomplish this task is to extend the state of each agent in $\mathcal{Q}$ by considering variables $\phi_k^{(i,j)}$ for all $u^{(j)} \in \mathcal{N}_i$ with dynamics $\phi_{k+1}^{(i,j)} = \phi_k^{(i,j)} + \tau \left(x_k^{(j)} - x_k^{(i)} \right)$ and with $\phi_0^{(i,j)} = 0$. In other words, the variable $\phi_k^{(i,j)}$ accumulates the sum of the overall influence of the j-th agent on the i-th one.

Based on the above augmented state variables, each node $v_i \in \mathcal{Q}$ sets $u_k^{(i)} = \sum_{j\in\mathcal{N}_i} \left(r_k^{(i,j)} - r_{k-1}^{(i,j)} \right) \phi_k^{(i,j)}$, where we define $r^{(i,j)}(-1) = 0$.

Overall, the dynamics of the i-th agent in $\mathcal{Q}$ is given by

$$\begin{cases} x_{k+1}^{(i)} = x_k^{(i)} + \tau \sum_{j\in\mathcal{N}_i} r_k^{(i,j)} \left(x_k^{(j)} - x_k^{(i)} \right) \\ \qquad\qquad + \sum_{j\in\mathcal{N}_i} \left(r_k^{(i,j)} - r_{k-1}^{(i,j)} \right) \phi_k^{(i,j)}, \\ \phi_{k+1}^{(i,j)} = \phi_k^{(i,j)} + \tau \left(x_k^{(j)} - x_k^{(i)} \right), \ \forall j \in \mathcal{N}_i. \end{cases} \tag{12.6}$$

In order to prove convergence of the error with respect to the ideal consensus, we need the following ancillary lemma, which inspects the structure of the state at time k.

Lemma 12.3.1

Assume that the reputation values reach a steady-state and that the agents in $\mathcal{Q}$ interact according to the protocol in Eq. (12.6), while the agents in $\mathcal{J}$ behave according to Eq. (12.3). Moreover, assume that $\widehat{G}$ is connected and consider the partition of the states $\boldsymbol{x}_k = \begin{bmatrix} \boldsymbol{x}_{\mathcal{Q},k}^T & \boldsymbol{x}_{\mathcal{J},k}^T \end{bmatrix}^T$. Then it holds

$$\boldsymbol{x}_k = \Gamma_k \boldsymbol{x}_0 + \sum_{h=0}^{k-1} \Gamma_{k-1-h} \boldsymbol{q}_h, \tag{12.7}$$

where $\Gamma_0 = I$ *while, for* $k > 1$

$$\Gamma_k = \left(I + L_{k-1} \sum_{s=1}^{k} (-1)^s \tau^s F(s,k) \right), \tag{12.8}$$

$$F(s,k) = \begin{cases} kI & \text{if } s = 1, \\ \prod_{u=k-s+1}^{k-2} L(u) \sum_{h=0}^{k-s} (h+1) L_h & \text{otherwise.} \end{cases}$$

and $\boldsymbol{q}_k = \begin{bmatrix} \mathbf{0}_{|\mathcal{R}|}^T & \boldsymbol{z}_k^T \end{bmatrix}^T$, *with* $\boldsymbol{z}_k \in \mathbb{R}^{\mathcal{J}}$ *defined as the stack of the terms* $z_k^{(i)}$ *for all* $v_i \in \mathcal{J}$. □

Proof. Let us use a proof by induction. To this end we observe that for $k = 1$ it holds $\boldsymbol{x}(1) = (I - \tau L_0)\boldsymbol{x}_0 + \boldsymbol{q}_0$, which has the structure of Eq. (12.7); hence, the basic step is verified. Let us now consider the inductive step. Specifically, let us assume that Eq. (12.7) holds true for all $h \in \{0, \ldots, k\}$, with $k \geq 1$, and and let us show that, then, it holds at the $_{k+1}$-th step. Let us use $\Phi_k \in \mathbb{R}^{n \times n}$ to collect the variables $\phi_k^{(i,j)}$, i.e., $\Phi_k^{(i,j)} = \phi_k^{(i,j)}$ if $v_i \in \mathcal{Q}$ and $(u^{(j)}, v_i) \in E$, while $\Phi_k^{(i,j)} = 0$, otherwise. We have that, by Eq. (12.6) $\Phi_k = -\tau \sum_{h=0}^{k-1} \Delta_h$, where $\Delta_k \in \mathbb{R}^{n \times n}$ is such that $\Delta_k^{(i,j)} = x_k^{(i)} - x_k^{(j)}$ if $v_i \in \mathcal{Q}$ and $(u^{(j)}, v_i) \in E$, while $\Delta_k^{(i,j)} = 0$, otherwise. Moreover, defining $R_k \in \mathbb{R}^{n \times n}$ as the matrix such that $r_k^{(i,j)} = r_k^{(i,j)}$ if $(u^{(j)}, v_i) \in E$ and $r_k^{(i,j)} = 0$, otherwise, we have that $(R_k \odot \Delta_k)\mathbf{1}_n = -\tau L_k \boldsymbol{x}_h$, where $\odot$ is the Hadamard (entrywise) product. At this point, stacking Eq. (12.6) for all agents in $\mathcal{R}$ and Eq. (12.3) for all agents in $\mathcal{J}$, we obtain

$$\begin{aligned} \boldsymbol{x}(k+1) &= (I - \tau L_k)\boldsymbol{x}_k + (R_k - R_{k-1}) \odot \Phi_k + \boldsymbol{q}_k \\ &= (I - \tau L_k)\boldsymbol{x}_k \\ &\quad - \tau \left((R_k - R_{k-1}) \odot \sum_{h=0}^{k-1} \Delta_h \right) \mathbf{1}_n + \boldsymbol{q}_k \\ &= (I - \tau L_k)\boldsymbol{x}_k - \tau (L_k - L_{k-1}) \sum_{h=0}^{k-1} \boldsymbol{x}_h + \boldsymbol{q}_k. \end{aligned}$$

Plugging the expression of $\boldsymbol{x}_h$ as in Eq. (12.7) for all $h \in \{0, \ldots, k\}$ in the above equation, we have that

$$\begin{aligned} \boldsymbol{x}_{k+1} &= \Gamma_k \boldsymbol{x}_0 + \sum_{h=0}^{k-1} \Gamma_{k-1-h} \boldsymbol{q}_h \\ &\quad - \tau (L_k - L_{k-1}) \sum_{h=0}^{k-1} \Gamma_h \boldsymbol{x}_0 \\ &\quad - \tau (L_k - L_{k-1}) \sum_{h=0}^{k-1} \sum_{\ell=0}^{h-1} \Gamma_{k-1-\ell} \boldsymbol{q}_\ell + \boldsymbol{q}_k. \end{aligned}$$

By some algebra, it can be shown that $\Gamma_{k+1} = \Gamma_k - \tau (L_k - L_{k-1}) \sum_{h=0}^{k-1} \Gamma_h$ and that

$$\begin{aligned} \sum_{h=0}^{k} \Gamma_{k-h} \boldsymbol{q}_h &= \sum_{h=0}^{k-1} \Gamma_{k-1-h} \boldsymbol{q}_h \\ &\quad - \tau (L_k - L_{k-1}) \sum_{h=0}^{k-1} \sum_{\ell=0}^{h-1} \Gamma_{k-1-\ell} \boldsymbol{q}_\ell + \boldsymbol{q}_k, \end{aligned}$$

and thus Eq. (12.7) holds also at $k + 1$. This completes our inductive proof.

□

We are now in position to prove the main result of this section.

Theorem 12.3.1

Let the assumptions of Lemma 12.3.1 hold true. Then, it holds $\lim_{k\to\infty} x_k^{(i)} = \widehat{\boldsymbol{w}}^T \boldsymbol{x}_{\mathcal{Q},0}$, *for all* $v_i \in \mathcal{Q}$. □

Proof. In order to prove the result we observe that, since we assumed the reputation values converge to a steady state, as k goes to infinity $r_k^{(i,j)} - r_{k-1}^{(i,j)}$ approaches zero. Similarly, for k approaching infinity the inputs $u_k^{(i)}$ approach zero, while L_k approaches $\overline{L}$. Therefore, we can equivalently express the dynamics of the agents in $\mathcal{Q}$ as $\boldsymbol{x}_{\mathcal{Q},k+1} = (I - \tau\widehat{L})\boldsymbol{x}_{\mathcal{Q},k} + \boldsymbol{\eta}_k$, for some $\boldsymbol{\eta}_k$ such that $\lim_{k\to\infty} \boldsymbol{\eta}_k = \mathbf{0}_n$, i.e., $\boldsymbol{\eta}_k$ is vanishing. Notice that, since the graph $\widehat{G}$ is connected, and it features bidirectional links (although with different weights in the two directions). Hence, the zero eigenvalue of $\widehat{L}$ is simple (e.g., see [107]) while all other eigenvalues have negative real parts. Therefore, as k goes to infinity, the state has, in general, nonzero projection only along the direction spanned by the eigenvector associated to the zero eigenvalue, while the projection along the other eigenvectors is zero. In particular, we have that $\lim_{k\to\infty} \overline{\boldsymbol{w}}^T L_{k-1} = \overline{\boldsymbol{w}}^T \overline{L} = \mathbf{0}_n^T$, therefore, by Lemma 12.3.1, it holds

$$\begin{aligned}\lim_{k\to\infty} \overline{\boldsymbol{w}}^T \boldsymbol{x}_k = \overline{\boldsymbol{w}}^T \boldsymbol{x}_0 + \\ + \underbrace{\lim_{k\to\infty} \overline{\boldsymbol{w}}^T L_{k-1} \sum_{s=1}^{k} (-1)^s \tau^s F(s,k) \boldsymbol{x}_0}_{\mathbf{0}_n} \\ + \lim_{k\to\infty} \overline{\boldsymbol{w}}^T \sum_{h=0}^{k-1} \Gamma_{k-1-h} \boldsymbol{q}_h.\end{aligned}$$

At this point we observe that, as k approaches infinity, for h finite it holds $\lim_{k\to\infty} \Gamma_{k-1-h} = \lim_{k\to\infty} \Gamma_k$; conversely, when also h approaches infinity we have that $\lim_{k\to\infty} \Gamma_{k-1-h} = \lim_{k\to\infty} \Gamma_0 = I_n$. As a consequence, we have that, in both cases, $\lim_{k\to\infty} \overline{\boldsymbol{w}}^T \Gamma_{k-1-h} \boldsymbol{q}_h = \lim_{k\to\infty} \overline{\boldsymbol{w}}^T \boldsymbol{q}_h = 0$ and thus $\lim_{k\to\infty} \overline{\boldsymbol{w}}^T \sum_{h=0}^{k-1} \Gamma_{k-1-h} \boldsymbol{q}_h = 0$. We have established that $\lim_{k\to\infty} \overline{\boldsymbol{w}}^T \boldsymbol{x}_k = \overline{\boldsymbol{w}}^T \boldsymbol{x}_0 = \widehat{\boldsymbol{w}}^T \boldsymbol{x}_{\mathcal{Q},0}$ and the proof is complete. □

12.4 Simulations

In this section, we propose a simulation campaign that will demonstrate the effectiveness of the proposed approach. In particular, we consider the undirected network topology $G = \{V, E\}$ with $|V| = 10$ nodes and $|E| = 19$ edges (reported in Figure 12.2a). Moreover we assume the agents have an initial condition equal to

$$\begin{aligned}\boldsymbol{x}_0 = \big[&11.3555 \quad 12.4915 \quad 13.8727 \quad 14.6132 \quad 15.6451 \\ &19.6826 \quad 10.0000 \quad 18.6496 \quad 20.8000 \quad 22.4395\big].\end{aligned}$$

We consider a scenario where $\mathcal{M} = \{7, 9\}$ and we assume that the malicious nodes do not follow the consensus protocol, but progressively increase their state by 0.1 at each iteration (i.e., they set $z_k^{(i)} = x_0^{(i)} + 0.1k$), with the aim to drag the states of the remaining agents towards larger and larger values. The effectiveness of this type of attack is shown in Figure 12.2b, where we can observe that, as the traditional consensus algorithm proceeds, the agents' states diverge following the lead of the two malicious nodes. Moreover, in our simulations, we assume that the function used by node v_i to assign a belief on the fact that a neighbor $u^{(j)}$ is not malicious is $f(d_k^{(i,j)}) = \xi_i/(\xi_i + d_k^{(i,j)})$, where ξ_i is a parameter locally chosen by each node to control the influence of the distance between the values v_i and $u^{(j)}$ on the credibility of node $u^{(j)}$, set in our case as $\xi_i = 2$. With the network

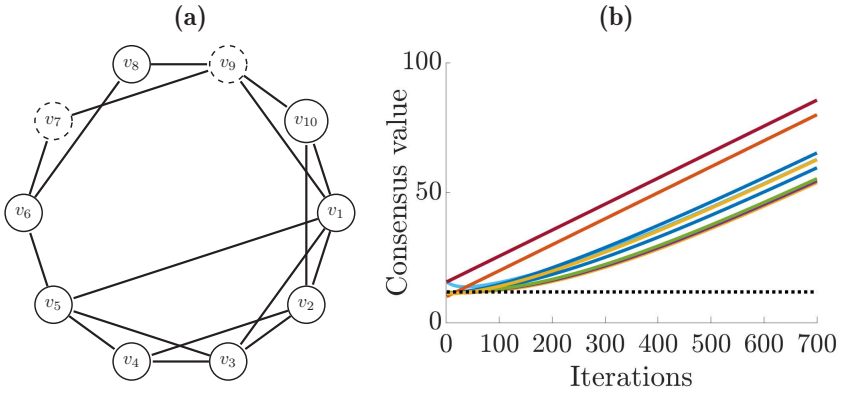


Figure 12.2: (a) Communication graph used in the simulations, where malicious nodes are highlighted with dashed boundaries. (b) Temporal evolution of the traditional consensus algorithm in the presence of two malicious nodes; the black dotted line indicates the average of the initial conditions of the legitimate agents.

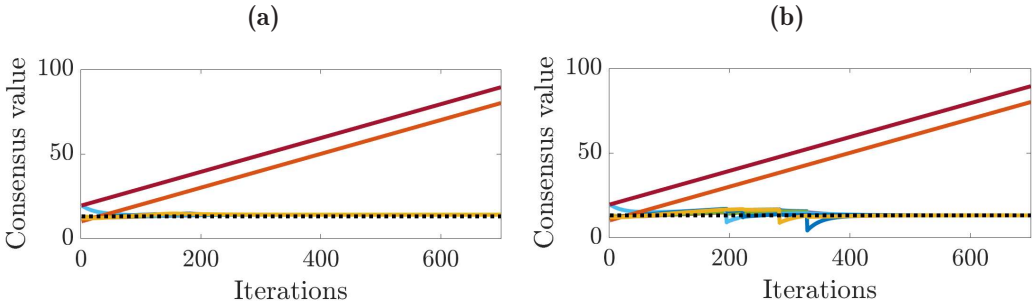


Figure 12.3: Temporal evolution of the weighted consensus algorithm employing (a) the reputation-based scheme and (b) the weight-correction mechanism, in the presence of two malicious nodes. The black dotted line indicates the average of the initial conditions of the legitimate agents.

and attack conditions just described, we then applied the reputation-weighted consensus strategy based on Evidence Theory. We set the value of the parameter $\alpha = 0.3$, i.e., the threshold triggering the disconnection of the link under consideration. In Figure 12.3a, it can be noted that, due to the effect of the reputations, the nodes actually manage to reach a consensus, although the result differs from the one they would have reached if the two malicious nodes were to be removed right from the start. Interestingly, according to the figure, we observe that the malicious nodes contribute to the dynamics of the other agents during the first iterations of the algorithm, i.e., before being removed. In order to further counter their effect, we complement the weighted consensus strategy based on reputations with the weight correction algorithm discussed in Section 12.3. Therefore, within Figure 12.3b, the weighted consensus obtained via the weight correction algorithm essentially amounts to the weighted average of the initial conditions of the legitimate agents, as if the malicious nodes were excluded from the beginning.

12.5 Discussion

In this chapter, we investigate a scenario in which a network of agents aims to implement a distributed consensus process, despite the presence of some malicious agents which attempt to steer the other agents' dynamics away from the intended consensus value. By resorting to Evidence Theory, we develop a methodology to allow agents to compute a

reputation value for their neighbors, and we show that such a reputation can be useful to reduce the influence of malicious nodes. However, before such reputation values converge to a steady state, malicious nodes are able to influence the other agents, preventing the reach of the intended agreement. To address this issue, we develop a weight correction algorithm that allows non-malicious agents to reach the same asymptotic state they would have reached if the asymptotic reputation values were to be used right from the start. Future work will be mainly aimed at identifying effective ways to tune the parameters of the proposed approach, e.g., the function $f(\cdot)$ and the threshold α, to guarantee effective detection of malicious agents. Moreover, we will extend our framework to the asynchronous case, and we will aim to provide guarantees in terms of robustness with respect to noise, disturbances and packet losses.

Chapter 13

Privacy-Preserving Consensus

In the previous chapters, we analyzed how adversarial actions can compromise the correctness and reliability of networked control systems, and we developed strategies for achieving *secure consensus*, where agents are able to coordinate correctly even in the presence of malicious or faulty nodes. However, while security and resilience protect against external manipulations, they do not necessarily guarantee the *confidentiality* of the information exchanged among agents. Even in attack-free scenarios, conventional consensus algorithms require participants to share their internal states or measurements over communication links, potentially revealing sensitive information to honest-but-curious peers or to external eavesdroppers.

In many cyber–physical networks, the local state of an agent inherently carries private data—for instance, energy consumption in smart grids, production levels in industrial plants, or individual preferences in distributed decision-making systems. Classical consensus protocols achieve global agreement through iterative averaging of exchanged states, but this transparency makes it possible to reconstruct the agents' initial conditions, especially when the network topology grants some nodes high observability over their neighbors. Thus, beyond resilience and security, there arises a crucial need for *privacy-preserving consensus* mechanisms that enable cooperation without disclosing private information.

This chapter addresses the problem of designing distributed consensus algorithms that ensure both correctness and confidentiality. Formally, the objective is to construct a protocol such that:

1. all agents asymptotically reach a common (possibly weighted) consensus value based on their initial states;
2. no agent, nor any external observer, can infer another agent's private initial state from the information exchanged during the process.

This formulation extends the notion of secure consensus by focusing on the protection of *privacy* rather than merely the integrity of transmitted data.

To achieve this goal, we develop a novel analytical framework that combines concepts from nonlinear synchronization, consensus dynamics, and control-based privacy mechanisms. The central idea is to exploit the natural synchronization properties of nonlinear dynamical systems as a means of *dynamic masking*: agents exchange only distorted versions of their internal variables, yet these distorted signals still lead to collective convergence toward the desired consensus value. Specifically, each agent embeds its private state within a hidden dynamic variable that evolves in a coupled nonlinear system. This variable is never directly transmitted, ensuring that the true initial state remains confidential while synchronization guarantees convergence to a consistent global value. In this way, the proposed approach achieves both resilience against observation attacks and privacy against honest-but-curious participants, bridging the gap between cooperative control and information confidentiality in distributed multi-agent systems.

Organization of the chapter.
The remainder of this chapter is organized as follows.

- **Section 13.1** reviews the theoretical background on synchronization in networks of nonlinear continuous-time dynamical systems, which forms the basis of our approach.

- **Section 13.2** introduces the proposed privacy-preserving consensus protocol and provides a convergence analysis.
- **Section 13.3** discusses the privacy and security guarantees of the proposed framework, highlighting its robustness against eavesdroppers and exposed nodes.
- **Section 13.4** presents numerical simulations demonstrating the effectiveness of the proposed method in both balanced and unbalanced network topologies, while
- **Section 13.5** concludes the chapter with a discussion of potential extensions and applications.

13.0.1 Privacy-Preserving Consensus

Suppose the agents aim to compute a weighted average of their initial conditions $w_i(0)$ with weights encoded by $\boldsymbol{\zeta}$ and, at the same time, want to prevent their neighbors from being able to ascertain their initial conditions. Over a strongly connected directed graph, a possible strategy to achieve this result [313], [316] is to modify the consensus protocol by setting

$$\dot{w}_i(t) = \sum_{j \in \mathcal{N}_i^{\text{in}}} (w_j(t) - w_i(t)) + u_i(t)$$

where $u_i(t)$ represents artificial noise injected by the i-th agent; in compact form, the dynamics reads

$$\dot{\boldsymbol{w}}(t) = -L\boldsymbol{w}(t) + \boldsymbol{u}(t).$$

In particular, suppose $u_i(t)$ is such that

$$\int_0^{t_i^*} u_i(\tau)d\tau = 0, \quad u_i(t) = 0 \text{ for } t > t_i^* \tag{13.1}$$

and $t_i^* \geq 0$. In other words, each agent selects a time instant t_i^* and designs the noise $u_i(t)$ so that the cumulative effect is zero at time t_i^*, while $u_i(t)$ becomes zero afterwards. Let us denote by $t_{\max}$ the maximum of the time instants t_i^*, i.e., the time at which the last cumulative effect reaches zero. Notably, since for all $t > t_{\max}$ the input $\boldsymbol{u}(t)$ is zero, the agents behave as if a standard consensus process was starting at time $t_{\max}$. Moreover, since

$$\boldsymbol{\zeta}^T \dot{\boldsymbol{w}}(t) = -\boldsymbol{\zeta}^T L\boldsymbol{w}(t) + \boldsymbol{\zeta}^T \boldsymbol{u}(t) = \boldsymbol{\zeta}^T \boldsymbol{u}(t),$$

it can be easily shown that

$$\boldsymbol{\zeta}^T \boldsymbol{w}(t) = \boldsymbol{\zeta}^T \boldsymbol{w}(0) + \boldsymbol{\zeta}^T \int_0^{\min\{t, t_{\max}\}} \boldsymbol{u}(\tau)d\tau;$$

hence, for all $t \geq t_{\max}$ it holds $\boldsymbol{\zeta}^T \boldsymbol{w}(t) = \boldsymbol{\zeta}^T \boldsymbol{w}(0)$, i.e., the weighted sum of the initial conditions, with weights encoded by $\boldsymbol{\zeta}$, is preserved for all $t \geq t_{\max}$. This, in turn, implies that the agents reach the desired consensus value, but without disclosing their actual initial conditions. However, the convergence time has a delay equal to $t_{\max}$. Notably, for all $t \geq t_{\max}$ it holds

$$\begin{aligned} w_i(t) &= w_i(0) + \int_0^t \dot{w}_i(\tau)d\tau \\ &= w_i(0) + \sum_{v_j \in \mathcal{N}_i} \int_0^t w_j(\tau)d\tau - d_i \int_0^t w_i(\tau)d\tau \\ &+ \int_0^t u_i(\tau)d\tau. \end{aligned}$$

Hence, since for all $t \geq t_{\max}$ it holds $\int_0^t u_i(\tau) = 0$, we have that for $t \geq t_{\max}$ it holds

$$w_i(0) = w_i(t) + d_i \int_0^t w_i(\tau)d\tau - \sum_{\mathcal{N}_i} \int_0^t w_j(\tau)d\tau.$$

Therefore, an honest but curious node v_ℓ can reconstruct the state of a node v_i provided that it directly observes the state of v_i and all its neighbors, i.e., if v_i is *exposed* to v_ℓ, according to the following definition [11], [27].

Definition 13.0.1

A node v_i is exposed *to a node v_ℓ if $\mathcal{N}_i^{in} \cup \{i\} \subseteq \mathcal{N}_\ell^{in} \cup \{\ell\}$.* □

Therefore, no privacy guarantees can be provided for exposed nodes; notably, most of the approaches available at the state of the art are prone to this vulnerability [11], [20], [313], [316], [319].

13.1 Synchronization in networks of nonlinear continuous-time dynamical systems

This section reviews the synchronization framework developed in [109] and references therein, which will be the basis for the later developments in this chapter. In particular, such a synchronization approach is able to handle time-varying coupling, directed graphs, and the presence of noise. In order to introduce the framework, let us first consider the following technical definition.

Definition 13.1.1

Let a function $\boldsymbol{\psi} : \mathbb{R}^n \times \mathbb{R} \to \mathbb{R}^n$ and a matrix $V \in \mathbb{R}^{n\times n}$ be given. The function $\boldsymbol{\psi}(\boldsymbol{x}, t)$ is said to be V-uniformly decreasing if, for some $c > 0$, we have

$$(\boldsymbol{x} - \boldsymbol{y})^T V(\boldsymbol{\psi}(\boldsymbol{x}, t) - \boldsymbol{\psi}(\boldsymbol{y}, t)) \leq -c \|\boldsymbol{x} - \boldsymbol{y}\|^2 ,$$

for all $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$, and for all $t \geq 0$. □

Let us now consider a dynamical system with the following structure

$$\dot{\boldsymbol{x}}(t) = \underbrace{\begin{bmatrix} \boldsymbol{f}(\boldsymbol{x}_1(t), t) \\ \vdots \\ \boldsymbol{f}(\boldsymbol{x}_n(t), t) \end{bmatrix}}_{\boldsymbol{f}(\boldsymbol{x}(t),t)} + (P \otimes D(t))\boldsymbol{x}(t) + \boldsymbol{u}(t) \tag{13.2}$$

where $\boldsymbol{x}_i(t), \boldsymbol{u}_i(t) \in \mathbb{R}^m$,

$$\boldsymbol{x}(t) = \left[\boldsymbol{x}_1^T(t), \ldots, \boldsymbol{x}_n^T(t)\right]^T \in \mathbb{R}^{nm},$$

and

$$\boldsymbol{u}(t) = \left[\boldsymbol{u}_1^T(t), \ldots, \boldsymbol{u}_n^T(t)\right]^T \in \mathbb{R}^{nm}.$$

Matrix $P \in \mathbb{R}^{n\times n}$ is a zero row-sum matrix that describes the coupling topology, whereas $D(t) \in \mathbb{R}^{m\times m}$ represents the time-varying linear coupling matrix between two systems. The following definition formalizes the idea of synchronization for the above dynamical system.

Definition 13.1.2 (*Synchronization Manifold*)

The synchronization manifold $\mathcal{M}$ is the linear subspace

$$\mathcal{M} = \{\boldsymbol{x} \in \mathbb{R}^{nm} \,|\, \boldsymbol{x}_i = \boldsymbol{x}_j, \forall i, j\}.$$

In particular, Eq. (13.2) is said to synchronize *if $\boldsymbol{x}(t)$ approaches the set $\mathcal{M}$, i.e., if for all i, j it holds*

$$\lim_{t\to\infty} \|\boldsymbol{x}_i - \boldsymbol{x}_j\| = 0.$$

□

Notice that, given the structure of the synchronization manifold $\mathcal{M}$, we have that $\boldsymbol{x} \in \mathcal{M}$ if and only if $\boldsymbol{x} = \mathbf{1} \otimes \boldsymbol{z}$ for some $\boldsymbol{z} \in \mathbb{R}^m$.

The next definition will be useful for proving synchronization of Eq. (13.2).

Definition 13.1.3

Given a real matrix P with nonpositive off-diagonal elements, $\mu(P)$ is the supremum of the set of real numbers μ such that $U(P - \mu I) \succeq 0$ for some $U \in \mathcal{W}_S$. In particular, if P is a Laplacian matrix, then $\mu(P)$ is the minimum over the real parts of the nonzero eigenvalues of P. □

Let us now review some conditions that guarantee the reach of synchronization in Eq. (13.2).

Theorem 13.1.1

Consider the dynamics in Eq. (13.2) with P being an $n \times n$ matrix with nonpositive off-diagonal elements. Moreover, let $V \in \mathbb{R}^{n \times n}$ be a given symmetric and positive definite matrix and suppose that the following conditions are satisfied:

1. $\lim_{t \to \infty} \|\boldsymbol{u}_i - \boldsymbol{u}_j\| = 0$ *for all* i, j,
2. $VD(t) = D^T(t)V \preceq 0$ *for all* t,
3. $\boldsymbol{f}(\boldsymbol{y}, t) + \mu(P)D(t)\boldsymbol{y}$ *is* V*-uniformly decreasing.*

Then, Eq. (13.2) synchronizes. □

As observed in [109], the condition that a function $\boldsymbol{\psi}(\boldsymbol{y}, t)$ is V-uniformly decreasing for a symmetric and positive definite V implies that $\dot{\boldsymbol{y}} = \boldsymbol{\psi}(\boldsymbol{y}, t)$ is globally asymptotically stable. In particular, when $\boldsymbol{\psi}(\boldsymbol{y}, t) = \boldsymbol{f}(\boldsymbol{y}, t) + K(t)\boldsymbol{y}$, we have that $K(t)\boldsymbol{y}$ is a term that globally stabilizes $\boldsymbol{f}(\boldsymbol{y}, t)$. A classical example in this sense is given by the Chua oscillator, for which $\boldsymbol{y} \in \mathbb{R}^3$ and

$$\boldsymbol{f}(\boldsymbol{y}, t) = \begin{bmatrix} c\alpha(y_2 - y_1 - \theta(y_1)) \\ c(y_1 - y_2 + y_3) \\ -c(\beta y_2 + \gamma y_3) \end{bmatrix},$$

with

$$\theta(y_1) = by_1 + \frac{1}{2}(a - b)\{|y_1 + 1| - |y_1 - 1|\}.$$

Notably, the Chua oscillator is known to synchronize [108] for $K(t) = \mu(P)D(t)$ with

$$V = I_3, \quad D(t) = -\begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \tag{13.3}$$

and for a sufficiently large $\mu(P) > 0$.

However, it might be challenging in general to identify $\boldsymbol{f}(\boldsymbol{y}, t)$, V, P and $D(t)$ that guarantee satisfaction of the conditions in Theorem 13.1.1. For this reason, in the next result we provide a constructive way to identify a V-uniformly decreasing function.

Theorem 13.1.2

Let a matrix $V \in \mathbb{R}^{n \times n}$ be given. A function $\boldsymbol{\psi} : \mathbb{R}^n \times \mathbb{R} \to \mathbb{R}^n$ is V-uniformly decreasing if and only if there is a function $g : \mathbb{R}^n \times \mathbb{R} \to \mathbb{R}$ such that

$$V\boldsymbol{\psi}(\boldsymbol{x}, t) = -\nabla_{\boldsymbol{x}} g(\boldsymbol{x}, t) \tag{13.4}$$

with $g(\boldsymbol{x}, t)$ being strongly convex in $\boldsymbol{x}$ for all $t \geq 0$. □

Proof. ($\Rightarrow$) In order to prove sufficiency we observe that if $g(\boldsymbol{x}, t)$ is strongly convex in $\boldsymbol{x}$ then it satisfies Eq. (1.1) for some $\mu > 0$. Therefore, if V and $\boldsymbol{\psi}(\boldsymbol{x}, t)$ satisfy Eq. (13.4), we have that

$$-(\boldsymbol{x} - \boldsymbol{y})^T V(\boldsymbol{\psi}(\boldsymbol{x}) - \boldsymbol{\psi}(\boldsymbol{y})) \geq \mu \|\boldsymbol{x} - \boldsymbol{y}\|^2$$

and thus

$$(\boldsymbol{x}-\boldsymbol{y})^T V(\boldsymbol{\psi}(\boldsymbol{x})-\boldsymbol{\psi}(\boldsymbol{y})) \leq -\mu\|\boldsymbol{x}-\boldsymbol{y}\|^2,$$

i.e., $\boldsymbol{\psi}(\cdot)$ is V-uniformly decreasing with $c=\mu$.
($\Leftarrow$) In order to prove necessity, suppose $\boldsymbol{\psi}(\cdot)$ is V-uniformly decreasing; then, by construction it holds

$$(\boldsymbol{x}-\boldsymbol{y})^T V(\boldsymbol{\psi}(\boldsymbol{x})-\boldsymbol{\psi}(\boldsymbol{y})) \leq -c\|\boldsymbol{x}-\boldsymbol{y}\|^2,$$

for some $c>0$. At this point, let us consider any continuous curve γ in $\mathbb{R}^n$ that starts at $\boldsymbol{x}$ and ends at $\boldsymbol{y}$. By the *gradient theorem* (e.g., see [320]), we have that there is a function $r(\boldsymbol{x},t)$, differentiable in $\boldsymbol{x}$, with

$$\nabla_{\boldsymbol{x}} r(\boldsymbol{x},t) = V\boldsymbol{\psi}(\boldsymbol{x},t)$$

that satisfies

$$\int_\gamma V\boldsymbol{\psi}(\boldsymbol{z},t)d\boldsymbol{z} = r(\boldsymbol{x},t)-r(\boldsymbol{y},t).$$

As a consequence, we have that $g(\boldsymbol{x},t)=-r(\boldsymbol{x},t)$ satisfies

$$(\nabla_{\boldsymbol{x}} g(\boldsymbol{x},t)-\nabla_{\boldsymbol{y}} g(\boldsymbol{y},t))^T(\boldsymbol{x}-\boldsymbol{y}) \geq c\|\boldsymbol{x}-\boldsymbol{y}\|^2,$$

and is thus strongly convex for $\mu=c$. This completes our proof. □

The above result allows designing a dynamics that guarantees synchronization by first identifying $g(\cdot)$ and then by computing its antigradient.

A few remarks are now in order.

Remark 13.1.1

Let us consider

$$\boldsymbol{\psi}(\boldsymbol{y},t) = \boldsymbol{f}(\boldsymbol{y},t) + K(t)\boldsymbol{y}$$

with $VK(t)$ symmetric and negative definite, and suppose

$$V\boldsymbol{f}(\boldsymbol{y},t) = -\nabla_{\boldsymbol{y}} g(\boldsymbol{y},t)$$

for some function $g(\boldsymbol{y},t)$. Since $VK(t)\boldsymbol{y}$ is the antigradient along $\boldsymbol{y}$ of the convex function

$$w(\boldsymbol{y},t) = -\frac{1}{2}\boldsymbol{y}^T VK(t)\boldsymbol{y},$$

according to the above theorem, synchronization is achieved if and only if $g(\boldsymbol{y},t)+w(\boldsymbol{y},t)$ is strongly convex. Therefore, $VK(t)\boldsymbol{y}$ can be regarded as the antigradient of a convexification term that is added to $g(\boldsymbol{y},t)$ in order to make it strongly convex, so that the dynamics $\dot{\boldsymbol{y}}=\boldsymbol{\psi}(\boldsymbol{y},t)$ essentially amounts to a gradient descent-like dynamics. □

Remark 13.1.2

A simple way to choose a strongly convex function $g(\boldsymbol{x},t)$ is to consider any (not necessarily strongly) convex function $g'(\boldsymbol{x},t)$ and add a strongly convex function, e.g.,

$$g(\boldsymbol{x},t) = g'(\boldsymbol{x},t) + \frac{1}{2}\boldsymbol{x}^T H\boldsymbol{x},$$

where H is symmetric and positive definite. In this case, we have that

$$\boldsymbol{h}(\boldsymbol{x},t) = -\nabla_{\boldsymbol{x}} g'(\boldsymbol{x},t) - H\boldsymbol{x}.$$

□

13.2 Proposed approach

Let us consider a scenario where a network of n agents is interconnected by a directed and strongly connected graph $G = \{V, E\}$, and each agent is provided with a hidden scalar[1] initial value $h_i(0)$, while

$$\boldsymbol{h}(0) = \big[h_1(0), \dots, h_n(0)\big]^T \in \mathbb{R}^n$$

denotes the stack of the hidden values for all agents. The purpose of each agent is to reach the consensus value of the initial conditions as in Eq. (11.4), without disclosing the value of those conditions to their *honest but curious* neighbors. In doing so, it is also intended to provide an additional layer of security against eavesdropping, by ensuring that messages transmitted between legitimate agents undergo a distortion capable of hiding their content from malicious entities.

To accomplish these tasks, let us consider that each agent maintains locally a state vector $\boldsymbol{x}_i(t) \in \mathbb{R}^m$ and interacts with neighboring agents by implementing the synchronization scheme discussed in Eq. (13.2). Further to that, let us consider a matrix $Q \in \mathbb{R}^{n \times nm}$ with the following structure

$$Q = I_n \otimes \boldsymbol{q}^T,$$

where $\boldsymbol{q} \in \mathbb{R}^m$ and $\boldsymbol{q} \neq \boldsymbol{0}_m$.

Notice that, by construction, for any $\boldsymbol{c} \in \mathbb{R}^m$, matrix Q satisfies

$$Q\left(\boldsymbol{1}_n \otimes \boldsymbol{c}\right) = \left(\boldsymbol{q}^T \boldsymbol{c}\right) \boldsymbol{1}_n. \tag{13.5}$$

Moreover, let us assume that each agent maintains an additional scalar variable $z_i(t)$ whose dynamics, in stacked form for all agents (i.e., considering $\boldsymbol{z}(t) = \begin{bmatrix} z_1(t) & \dots z_n(t) \end{bmatrix}^T$), reads as follows

$$\dot{\boldsymbol{z}}(t) = -L\boldsymbol{z}(t) + Q\left(\boldsymbol{f}(\boldsymbol{x}(t), t) + (P \otimes D(t))\boldsymbol{x}(t) + \boldsymbol{u}(t)\right),$$

where the initial conditions $\boldsymbol{x}(0)$ for the synchronizing dynamics and $\boldsymbol{z}(0)$ for the additional dynamics are chosen so that

$$h_i(0) = z_i(0) - \boldsymbol{q}^T \boldsymbol{x}_i(0),$$

which, in stacked form, reads

$$\boldsymbol{h}(0) = \boldsymbol{z}(0) - Q\boldsymbol{x}(0).$$

Finally, we assume that each agent locally computes

$$\boldsymbol{h}(t) = \boldsymbol{z}(t) - Q\boldsymbol{x}(t).$$

Overall, the agents' dynamics can be summarized as follows

$$\begin{cases} \dot{\boldsymbol{x}}(t) = \boldsymbol{f}(\boldsymbol{x}(t), t) + (P \otimes D(t))\boldsymbol{x} + \boldsymbol{u}(t) \\ \dot{\boldsymbol{z}}(t) = -L\boldsymbol{z}(t) + Q\left(\boldsymbol{f}(\boldsymbol{x}(t), t) + (P \otimes D(t))\boldsymbol{x}(t) + \boldsymbol{u}(t)\right) \\ \boldsymbol{h}(t) = \boldsymbol{z}(t) - Q\boldsymbol{x}(t) \end{cases} \tag{13.6}$$

where

$$\boldsymbol{u}(t) = [\boldsymbol{u}_1^T(t), \dots, \boldsymbol{u}_n^T(t)]^T \in \mathbb{R}^{nm}$$

and $\boldsymbol{u}_i(t) \in \mathbb{R}^m$ is a time-varying input locally selected by agent i, such that the condition 1) of Theorem 13.1.1 holds. For simplicity, we assume that $P = \eta L$, where L is the

[1] Note that all the results presented can be trivially generalized to the vectorial case.

Laplacian matrix corresponding to the agents' interaction graph, and $\eta > 0$ is a suitable gain.

Notice that the dynamics of the state vector $\boldsymbol{z}(t)$, dependent on the exchange of $z_i(t)$ with neighboring agents, is distorted by the nonlinear term featuring $\boldsymbol{x}(t)$, which influences the consensus dynamics via the matrix Q. Moreover, for the nonlinear dynamics of the state vector $\boldsymbol{x}(t)$ the agents need to exchange only a subset of the state variables governed by the time-varying matrix $D(t)$. The dynamics of the hidden consensus vector $\boldsymbol{h}(t)$, on the other hand, is calculated only locally and is never exchanged.

Once the nonlinear dynamics have reached synchronization, thanks to the structure proposed in Eq. (13.6), it will be possible for legitimate agents to trace the consensus value $\boldsymbol{h}(t)$, while maintaining security on the exchanged messages and privacy on the initial conditions. In order to achieve this, we need the following assumption.

Assumption 13.2.1

The agents' interaction topology is such that the set of the exposed nodes is empty. □

In the remainder of this section, we aim to study the convergence of the hidden state variable $\boldsymbol{h}(t)$. In the next section, we analyze possible strategies that the attacker may deploy in an attempt to breach privacy and security.

The following theorem shows that $h_i(t)$ approaches the weighted average of the initial hidden values for all agents.

Theorem 13.2.1

Let us consider a network of n agents interconnected by a directed and strongly connected graph $G = \{V, E\}$ and interacting according to the dynamics in Eq. (13.6). Moreover, let us assume that the dynamics for $\boldsymbol{x}(t)$ satisfies Theorem 13.1.1 (and thus synchronizes). Then, it holds

$$\lim_{t\to\infty} \boldsymbol{h}(t) = \mathbf{1}_n \boldsymbol{\zeta}^T \boldsymbol{h}_0,$$

i.e., the local outputs $h_i(t)$ asymptotically converge to a weighted average of the hidden initial states $\boldsymbol{h}(0)$, considering the weights encoded by $\boldsymbol{\zeta}$. □

Proof. In order to prove the result, we observe that

$$\begin{aligned}
\boldsymbol{z}(t) &= e^{-Lt}\boldsymbol{z}(0) + \int_0^t e^{-L(t-\tau)} Q\dot{\boldsymbol{x}}(\tau)\, d\tau \\
&= e^{-Lt}\boldsymbol{z}(0) + e^{-Lt}\int_0^t e^{L\tau} Q\dot{\boldsymbol{x}}(\tau)\, d\tau \\
&= e^{-Lt}\boldsymbol{z}(0) + e^{-Lt}(e^{Lt}Q\boldsymbol{x}(t) - e^{L0}Q\boldsymbol{x}(0)) \\
&- e^{-Lt}\int_0^t Le^{L\tau}Q\boldsymbol{x}(\tau)\, d\tau \\
&= e^{-Lt}(\boldsymbol{z}(0) - Q\boldsymbol{x}(0)) + Q\boldsymbol{x}(t) \\
&- \int_0^t \underbrace{e^{-Lt}Le^{L\tau}Q\boldsymbol{x}(\tau)}_{s(t,\tau)}\, d\tau.
\end{aligned} \tag{13.7}$$

At this point, we claim that

$$\lim_{t\to\infty} \int_0^t s(t,\tau) = \mathbf{0}_n.$$

In fact we observe that, when t approaches infinity but τ is finite, we have that e^{-Lt} approaches $\mathbf{1}_n\boldsymbol{\zeta}^T$ and thus $e^{-Lt}L$ approaches $\mathbf{1}_n\boldsymbol{\zeta}^T L$, which is zero; since for finite τ, by construction, $e^{L\tau}Q\boldsymbol{x}(\tau)$ is bounded, we conclude that $\boldsymbol{s}(t,\tau)$ approaches zero. If, conversely, both t and τ approach infinity we have that (i) $\boldsymbol{x}(t)$ reaches synchronization

by Theorem 13.1.1, (ii) $Q\boldsymbol{x}(\tau)$ approaches a vector $\alpha(\tau)\mathbf{1}_n$ for some $\alpha(\tau)$ by Eq. (13.5), and (iii) $e^{L\tau}\mathbf{1}_n = \mathbf{1}_n$ by construction. Thus, we have that $\boldsymbol{s}(t,\tau)$ approaches

$$\mathbf{1}_n\boldsymbol{\zeta}^T L e^{L\tau}\alpha(\tau)\mathbf{1}_n = \alpha(\tau)\mathbf{1}_n\boldsymbol{\zeta}^T L\mathbf{1}_n = \mathbf{0}_n$$

and, also in this latter case, $\boldsymbol{s}(t,\tau)$ approaches zero. In conclusion, in the limit of t approaching infinity, $\boldsymbol{s}(t,\tau)$ is zero for all finite values of τ and when τ approaches infinity. This implies that

$$\lim_{t\to\infty}\int_0^t \boldsymbol{s}(t,\tau)\,d\tau = \mathbf{0}_n$$

and our claim is verified.

As a consequence, we have that

$$\begin{aligned}\lim_{t\to\infty}\boldsymbol{z}(t) - Q\boldsymbol{x}(t) &= \lim_{t\to\infty} e^{-Lt}\left(\boldsymbol{z}(0) - Q\boldsymbol{x}(0)\right)\\ &= \lim_{t\to\infty} e^{-Lt}\boldsymbol{h}(0)\\ &= \boldsymbol{\zeta}_n^T\boldsymbol{h}(0)\mathbf{1}_n\end{aligned}$$

and thus the entries of $\boldsymbol{h}(t) = \boldsymbol{z}(t) - Q\boldsymbol{x}(t)$ converge to the weighted average of $\boldsymbol{h}(0)$, considering the weights encoded by $\boldsymbol{\zeta}$. This completes our proof. □

13.3 Security and Privacy

13.3.1 Security Analysis

In this section, we discuss the limitations of possible strategies that can be implemented by a malicious eavesdropper. Specifically, we assume that a malicious entity aims to identify the consensus value, without making changes to it and without disclosing its presence; this behavior is captured by the following definition.

Definition 13.3.1 (*Security*)

A consensus algorithm is said to be secure, *if a third party (or eavesdropper) able to access all messages exchanged over the network cannot reconstruct the state of the legitimate agents without being identified or actively modifying the consensus value.* □

It is worth remembering that the slightest injection of noise in the network is capable of causing a permanent deviation from the consensus value. Therefore, an attack with an injection of false data would potentially disrupt synchronization and the achievement of consensus, which is out of the scope of our attack model as this would make it impossible for the attacker to get to know the consensus value.

Let us now investigate some possible strategies that the attacker can put into practice. To this end, we want to place ourselves in the worst case in which the attacker has partial knowledge of some piece of global information about the network. In particular, from Eq. (13.6), for the hidden consensus value we have

$$\boldsymbol{h}(t) = \boldsymbol{z}(t) - Q\boldsymbol{x}(t)$$

in which we notice a dependence on the state variables $\boldsymbol{x}(t)$ and $\boldsymbol{z}(t)$. Since the state vector $\boldsymbol{z}(t)$ is directly exchanged at each step, a malicious agent could wait for all legitimate agents to converge on the consensus and read this value on one of the communication channels. On the other hand, since $\boldsymbol{x}(t)$ it is never directly exchanged but undergoes a distortion due to the product with $(P \otimes D(t))$ and the addition of the noise $u_i(t)$, the only way the malicious agent would have to reconstruct this state vector would be to locally implement a system identical to the one relating to nonlinear dynamics of $\boldsymbol{x}(t)$, in the hope of synchronizing with the other agents as if it were a receiving slave system. This, however, implies the strong assumption that the malicious agent knows both the

model of nonlinear dynamics $f(\boldsymbol{x}(t), t)$ and the value to which it must tend its local input noise in order for hypothesis 1) of Theorem 13.1.1 to be verified.

Moreover, even if we admit that the eavesdropper has such knowledge, in order to be able to know the hidden consensus value $h(t)$, it should also be able to estimate the value of the transformation matrix Q, which is only applied locally to each agent.

Let us assume an eavesdropper node is able to locally synchronize its nonlinear dynamics to the others in the network. Then, the eavesdropper is not able to reconstruct the consensus value unless it has global knowledge of the network structure and without introducing errors. In fact, one possible strategy would be to start from the dynamics of $\boldsymbol{z}(t)$, i.e., the variable that is exchanged by each agent

$$\begin{aligned}\dot{\boldsymbol{z}}(t) =& - L\boldsymbol{z}(t) + Q\underbrace{(\boldsymbol{f}(\boldsymbol{x}(t), t) + (P \otimes D(t))\boldsymbol{x}(t) + \boldsymbol{u}(t))}_{\dot{\boldsymbol{x}}(t)} \\ =& - L\boldsymbol{z}(t) + Q\dot{\boldsymbol{x}}(t).\end{aligned} \tag{13.8}$$

Interestingly, by premultiplying both sides of the above equation by $\boldsymbol{\zeta}^T$, we obtain

$$\boldsymbol{\zeta}^T \dot{\boldsymbol{z}}(t) = -\underbrace{\boldsymbol{\zeta}^T L\boldsymbol{z}(t)}_{\mathbf{0}_n} + \boldsymbol{\zeta}^T Q\dot{\boldsymbol{x}}(t) = \boldsymbol{\zeta}^T \left(\sum_{i=1}^{n} \boldsymbol{q}^T \dot{\boldsymbol{x}}_i(t) \right). \tag{13.9}$$

At this point, considering the vectorization of both of the terms of Eq. (13.10) and noting that $(\boldsymbol{q}^T) = \boldsymbol{q}$, we obtain

$$\boldsymbol{\zeta}^T \dot{\boldsymbol{z}}(t) = \sum_{i=1}^{n} \left(\dot{\boldsymbol{x}}_i(t)^T \otimes \boldsymbol{\zeta}^T \right) \boldsymbol{q}. \tag{13.10}$$

Therefore, in order to derive q, an attacker will need at least m linearly independent observations of $\dot{\boldsymbol{x}}(t)$ and m observations of $\dot{\boldsymbol{z}}(t)$. Specifically, it must be able to compute

$$\underbrace{\begin{bmatrix} \sum_{i=1}^{n} \left(\dot{\boldsymbol{x}}_i(t_1)^T \otimes \boldsymbol{\zeta}^T \right) \\ \vdots \\ \sum_{i=1}^{n} \left(\dot{\boldsymbol{x}}_i(t_m)^T \otimes \boldsymbol{\zeta}^T \right) \end{bmatrix}}_{\dot{\mathcal{X}}} \boldsymbol{q} = \underbrace{\begin{bmatrix} \boldsymbol{\zeta}^T \dot{\boldsymbol{z}}(t_1) \\ \vdots \\ \boldsymbol{\zeta}^T \dot{\boldsymbol{z}}(t_m) \end{bmatrix}}_{\dot{\mathcal{Z}}}, \tag{13.11}$$

from which, assuming $\dot{\mathcal{X}}$ is nonsingular,

$$\boldsymbol{q} = \dot{\mathcal{X}}^{-1} \dot{\mathcal{Z}}. \tag{13.12}$$

Thus, in order to be able to estimate $\boldsymbol{q}$, the malicious entity would need to know $\boldsymbol{\zeta}$, which requires global knowledge on the graph topology. Moreover, while the agents exchange $z_i(t)$, they do not directly exchange $\dot{z}_i(t)$; estimating $\dot{z}_i(t)$ would require numerical differentiation of $z_i(t)$, thus introducing errors.

13.3.2 Privacy Analysis

In this section, we aim to evaluate a possible strategy for the violation of privacy on the initial conditions of legitimate nodes and show its failure in the event that Assumption 13.2.1 holds. The concept of privacy preservation to which we refer is expressed by the following definition.

Definition 13.3.2 (*Privacy*)

A consensus algorithm is said to be private *if each node involved is unable to reconstruct the initial conditions of its neighbors, or of any other node.* □

Notice that the most interesting aspect of the proposed approach is that the agents exchange $z_i(t)$ and $D(t)\boldsymbol{x}_i(t)$, but locally compute $h_i(t)$, which, under mild hypotheses, is a different combination of the state variables, and is thus not directly shared.

Let us assume an honest but curious node ℓ is an out neighbor of a target node i. Then, node ℓ is not able to retrieve the hidden initial value of i if Assumption 13.2.1 holds true. In fact, we observe that

$$\begin{aligned}\boldsymbol{z}(t) &= \boldsymbol{z}(0) + \int_0^t \dot{\boldsymbol{z}}(\tau)d\tau \\ &= \boldsymbol{z}(0) - L\int_0^t \boldsymbol{z}(\tau)d\tau + Q\boldsymbol{x}(t) - Q\boldsymbol{x}(0).\end{aligned}$$

Therefore, considering the i-th agent and defining $\boldsymbol{e}_i$ as the i-th vector in the canonical basis of $\mathbb{R}^n$, we have that

$$\begin{aligned}z_i(t) &= z_i(0) + \int_0^t \sum_{j\in\mathcal{N}_i^{\text{in}}} (z_j(\tau) - z_i(\tau))\, d\tau \\ &\quad + \left(\boldsymbol{e}_i^T \otimes \boldsymbol{q}^T\right)(\boldsymbol{x}(t) - \boldsymbol{x}(0)) \\ &= z_i(0) + \int_0^t \sum_{j\in\mathcal{N}_i^{\text{in}}} (z_j(\tau) - z_i(\tau))\, d\tau \\ &\quad + \boldsymbol{q}^T(\boldsymbol{x}_i(t) - \boldsymbol{x}_i(0)) \\ &= z_i(0) - d_i\int_0^t z_i(\tau)d\tau + \sum_{j\in\mathcal{N}_i^{\text{in}}}\int_0^t z_j(\tau)d\tau \\ &\quad + \boldsymbol{q}^T(\boldsymbol{x}_i(t) - \boldsymbol{x}_i(0)),\end{aligned}$$

since $h_i(0) = z_i(0) - \boldsymbol{q}^T\boldsymbol{x}_i(0)$, we have that

$$h_i(0) = z_i(t) + d_i\int_0^t z_i(\tau)d\tau - \sum_{j\in\mathcal{N}_i^{\text{in}}}\int_0^t z_j(\tau)d\tau - \boldsymbol{q}^T\boldsymbol{x}_i(t). \tag{13.13}$$

Now we can observe from Eq. (13.13) that in order for the curious node to reconstruct $h_i(0)$ according to the above scheme, it must be able to compute

$$\sum_{j\in\mathcal{N}_i^{\text{in}}}\int_0^t z_j(\tau)d\tau, \quad \forall\, v_j \in \mathcal{N}_i^{\text{in}} \cup \{v_i\},$$

i.e., node i must be exposed to the curious node thus violating Assumption 13.2.1.

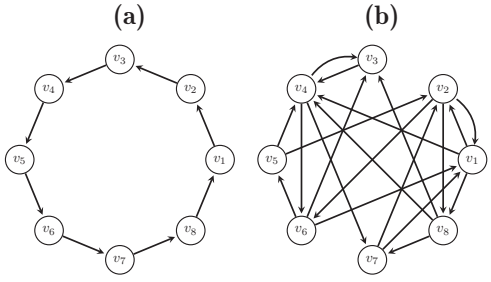


Figure 13.1: (a) Balanced ring graph topology with 8 agents. (b) Unbalanced graph topology with 8 agents.

13.4 Simulations

In this section, we aim to demonstrate the effectiveness of the proposed approach via a simulation scenario. In particular, we choose as a case study for the nonlinear function the dynamics of the chaotic Chua oscillator, setting $\boldsymbol{f}(\boldsymbol{x}(t), t)$, $D(t)$ and V as in Eq. (13.3), and we consider two directed network topologies that satisfy Assumption 13.2.1. The first one is a directed ring graph $G_1 = \{V_1, E_1\}$ with $|V_1| = 8$ nodes and $|E_1| = 8$ directed edges (reported in Figure 13.1a), while the second one is a random topology $G_2 = \{V_2, E_2\}$ with $|V_2| = 8$ nodes and $|E_2| = 20$ directed edges (reported in Figure 13.1b). Notably, G_2 is *unbalanced*, i.e., the in- and out-degrees do not coincide for all nodes.

For both topologies, the agents aim to cooperatively achieve the average of their hidden initial conditions

$$\begin{aligned}\boldsymbol{h}_0 = \big[&-32.37 \quad -10.10 \quad -27.28 \quad -24.67 \\ &-13.96 \quad -23.94 \quad -21.08 \quad -20.14\big]^T\end{aligned}$$

without disclosing them to each other. To this end, each agent randomly selects a set of initial conditions for the Chua oscillator variables

$$\boldsymbol{x}_{10} = \big[2.60 \quad 7.37 \;\; 7.97 \;\; 0.26 \;\; 4.97 \;\; 5.51 \;\; 3.13 \;\; 6.58\big]^T,$$

$$\boldsymbol{x}_{20} = \big[3.87 \;\; 3.79 \;\; 2.69 \;\; 6.62 \;\; 4.15 \;\; 9.73 \;\; 0.42 \;\; 2.15\big]^T,$$

$$\boldsymbol{x}_{30} = \big[8.32 \;\; 0.13 \;\; 5.83 \;\; 3.88 \;\; 3.51 \;\; 1.13 \;\; 7.38 \;\; 4.17\big]^T.$$

Then, to also ensure the security of the approach, agents select a common vector[2]

$$\boldsymbol{q} = \big[1 \quad 2 \quad 3\big]^T$$

and each agent i calculates the initial values for the shared variables as

$$z_{0i} = h_{0i} + \boldsymbol{q}^T \boldsymbol{x}_{0i},$$

thus obtaining

$$\boldsymbol{z}_0 = \big[2.93 \quad 5.24 \;\; 3.57 \;\; 0.46 \;\; 9.83 \;\; 4.41 \;\; 5.04 \;\; 3.24\big]^T.$$

Notice that, in the simulations, the parameters of the Chua oscillators are set as follows:

$$\alpha = 10,\ \beta = 15,\ \gamma = 0.0385,\ a = -1.5,\ b = -0.9,\ c = 3,$$

while each component of the input noise $\boldsymbol{u}_i(t) \in \mathbb{R}^3$ is initially set as $u_{ij}(t) = \sin(10t + \phi_{ij})$, where ϕ_{ij} is chosen uniformly at random in $[0, 2\pi]$, and then becomes a constant value $u_{ij}(t) = 2$ for $t > 2[s]$. Finally, we assume the matrix $P = \eta L$ with $\eta = 100$.

Considering the ring topology (in this case $\boldsymbol{\zeta} = \mathbf{1}_n/n$, where $n = 8$), we can observe from Figure 13.2 that, as time approaches infinity, each $h_i(t)$ approaches the average of the entries of the initial condition vector $\boldsymbol{h}_0$, i.e., the value -21.69. On the other hand, what circulates in the network and is seen from an external eavesdropper is equal to a chaotically oscillating synchronized variable (the variable $\boldsymbol{z}(t)$ in Figure 13.3). Figure 13.4 shows the trend of the three state variables of the Chua oscillators at each agent, which are shown to synchronize.

Considering the random topology, since the graph is unbalanced, the consensus hidden variable will tend to a weighted average according to the dominant left eigenvector ζ. In particular, as shown in Figure 13.5, we have that $\lim_{t\to\infty} h_i(t) = -23.76$. Also in this case, the dynamics of the exchanged variables $\boldsymbol{z}(t)$ remain oscillatory and synchronized (see Figure 13.6), and the state variables of the oscillator $\boldsymbol{x}(t)$ also synchronize (see Figure 13.7).

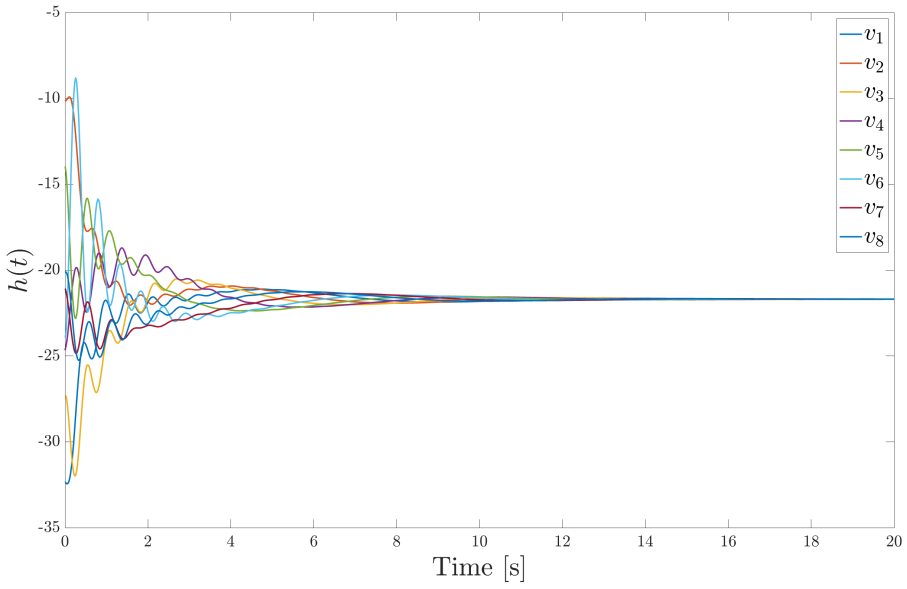


Figure 13.2: Temporal evolution of the local state variable $\boldsymbol{h}(t)$ carrying the hidden consensus value for a network of Chua oscillators arranged in a ring topology.

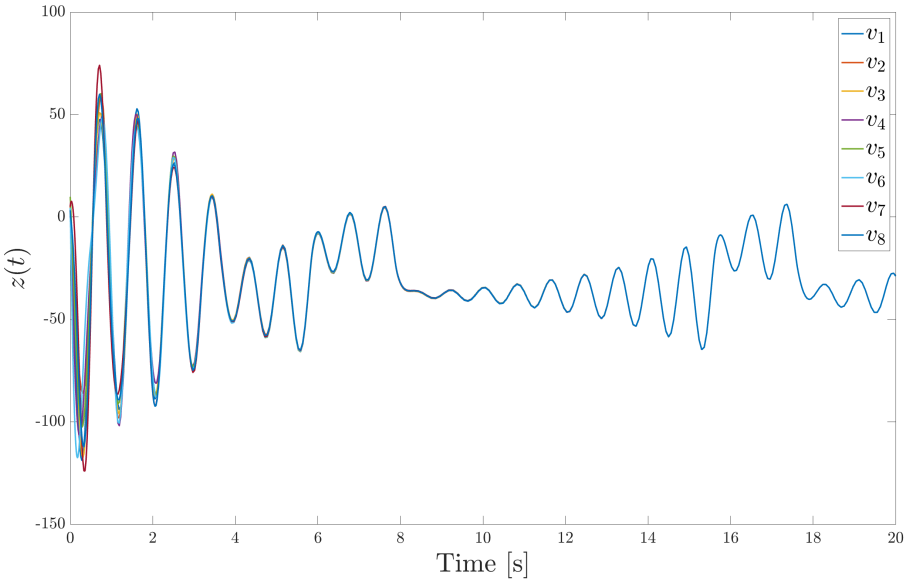


Figure 13.3: Temporal evolution of the shared state variable $\boldsymbol{z}(t)$ for a network of Chua oscillators arranged in a ring topology.

To conclude the section, let us show how the results in Theorem 13.1.2 can be used to guide the design of the synchronization dynamics $\boldsymbol{f}(\boldsymbol{x}(t), t)$.

Specifically, let us consider $n = 6$ agents, and let us assume that the synchronization dynamics at each agent features $m = 3$ state variables. Moreover, for all $t \geq 0$, let us

[2]Note that the sharing of this information over the network can be done securely in an initialization phase in which this vector is sent encrypted.

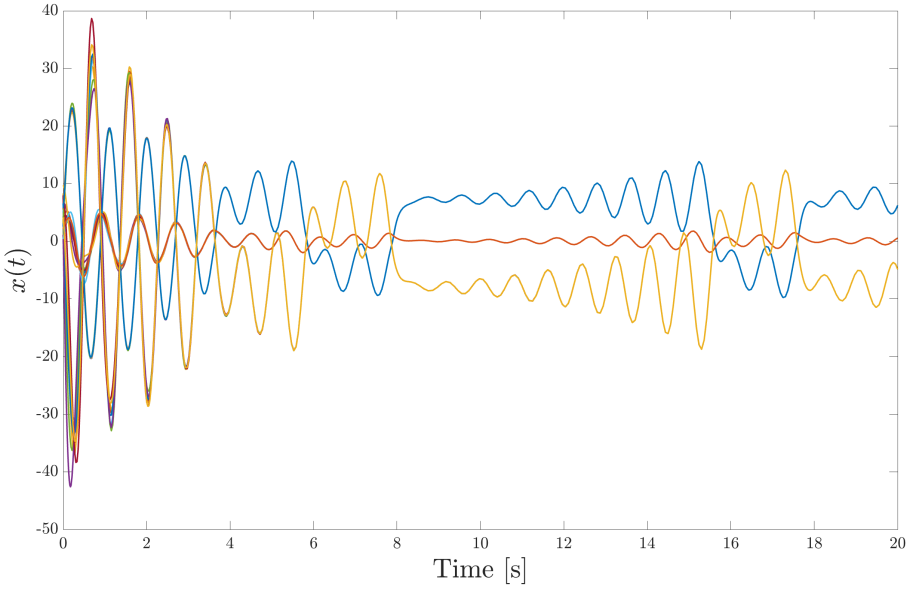


Figure 13.4: Temporal evolution of the state dynamics $\boldsymbol{x}(t)$ for a network of Chua oscillators arranged in a ring topology.

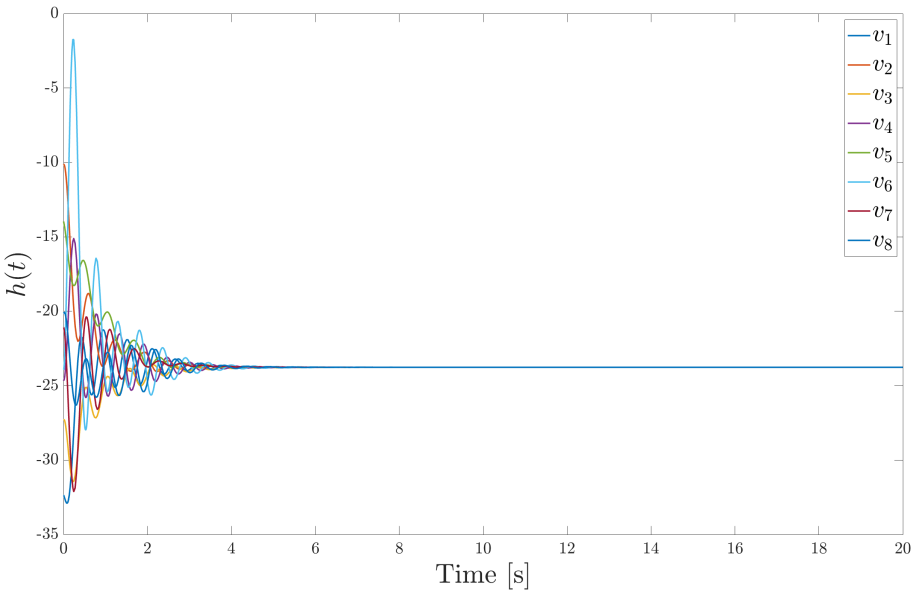


Figure 13.5: Temporal evolution of the local state variable $\boldsymbol{h}(t)$ carrying the hidden consensus value for a network of Chua oscillators arranged in an unbalanced random topology.

consider the $n \times n$ matrix $S(t) = (20 + 5sin(20t))\mathcal{S}$ with

$$\mathcal{S} = \begin{bmatrix} 12 & 0.5 & 5 \\ 0.2 & 3 & 0.3 \\ 3 & 0.1 & 20 \end{bmatrix}.$$

Notably, for all $t \geq 0$, the eigenvalues of $S(t)$ are lower bounded by $\delta \approx 44.85$; therefore,

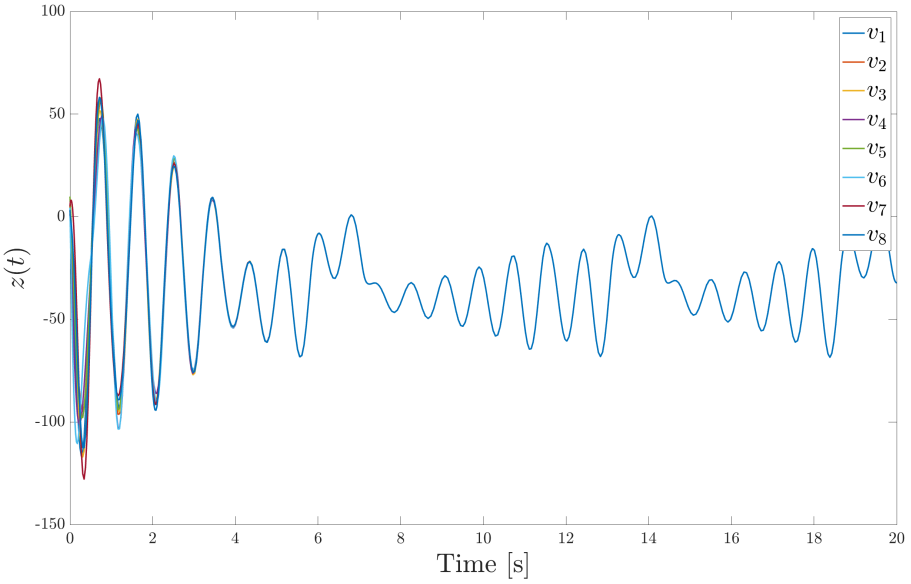


Figure 13.6: Temporal evolution of the shared state variable $\boldsymbol{z}(t)$ in the case of Chua oscillator considering the unbalanced random graph.

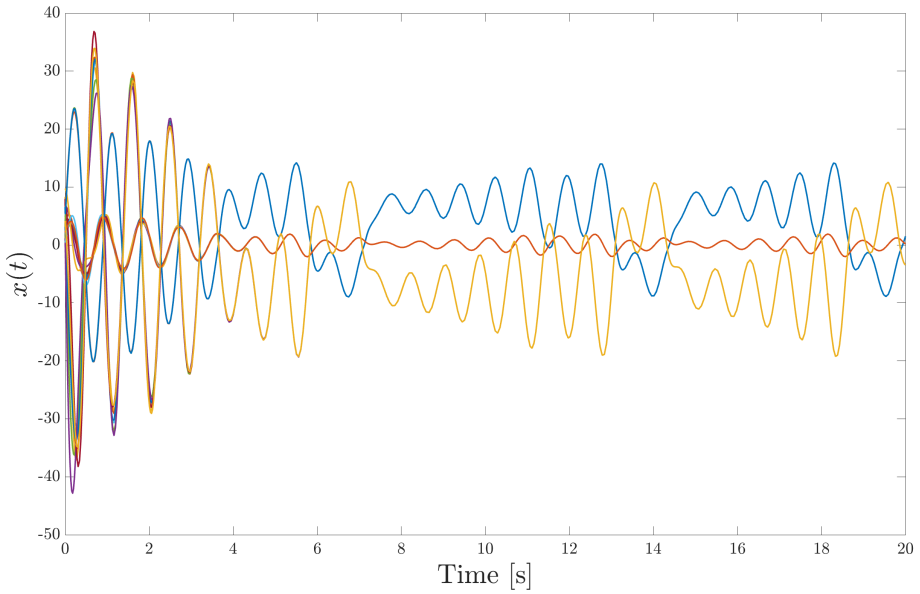


Figure 13.7: Temporal evolution of the dynamics $\boldsymbol{x}(t)$ in the case of Chua oscillator considering the unbalanced random graph.

the quadratic time-varying function

$$g(\boldsymbol{x}, t) = \frac{1}{2}\boldsymbol{x}^T S(t)\boldsymbol{x} + \boldsymbol{\omega}^T(t)\boldsymbol{x}$$

is strongly convex. As a consequence, according to Theorem 13.1.2,

$$\boldsymbol{f}(\boldsymbol{x}, t) = -\nabla_{\boldsymbol{x}} g(\boldsymbol{x}, t) = -S(t)\boldsymbol{x} - \boldsymbol{\omega}(t) \tag{13.14}$$

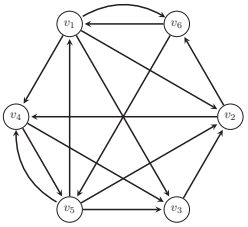


Figure 13.8: Unbalanced graph topology with 6 agents.

is V-uniformly decreasing with $V = I_n$. In this example, we set

$$\boldsymbol{\omega}(t) = 10e^{2\boldsymbol{x}(t)}sin(10t).$$

and we consider the directed unbalanced topology $G_3 = \{V_3, E_3\}$ in Figure 13.8 with $|V_3| = 6$ nodes and $|E_3| = 15$ directed edges.

The initial conditions selected by the agents are

$$\boldsymbol{h}_0 = \begin{bmatrix} 4.70 & 5.20 & -0.42 & 5.60 & 7.57 & 4.06 \end{bmatrix}^T,$$

$$\boldsymbol{x}_{10} = \begin{bmatrix} 0.28 & 0.45 & 0.29 & 0.68 & 0.49 & 0.15 \end{bmatrix}^T,$$

$$\boldsymbol{x}_{20} = \begin{bmatrix} 0.79 & 0.90 & 0.13 & 0.21 & 0.05 & 0.59 \end{bmatrix}^T,$$

$$\boldsymbol{x}_{30} = \begin{bmatrix} 0.10 & 0.29 & 0.02 & 0.27 & 0.57 & 0.70 \end{bmatrix}^T,$$

$$\boldsymbol{z}_0 = \begin{bmatrix} 6.87 & 8.35 & 0.18 & 7.5 & 9.89 & 7.48 \end{bmatrix}^T.$$

In this example, we also consider a time-varying matrix $D(t)$, which assumes three possible values over time. Specifically, we assume that

$$D(t) = \begin{bmatrix} -9 & 1 & 0 \\ 1 & -5 & 3 \\ 0 & 3 & -10 \end{bmatrix} \quad \text{for } t < 2[s],$$

$$D(t) = \begin{bmatrix} -2 & 0 & 1 \\ 0 & -1 & 0 \\ 0 & 0 & 0 \end{bmatrix} \quad \text{for } 2 < t < 5[s],$$

and

$$D(t) = \begin{bmatrix} -5 & 0 & 0 \\ 2 & -9 & 6 \\ 0.01 & 0.04 & 0.30 \end{bmatrix} \quad \text{for } t > 5[s].$$

Moreover, $\boldsymbol{q}, \boldsymbol{u}(t)$ and η are chosen as in the previous case. Similarly to what was shown in the previous simulations, the hidden state $\boldsymbol{h}(t)$ reaches the consensus value of 3.49 (Figure 13.9), i.e. the weighted average of the initial conditions $\boldsymbol{h}_0$, while the exchanged variable $\boldsymbol{z}(t)$ is synchronized but continues to oscillate (Figure 13.10). Finally, Figure 13.11 shows the trend of the variable x, which expresses the dynamics of the chosen nonlinear function.

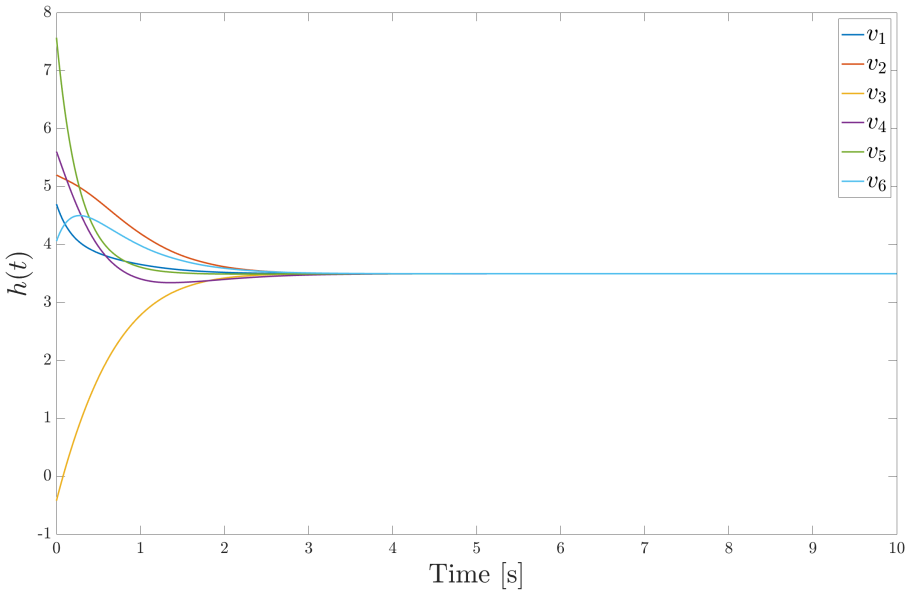


Figure 13.9: Temporal evolution of local state variable $\boldsymbol{h}(t)$ carrying the hidden consensus value considering the generic nonlinear function.

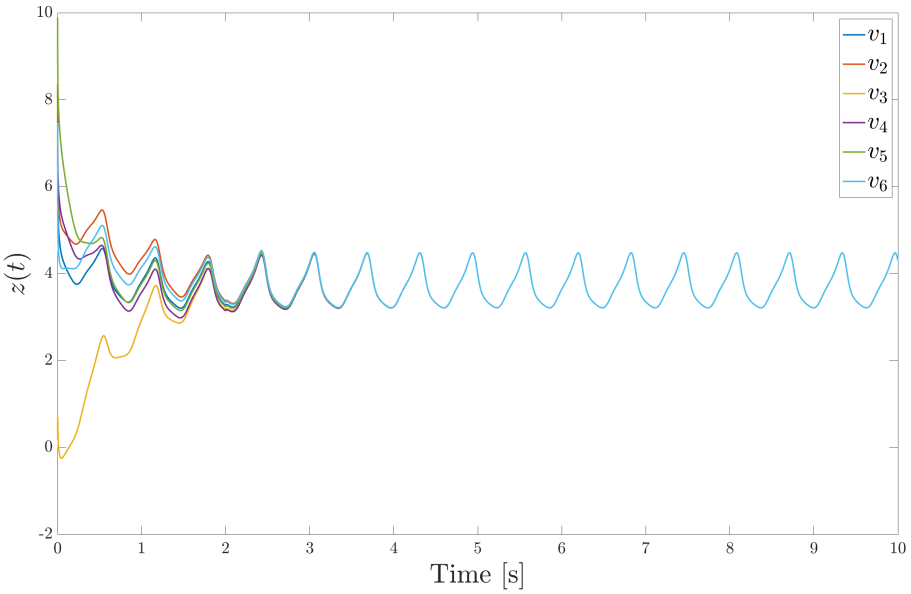


Figure 13.10: Temporal evolution of the shared state variable $\boldsymbol{z}(t)$ considering the generic nonlinear function.

13.5 Discussion

In this chapter, we presented a novel consensus methodology that utilizes the synchronization of a network of nonlinear dynamic systems to guarantee privacy and security against eavesdroppers and malicious third parties. Specifically, we considered a directed and strongly connected graph topology, where agents aim to compute a weighted average of their initial conditions. Our approach splits the initial value to initialize both a distributed synchronization process for a network of nonlinear dynamical systems

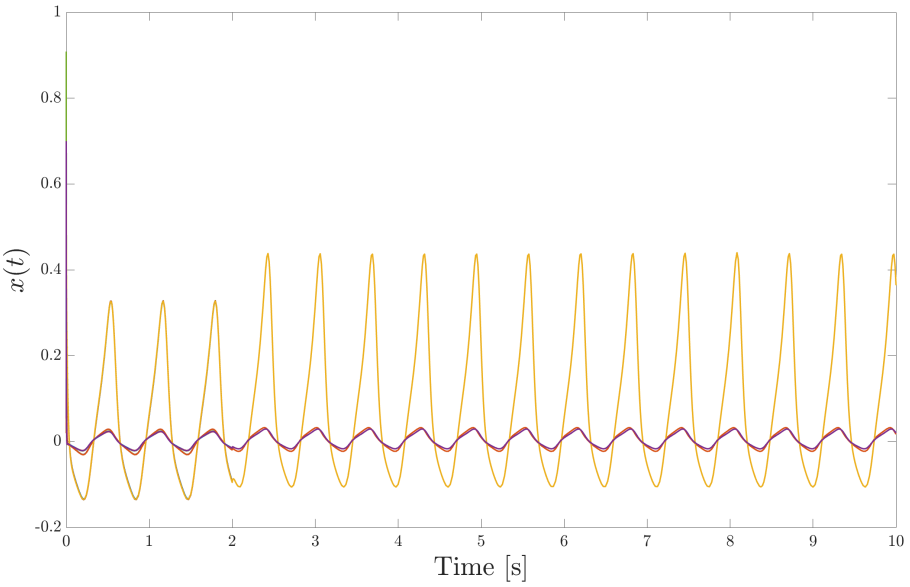


Figure 13.11: Temporal evolution of the dynamics $\boldsymbol{x}(t)$ when the nonlinear function $\boldsymbol{f}(\cdot)$ is as in Eq. (13.14).

characterized by noise, partial information sharing, and time-varying coupling, and a linear average consensus process that is influenced by the nonlinear synchronization.

By relying on nonlinear synchronization, the proposed approach provides a layer of security against malicious third parties who lack knowledge of the nonlinear models and the matrix used to reconstruct the consensus value. Furthermore, the approach ensures privacy by introducing noise and different transients for the agents before synchronization is achieved.

Compared to our preliminary conference chapter, this chapter presents significant improvements in terms of generality and robustness. We extended the approach from the synchronization of a network of Chua oscillators over an undirected graph with static coupling and in the absence of noise to a general framework for the synchronization of networks of nonlinear systems, characterized by possibly directed interaction topologies, time-varying coupling, and presence of noise. We also investigated the privacy and security characteristics of the framework and established a relation between synchronization and the overall dynamics as a gradient descent along a strongly convex function.

Overall, the proposed approach has promising potential for real-world applications that require secure and private consensus in distributed networks. Expected future work directions include the possibility to extend the framework in order to withstand communication delays, and the application of the methodology to distributed estimation or optimization algorithms.

References — Secure and Resilient Consensus in Networked Multi-Agent Systems

[8] Olfati-Saber, R., Fax, J. A., and Murray, R. M., "Consensus and cooperation in networked multi-agent systems," *Proceedings of the IEEE*, vol. 95, no. 1, pp. 215–233, 2007.

[11] Fioravanti, C., Oliva, G., Faramondi, L., and Hadjicostis, C. N., "A geometrical approach for consensus security," *Systems & Control Letters*, Submitted.

[13] Olfati-Saber, R. and Murray, R. M., "Consensus problems in networks of agents with switching topology and time-delays," *IEEE Transactions on automatic control*, vol. 49, no. 9, pp. 1520–1533, 2004.

[14] Sabattini, L., Secchi, C., Chopra, N., and Gasparri, A., "Distributed control of multirobot systems with global connectivity maintenance," *IEEE Transactions on Robotics*, vol. 29, no. 5, pp. 1326–1332, 2013.

[16] Olfati-Saber, R., "Kalman-consensus filter: Optimality, stability, and performance," in *Proceedings of the 48h IEEE Conference on Decision and Control (CDC) held jointly with 2009 28th Chinese Control Conference*, Ieee, 2009, pp. 7036–7042.

[18] Oliva, G., Rikos, A. I., Gasparri, A., and Hadjicostis, C. N., "Distributed negotiation for reaching agreement among reluctant players in cooperative multi-agent systems," *IEEE Transactions on Automatic Control*, 2022.

[20] Mo, Y. and Murray, R. M., "Privacy preserving average consensus," *IEEE Transactions on Automatic Control*, vol. 62, no. 2, pp. 753–765, 2016.

[21] Nozari, E., Tallapragada, P., and Cortés, J., "Differentially private average consensus: Obstructions, trade-offs, and optimal algorithm design," *Automatica*, vol. 81, pp. 221–231, 2017.

[22] Hadjicostis, C. N. and Domínguez-García, A. D., "Privacy-preserving distributed averaging via homomorphically encrypted ratio consensus," *IEEE Transactions on Automatic Control*, vol. 65, no. 9, pp. 3887–3894, 2020.

[23] Fang, W., Zamani, M., and Chen, Z., "Secure and privacy preserving consensus for second-order systems based on paillier encryption," *Systems & Control Letters*, vol. 148, p. 104 869, 2021.

[24] Calis, M., Heusdens, R., and Hendriks, R. C., "A privacy-preserving asynchronous averaging algorithm based on state decomposition," in *2020 28th European Signal Processing Conference (EUSIPCO)*, IEEE, 2021, pp. 2115–2119.

[25] Kishida, M., "Encrypted average consensus with quantized control law," in *2018 IEEE Conference on Decision and Control (CDC)*, IEEE, 2018, pp. 5850–5856.

[26] Rikos, A. I., Charalambous, T., Johansson, K. H., and Hadjicostis, C. N., "Distributed event-triggered algorithms for finite-time privacy-preserving quantized average consensus," *IEEE Transactions on Control of Network Systems*, 2022.

[27] Fioravanti, C., Oliva, G., Panzieri, S., and Hadjicostis, C. N., "Private consensus using chaotic oscillator-based encryption," in *2022 30th Mediterranean Conference on Control and Automation (MED)*, IEEE, 2022, pp. 927–932.

[28] Xiong, Y. and Li, Z., "Privacy preserving average consensus by adding edge-based perturbation signals," in *2020 IEEE Conference on Control Technology and Applications (CCTA)*, IEEE, 2020, pp. 712–717.

[29] Boopalan, P. et al., "Fusion of federated learning and industrial internet of things: A survey," *Computer Networks*, p. 109 048, 2022.

[30] Abdel-Basset, M., Moustafa, N., and Hawash, H., "Privacy-preserved generative network for trustworthy anomaly detection in smart grids: A federated semisupervised approach," *IEEE transactions on industrial informatics*, vol. 19, no. 1, pp. 995–1005, 2022.

[105] Kumar, P., Lin, Y., Bai, G., Paverd, A., Dong, J. S., and Martin, A., "Smart grid metering networks: A survey on security, privacy and open research issues," *IEEE Communications Surveys & Tutorials*, vol. 21, no. 3, pp. 2886–2927, 2019.

[106] Gebremichael, T. et al., "Security and privacy in the industrial internet of things: Current standards and future challenges," *IEEE Access*, vol. 8, pp. 152 351–152 366, 2020.

[107] Godsil, C. and Royle, G. F., *Algebraic graph theory*. Springer Science & Business Media, 2001, vol. 207.

[108] Wu, C. W. and Chua, L. O., "A unified framework for synchronization and control of dynamical systems," *International Journal of Bifurcation and Chaos*, vol. 4, no. 04, pp. 979–998, 1994.

[109] Wu, C. W., *Synchronization in complex networks of nonlinear dynamical systems*. World scientific, 2007.

[110] Liu, X., Li, C., Ge, S. S., and Li, D., "Time-synchronized control of chaotic systems in secure communication," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 69, no. 9, pp. 3748–3761, 2022.

[111] Fiore, D. and Russo, G., "Resilient consensus for multi-agent systems subject to differential privacy requirements," *Automatica*, vol. 106, pp. 18–26, 2019.

[112] Rezazadeh, N. and Kia, S. S., "Privacy preservation in a continuous-time static average consensus algorithm over directed graphs," in *2018 Annual American Control Conference (ACC)*, IEEE, 2018, pp. 5890–5895.

[113] Matthews, R., "On the derivation of a "chaotic" encryption algorithm," *Cryptologia*, vol. 13, no. 1, pp. 29–42, 1989.

[114] Schmitz, R., "Use of chaotic dynamical systems in cryptography," *Journal of the Franklin Institute*, vol. 338, no. 4, pp. 429–441, 2001.

[115] De la Hoz, M. Z., Acho, L., and Vidal, Y., "A modified chua chaotic oscillator and its application to secure communications," *Applied Mathematics and Computation*, vol. 247, pp. 712–722, 2014.

[116] Loan, N. A., Hurrah, N. N., Parah, S. A., Lee, J. W., Sheikh, J. A., and Bhat, G. M., "Secure and robust digital image watermarking using coefficient differencing and chaotic encryption," *IEEE Access*, vol. 6, pp. 19 876–19 897, 2018.

[117] Wang, M., Wang, X., Zhang, Y., and Gao, Z., "A novel chaotic encryption scheme based on image segmentation and multiple diffusion models," *Optics & Laser Technology*, vol. 108, pp. 558–573, 2018.

[119] Jia, P., MirTabatabaei, A., Friedkin, N. E., and Bullo, F., "Opinion dynamics and the evolution of social power in influence networks," *SIAM review*, vol. 57, no. 3, pp. 367–397, 2015.

[291] Costanzo, E. D., Menci, M., Messina, E., Natalini, R., and Vecchio, A., "A hybrid model of collective motion of discrete particles under alignment and continuum chemotaxis.," *Discrete & Continuous Dynamical Systems-Series B*, vol. 25, no. 1, 2020.

[292] Peng, H., Wang, J., Wang, S., Shen, W., Shi, D., and Liu, D., "Coordinated motion control for a wheel-leg robot with speed consensus strategy," *IEEE/ASME Transactions on Mechatronics*, vol. 25, no. 3, pp. 1366–1376, 2020.

[293] Xiao, Y., Zhang, N., Lou, W., and Hou, Y. T., "A survey of distributed consensus protocols for blockchain networks," *IEEE Communications Surveys & Tutorials*, vol. 22, no. 2, pp. 1432–1465, 2020.

[294] Bhardwaj, R. and Datta, D., "Consensus algorithm," in *Decentralised Internet of Things*, Springer, 2020, pp. 91–107.

[295] Patel, S., Khatana, V., Saraswat, G., and Salapaka, M. V., "Distributed detection of malicious attacks on consensus algorithms with applications in power networks," in *2020 7th International Conference on Control, Decision and Information Technologies (CoDIT)*, IEEE, vol. 1, 2020, pp. 397–402.

[296] Maglaras, L. A. et al., "Cyber security of critical infrastructures," *Ict Express*, vol. 4, no. 1, pp. 42–45, 2018.

[297] Peng, C., Sun, H., Yang, M., and Wang, Y.-L., "A survey on security communication and control for smart grids under malicious cyber attacks," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 49, no. 8, pp. 1554–1569, 2019.

[298] Saldana, D., Prorok, A., Sundaram, S., Campos, M. F., and Kumar, V., "Resilient consensus for time-varying networks of dynamic agents," in *2017 American control conference (ACC)*, IEEE, 2017, pp. 252–258.

[299] Zhai, Y., Liu, Z.-W., Ge, M.-F., Wen, G., Yu, X., and Qin, Y., "Trusted-region subsequence reduction for designing resilient consensus algorithms," *IEEE Transactions on Network Science and Engineering*, vol. 8, no. 1, pp. 259–268, 2020.

[300] Zhao, C., He, J., and Chen, J., "Resilient consensus with mobile detectors against malicious attacks," *IEEE Transactions on Signal and Information Processing over Networks*, vol. 4, no. 1, pp. 60–69, 2017.

[301] Yuan, L. and Ishii, H., "Secure consensus with distributed detection via two-hop communication," *Automatica*, vol. 131, p. 109 775, 2021.

[302] Singh, D. P., Sowerby, K. W., and Austin, A. C., "Distributed trust and reputation management for future wireless systems," *IEEE Communications Magazine*, 2022.

[303] Kang, J., Xiong, Z., Niyato, D., Ye, D., Kim, D. I., and Zhao, J., "Toward secure blockchain-enabled internet of vehicles: Optimizing consensus management using reputation and contract theory," *IEEE Transactions on Vehicular Technology*, vol. 68, no. 3, pp. 2906–2920, 2019.

[304] Biryukov, A., Feher, D., and Khovratovich, D., "Guru: Universal reputation module for distributed consensus protocols," *Cryptology ePrint Archive*, 2017.

[305] Hallajiyan, M., Hassani, H., Razavi-Far, R., and Saif, M., "Consensus and reputation-based resilient control of networked microgrids," in *2021 4th IEEE International Conference on Industrial Cyber-Physical Systems (ICPS)*, IEEE, 2021, pp. 619–624.

[306] Ramos, G., Silvestre, D., and Silvestre, C., "A discrete-time reputation-based resilient consensus algorithm for synchronous or asynchronous communications," *arXiv preprint arXiv:2107.00431*, 2021.

[307] Shafer, G., "A mathematical theory of evidence," *Princeton University pres*, pp. 1–6, 1976.

[308] Dempster, A., "Upper and lower probabilities induced by mutivalued mapping," *The annals of mathematical statistics*, vol. 38, no. 2, pp. 325–339, 1967.

[309] Yager, R., "On the dempster-shafer framework and new combination rules," *Information Sciences*, vol. 41, pp. 93–137, 1987.

[310] Gasparri, A., Fiorini, F., Di Rocco, M., and Panzieri, S., "A networked transferable belief model approach for distributed data aggregation," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 42, no. 2, pp. 391–405, 2011.

[311] Ducourthial, B., Cherfaoui, V., and Denoeux, T., "Self-stabilizing distributed data fusion," in *Symposium on Self-Stabilizing Systems*, Springer, 2012, pp. 148–162.

[312] C. Foglietta, A. G. and Panzieri, S., "A networked evidence theory framework for critical infrastructure modeling," *Critical Infrastructure Protection VI*, pp. 205–215, 2012.

[313] Manitara, N. E. and Hadjicostis, C. N., "Privacy-preserving asymptotic average consensus," in *2013 European Control Conference (ECC)*, IEEE, 2013, pp. 760–765.

[314] Charalambous, T., Manitara, N. E., and Hadjicostis, C. N., "Privacy-preserving average consensus over digraphs in the presence of time delays," in *2019 57th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, IEEE, 2019, pp. 238–245.

[315] Pilet, A. B., Frey, D., and Taıani, F., "Robust privacy-preserving gossip averaging," in *International Symposium on Stabilizing, Safety, and Security of Distributed Systems*, Springer, 2019, pp. 38–52.

[316] Hadjicostis, C. N., "Privary preserving distributed average consensus via homomorphic encryption," in *2018 IEEE Conference on Decision and Control (CDC)*, IEEE, 2018, pp. 1258–1263.

[317] Ferrari, R. M. and Teixeira, A. M., "Detection of cyber-attacks: A multiplicative watermarking scheme," in *Safety, Security and Privacy for Cyber-Physical Systems*, Springer, 2021, pp. 173–201.

[318] Smets, P., "Decision making in the tbm: The necessity of the pignistic transformation," *International journal of approximate reasoning*, vol. 38, no. 2, pp. 133–147, 2005.

[319] Ridgley, I. L. D., Freeman, R. A., and Lynch, K. M., "Private and hot-pluggable distributed averaging," *IEEE Control Systems Letters*, vol. 4, no. 4, pp. 988–993, 2020.

[320] Stewart, J., *Calculus*. Cengage Learning, 2011.

Observering and Simulation of Critical Infrastructure Interdependencies

Chapter 14

Introduction to Critical Infrastructure Interdependencies Modelling

In the previous parts of this thesis, we developed a progression of analytical and algorithmic frameworks to enhance the security, reliability, and awareness of distributed cyber–physical systems. Starting from the *anomaly detection and isolation* methodology, we established the foundations for identifying and reconstructing adversarial behavior in estimation and control loops. We then introduced the *Age of Incorrect Information* framework to characterize how malicious interference affects the timeliness and correctness of exchanged data, providing a principled way to quantify information degradation under attack. Building upon these insights, we examined the problem of *secure and resilient consensus*, where multiple agents must cooperatively reach agreement in the presence of untrusted or privacy-constrained communication. Together, these contributions form a coherent toolkit for protecting the fundamental building blocks of networked decision-making: detection, mitigation, and coordination.

The final part of this thesis broadens the perspective from individual systems and networked agents to the macroscopic scale of *Critical Infrastructures* (CIs). These infrastructures form the backbone of modern societies, encompassing essential services such as energy, transport, water, telecommunications, finance, and public administration. Their continuous and secure operation underpins economic stability, national security, and citizens' well-being. However, the growing *interconnection* among these infrastructures—driven by digitalization, automation, and the proliferation of cyber–physical systems—has created a complex web of dependencies that can amplify both efficiency and vulnerability. A disruption that once remained local can now cascade across multiple sectors, causing widespread societal and economic impacts.

This increasing complexity has reshaped the way resilience and risk are conceptualized. In today's interconnected environment, it is no longer sufficient to protect individual assets or systems in isolation. Instead, it is necessary to understand how infrastructures *interact*, how failures propagate through physical, cyber, and logical channels, and how feedback loops can intensify or mitigate cascading effects. Such understanding requires not only sectoral expertise but also *integrated models* capable of representing the multi-layered nature of interdependencies.

Cyber threats have further exacerbated this challenge. Modern infrastructures depend on digital control, monitoring, and communication systems—such as SCADA networks, IoT platforms, and cloud-based services—that introduce new attack vectors and interconnectivity risks. A cyberattack on one component of a critical infrastructure can have physical consequences across others, bridging the digital and physical domains. This has shifted the focus of research and policy from traditional reliability analysis toward a systemic view of *cyber–physical interdependencies*. Modeling and simulating these interconnections are therefore crucial to anticipate cascading failures, evaluate vulnerabilities, and design effective resilience and response strategies.

Over the past two decades, a vast body of literature has emerged to capture the complexity of interdependent infrastructures. Early studies emphasized qualitative frameworks for mapping dependencies and assessing risk propagation. More recent works have introduced quantitative, simulation-based approaches that integrate physical processes, cyber dependencies, and human factors. Despite these advances, several challenges persist: representing both the technical and organizational layers of interdependence, incorporating uncertainty and incomplete data, and bridging the gap between academic modeling and the needs of policy and governance.

By situating the previously developed methods for anomaly detection, adversarial resilience, and secure coordination within the broader context of interdependent infrastructures, this final part seeks to connect *micro-level mechanisms* with *macro-level systemic behavior*. The ultimate objective is to show how the analytical and algorithmic tools proposed in earlier chapters can inform the design of *integrated resilience models* for complex socio-technical systems, where local security decisions and global interdependencies jointly determine the robustness of modern critical infrastructure networks.

Relation to the Author's Publications. The research presented in this chapter builds upon the author's contributions to the modeling and simulation of interdependent Critical Infrastructures (CIs) developed through several international conference publications. The foundations of this work were established in [C6], which introduced the use of the *CISIApro 2.0* simulator to analyze the Italian *National Security Perimeter for Cyber*. That study formalized the modeling of cross-sector dependencies—encompassing physical, cyber, and logical relations—and demonstrated how interdependency analysis can support national-level cybersecurity assessments and policy design.

Building upon this foundation, further developments were presented in [C13] and [C14], which extended the CISIApro framework to represent uncertainty and ambiguity in infrastructure dependencies. In particular, [C13] explored the integration of *fuzzy logic* for impact assessment and risk evaluation, providing a flexible mechanism to quantify the effects of incomplete or imprecise data on systemic resilience analysis. Subsequently, [C14] demonstrated how these modeling principles could be applied to large-scale, multi-sector simulations, enabling the representation of cyber–physical interdependencies and cascading effects across heterogeneous domains.

Organization of the Chapter. The remainder of this chapter is organized as follows:

- **Section 14.1** reviews the main approaches to modeling interdependencies in Critical Infrastructures. It discusses the evolution from qualitative frameworks to quantitative, simulation-based methods, emphasizing how cyber–physical–social interactions and cascading effects are represented across different modeling paradigms.
- **Section 14.2** introduces the adopted modeling framework and the *CISIApro 2.0* simulator. It describes the agent-based multilayer architecture used to capture interdependencies, the representation of cascading failures, and the mechanisms for redundancy, recovery, and resilience assessment.
- **Section 14.3** outlines the specific contributions of this dissertation to the field of interdependency modeling. It details the integration of regulatory data from the National Cybersecurity Perimeter, the introduction of uncertainty management through possibility theory, and the validation of the extended framework on realistic cyber–physical scenarios.

14.1 Literature Review on Modeling Interdependencies and Cascading Effects in Critical Infrastructures

The modeling of interdependencies among Critical Infrastructures has long been recognized as a cornerstone of resilience engineering and systemic risk assessment in complex socio–technical systems. Modern societies depend on tightly interconnected networks that deliver essential services such as energy, water, transportation, communication, finance, and healthcare. While each sector performs distinct functions, their increasing integration through digital control, automation, and shared information systems has generated a dense web of interdependencies. Consequently, disruptions originating in one sector can propagate to others through a combination of *physical*, *cyber*, *informational*,

or *geographic* links, producing *cascading failures* that amplify the scope, duration, and societal impact of an initial event [120]. Such phenomena—triggered by natural hazards, technical malfunctions, cyberattacks, or human errors—demonstrate that infrastructures cannot be analyzed or protected in isolation. Understanding and modeling these complex relationships has therefore become essential to safeguard national resilience, support evidence-based policymaking, and guide emergency and risk management.

14.1.1 Conceptual Foundations of Interdependency Modeling

The study of interdependencies among Critical Infrastructures originates from the recognition that the functionality and resilience of modern societies depend not only on the performance of individual infrastructures, but on the network of linkages that connect them. Early research efforts aimed to identify and classify these relationships, laying the conceptual groundwork for what would become a multidisciplinary field bridging engineering, economics, and risk science. The seminal contribution of [120] represents a cornerstone in this evolution: it formalized the notion that infrastructures are not isolated entities, but components of an interconnected "system-of-systems" whose global behavior results from both direct and indirect dependencies. The framework introduced four fundamental categories of interconnection—*physical, cyber, geographic, and logical*—which together define the main pathways through which perturbations may propagate across sectors. This taxonomy, which remains a reference standard in the field, provided a structured vocabulary for describing cross-sector relationships and laid the basis for systematic analysis of cascading effects.

Following this conceptual foundation, subsequent works—such as the comprehensive surveys by [121] and [5]—contributed to the formalization of interdependency modeling as a scientific discipline. These studies organized the growing body of literature into distinct methodological categories, distinguishing between *analytical models*, which use mathematical abstractions to capture dependency functions; *topological models*, which apply network theory to quantify connectivity and vulnerability; and *simulation-based approaches*, which employ computational experiments to explore the dynamic propagation of failures. Through this classification, the authors highlighted an essential insight: CIs behave as *non-linear, adaptive systems* where local disturbances can produce disproportionate or emergent global effects. Hence, traditional linear reliability analysis—suitable for isolated assets—was insufficient to capture the feedback-rich and dynamic nature of interdependent infrastructures.

This systemic perspective was further developed by works such as [122], [123], [321], which characterized infrastructures as *complex adaptive systems* (CAS). In this view, each infrastructure—be it power, water, or communication—operates as an autonomous subsystem capable of adaptation and local decision-making, while the overall system exhibits emergent properties not directly predictable from individual behavior. The CAS paradigm introduced key concepts such as multi-scale feedback, self-organization, and resilience through adaptation, which became fundamental for subsequent generations of interdependency models. This transition marked a decisive shift from *static* representations—focused on structural connectivity or steady-state robustness—to *dynamic* frameworks that explicitly account for temporal evolution, cascading propagation, recovery processes, and the role of uncertainty.

In parallel, the technological transformation of industrial systems profoundly reshaped the meaning of "dependency." The progressive digitalization of Industrial Control Systems, SCADA platforms, and IoT-based automation has blurred the boundaries between physical processes and cyber infrastructures [6]. This convergence introduced new classes of vulnerabilities, where failures or attacks in digital control layers can directly affect physical processes, and vice versa. Consequently, interdependency evolved from a purely physical notion—centered on the exchange of material resources such as energy or water—into a truly *cyber–physical–organizational phenomenon.* Within this expanded paradigm,

dependencies are no longer limited to the transfer of goods or energy, but also include information flows, decision hierarchies, and institutional linkages that shape the collective behavior of infrastructure systems. This realization has driven a methodological shift toward multidisciplinary frameworks that integrate systems engineering, network science, control theory, and cybersecurity, providing the foundation for modern cyber–physical risk modeling and resilience assessment.

Overall, the conceptual evolution of interdependency modeling reveals a clear trajectory: from qualitative taxonomies and static network representations, toward dynamic, adaptive, and cyber–physical frameworks that capture the complex feedbacks, temporal evolution, and uncertainty of real-world infrastructures. This theoretical maturation has not only advanced scientific understanding but also paved the way for simulation platforms and analytical tools capable of supporting evidence-based resilience planning and policy design.

14.1.2 Regulatory and Governance Frameworks

The growing scientific understanding of interdependencies among Critical Infrastructures has been accompanied by a parallel evolution in the regulatory and institutional domains, which increasingly frame resilience as a strategic, operational, and legal obligation. In the European Union, the recognition that systemic risks can propagate across sectors and borders has led to the creation of a harmonized policy framework designed to strengthen both cybersecurity and the resilience of essential services. This framework reflects a progressive convergence between *technical science* and *policy design*: the former providing quantitative insight into systemic vulnerabilities, the latter establishing governance mechanisms to translate such insight into coordinated action.

A major milestone in this process was the adoption of the *Directive on Security of Network and Information Systems (NIS Directive)* [322], which introduced the first EU-wide legal framework for cybersecurity governance. The NIS Directive represents a paradigm shift: it moved cybersecurity from a voluntary, sectoral concern to a matter of regulatory compliance and collective responsibility across Member States. By imposing harmonized obligations for operators of essential services (OESs) and digital service providers (DSPs), the Directive required systematic *risk assessment*, *incident reporting*, and the implementation of appropriate *technical and organizational measures*. Its transposition into national legislation—such as the Italian implementation [323]—extended these principles to domestic infrastructures, defining the national perimeter of essential operators and identifying the authorities responsible for oversight, incident coordination, and enforcement. Beyond compliance, the NIS Directive encouraged the adoption of proactive risk management practices and opened the path toward quantitative approaches to resilience assessment, thereby aligning with the research trends that view CIs as dynamically coupled, risk-propagating systems.

To consolidate this strategic direction, the European Union established the *European Cybersecurity Competence Centre (ECCC)* [324], an institution explicitly designed to bridge the gap between research, innovation, and operational policy. The ECCC coordinates a network of national competence centres, creating a pan-European ecosystem for cybersecurity capacity building, knowledge exchange, and technological innovation. Its mandate extends beyond coordination: it actively promotes the development of advanced analytical tools—such as modeling, simulation, and AI-based risk assessment platforms—that enable Member States to implement *evidence-based cybersecurity governance*. By fostering cross-border cooperation and the transfer of scientific results into strategic policy frameworks, the ECCC operationalizes the idea that resilience must be informed by quantitative, data-driven methodologies—a principle also at the core of current research on cyber–physical interdependencies [6], [138]. In this respect, the ECCC symbolizes the institutionalization of the dialogue between *resilience science* and *European cybersecurity policy*.

Within this European context, Italy has played a particularly proactive role in developing a coherent national cybersecurity architecture. The *Law Decree n.133/2019* [325] initiated the formal establishment of a coordinated cybersecurity governance system, introducing the concept of a national *cybersecurity perimeter* that encompasses both public and private operators of critical infrastructures. This perimeter-based approach reflects the systemic logic of interdependency analysis: rather than treating infrastructures as isolated entities, it recognizes the networked nature of risk propagation. Subsequent *Decrees of the President of the Council of Ministers—DPCM n.131/2020, n.81/2021, n.15/2021, and n.92/2022* [326], [327], [328], [329]—further refined this architecture by defining operational procedures, delineating institutional responsibilities, and specifying coordination protocols among government agencies, security services, and sectoral regulators. Together, these decrees translated the high-level principles of the NIS Directive into an actionable national governance structure.

A decisive institutional milestone in the Italian context was the establishment of the *Agenzia per la Cybersicurezza Nazionale (ACN)* through *Law Decree n.82/2021* [330]. The ACN represents the cornerstone of Italy's national cybersecurity architecture, acting as the central authority responsible for coordinating activities of prevention, preparedness, and incident response across all critical sectors. Within the Italian governance framework, it ensures the integration of technical, strategic, and regulatory functions, unifying competencies that were previously distributed among multiple institutions. At the same time, the Agency serves as Italy's official liaison with the European Union and international cybersecurity organizations, facilitating alignment with EU directives, such as the NIS and NIS2 frameworks, and with the initiatives of the *European Cybersecurity Competence Centre (ECCC)*. Through this dual national and European role, the ACN consolidates Italy's capacity to address systemic cyber–physical risks and to promote a coherent, data-driven approach to critical infrastructure resilience. In line with the European policy shift toward data-driven governance, the ACN actively promotes the use of modeling, simulation, and analytical tools to assess systemic risk and support strategic planning. The Agency's mandate explicitly acknowledges that resilience requires both *organizational integration* and *analytical capability*, highlighting the importance of simulation platforms and quantitative frameworks to anticipate and mitigate cascading failures in cyber–physical infrastructures.

This institutional trajectory culminated in the *Italian Cybersecurity Strategy 2022–2026* [331], which consolidates the principles of risk-based governance and systemic resilience into a unified strategic vision. The Strategy adopts a holistic approach that merges technological protection, human capability, and analytical foresight, establishing five strategic pillars that include risk prevention, technological autonomy, and operational readiness. A key innovation is the explicit integration of modeling and simulation into the national decision-making process: the document identifies digital-twin and agent-based simulation tools—such as *CISIApro* [59], [60], [61], [62], [128]—as essential enablers for systemic risk analysis, scenario-based planning, and policy validation. *CISIApro*, in particular, is highlighted as a reference platform for reproducing cross-sector dependencies, testing mitigation strategies, and quantifying the impact of cascading failures within the national cyber and infrastructure protection perimeter. By formally embedding analytical modeling within national strategy, Italy positions itself among the few EU Member States that explicitly recognize simulation as a *policy instrument* for resilience governance.

14.1.3 Modeling Cascading Effects in Interdependent Systems

A major consequence of the conceptual and regulatory advances discussed above has been the development of increasingly sophisticated modeling frameworks aimed at capturing *cascading effects* in interconnected infrastructures. As Critical Infrastructures have come to be recognized as tightly coupled, multi-layered networks—where physical, cyber, and organizational processes interact—modeling approaches have evolved from purely

structural representations to dynamic, probabilistic, and agent-based formulations capable of reproducing feedback, adaptation, and emergent behaviors. This methodological diversification reflects a broader evolution in resilience science: from assessing static robustness to analyzing dynamic, time-dependent processes of disturbance propagation and recovery.

The earliest efforts to model interdependencies drew heavily from *graph theory* and *percolation analysis*, which offered foundational tools to quantify network robustness and identify critical elements [5], [122]. In these models, infrastructures are represented as nodes connected by edges that encode functional or spatial dependencies. Cascading failures are studied by progressively removing nodes or links, thereby simulating the loss of components and observing how the connectivity of the overall system degrades. This approach provides key insights into structural fragility—allowing the estimation of robustness metrics such as the percolation threshold—but it abstracts away the temporal and operational dynamics of real infrastructures. As a result, these models are well suited for topological vulnerability assessment and for highlighting "critical nodes" whose failure might compromise overall connectivity, yet they cannot reproduce dynamic control processes or the role of information feedback in recovery.

To overcome these limitations, a second generation of approaches introduced *flow-based* and *coupled cyber–physical models*, which explicitly represent the transfer of energy, water, or data between subsystems through physical laws or control rules. These models integrate process dynamics—often formulated through systems of differential equations or discrete-event simulations—with communication and control architectures, thereby bridging the physical and cyber layers of infrastructures. For instance, [124] proposed a coupled simulation framework for electric power systems and SCADA networks, demonstrating how false data injection attacks can induce physical overloads and cascading blackouts. Similarly, [125] and [126] analyzed the propagation of cyber failures across interdependent networks, showing that latency, packet loss, and misinformation can substantially amplify system degradation. These studies highlight a key advancement: the transition from purely structural to *dynamic interdependency modeling*, where feedback, latency, and control-layer interactions play a central role. However, such high-fidelity models typically require detailed data and are computationally expensive, which limits their application to large-scale or cross-sector scenarios.

A complementary stream of research focuses on *system-dynamics* and *input–output* formulations, which capture interdependencies at a macroscopic level using aggregated variables and causal feedback loops [6]. These approaches extend methods originally developed in economics and ecology to the context of critical infrastructures, modeling how perturbations in one sector (e.g., energy production) affect others (e.g., transport or water) through dependency coefficients or flow matrices. System-dynamics models allow analysts to simulate feedback-rich phenomena—such as demand shocks or capacity degradation—over longer timescales, making them particularly suitable for strategic risk assessment and policy evaluation. Their key advantage lies in computational tractability and interpretability, as they can capture high-level cause–effect chains with limited data requirements. Nonetheless, these models are often based on linear or quasi-linear relationships and may overlook the non-linear dynamics, control adaptation, and discrete interactions characteristic of complex infrastructure systems.

In response to the need for models that balance detail, adaptability, and scalability, the research community has increasingly adopted *agent-based modeling* (ABM) as a unifying paradigm [127], [128]. In ABM, each infrastructure, subsystem, or even individual asset is represented as an autonomous agent endowed with rules governing its operation, communication, and response to disturbances. Dependencies are modeled as directed interactions among agents—representing physical flows, cyber communication, or organizational coordination—allowing emergent collective behaviors such as cascading failures, adaptive recovery, and coordinated defense to be observed. This approach naturally accommodates heterogeneity, stochasticity, and decentralized decision-making,

providing a realistic representation of modern cyber–physical–organizational systems. Agent-based frameworks thus serve as a bridge between analytical rigor and systemic realism, supporting both fine-grained simulations and large-scale policy exploration.

Several simulation environments have operationalized these paradigms into practical tools for risk and resilience analysis, each embodying a different balance between realism, scalability, and analytical tractability. Among the earliest examples, the *Critical Infrastructure Protection Modeling and Analysis (CIPMA)* program in Australia **ausgov** was designed as a national-level decision-support system combining economic and physical dependency modeling. CIPMA integrates data from multiple infrastructure sectors and employs input–output relationships to assess the macroeconomic and service-level consequences of disruptions, providing high-level scenario analysis for government and emergency agencies. Although primarily top-down and aggregated, it represents one of the first operational attempts to institutionalize interdependency modeling within public policy.

At a finer level of granularity, the *I2Sim* platform [332], developed at the University of British Columbia, introduced an agent-based paradigm in which infrastructures are modeled as interacting *resource-exchange agents*. Each agent represents a facility or service provider that consumes and produces resources—such as energy, water, or transport capacity—and interacts with others according to dependency rules. I2Sim enables analysts to quantify degradation propagation, recovery dynamics, and performance trade-offs across sectors. Its modular structure and resource-centric formalism make it well suited for exploring localized cascading effects and for validating mitigation strategies in both emergency management and restoration planning contexts.

Within the European research landscape, the *IRRIIS* and *DIESIS* projects [6] marked a significant step forward by introducing *federated simulation frameworks*. These initiatives established the High-Level Architecture as a standard for interoperability among heterogeneous simulators—each representing a specific infrastructure domain (e.g., power, water, or ICT). The federated approach allows domain-specific models, often with different temporal resolutions and abstraction levels, to interact through standardized interfaces, thereby enabling multi-sector, multi-scale analysis. This architecture paved the way for collaborative simulation environments in which physical processes, communication networks, and decision mechanisms can be jointly assessed under realistic operational conditions.

Building upon the DIESIS experience, the *DOMINO* simulator [333] was developed to analyze interdependent infrastructures from a cyber–physical perspective. DOMINO combines the structural rigor of network-based representations with the behavioral flexibility of agent-based modeling. It explicitly models both physical interconnections (e.g., resource flows, load transfers) and logical dependencies (e.g., information exchange, control relationships), allowing analysts to study how faults or cyberattacks propagate across coupled domains. A key feature of DOMINO is its modular architecture, which supports the integration of sector-specific models and enables *scenario-based analysis* for cascading failure and resilience assessment. Through this hybrid design, DOMINO serves as a reference platform in several European initiatives on critical infrastructure protection, offering quantitative tools for evaluating systemic risk and testing mitigation strategies.

More recent simulation platforms have further expanded this methodological spectrum by incorporating advanced concepts such as cyberattack propagation, human-in-the-loop behavior, and uncertainty quantification. For example, *CyberGRID* [132] focuses on coupled power and communication networks, integrating physical power flow equations with network traffic dynamics to evaluate vulnerabilities under cyberattacks. The *EMRALD* framework [133] provides a stochastic environment for reliability and resilience analysis, where fault trees and probabilistic models are used to simulate cascading disruptions across cyber–physical systems. Similarly, *InfraRisk* [134] introduces techniques for handling incomplete and imprecise dependency information, combining graph-based analysis with uncertainty modeling to estimate the likelihood and severity of

cascading failures even in data-scarce contexts. Together, these platforms exemplify the ongoing evolution from deterministic, single-domain simulators toward hybrid, data-driven environments that integrate cyber, physical, and organizational dimensions of infrastructure resilience.

Each of these tools embodies a distinct modeling philosophy: from top-down economic models such as CIPMA, which emphasize aggregated interdependencies, to bottom-up, dynamic environments like I2Sim, DOMINO, or CyberGRID, which capture feedback-rich, multi-domain interactions. Federated frameworks like DIESIS provide interoperability across specialized simulators, while probabilistic platforms such as EMRALD and InfraRisk focus on uncertainty management and resilience quantification. Collectively, they demonstrate the growing maturity of interdependency modeling, highlighting the trade-offs between abstraction, realism, and computational scalability that define current research and operational practice in critical infrastructure resilience.

Within this methodological landscape, this thesis adopts an *agent-based and multilayer approach*, conceptually aligned with models such as *I2Sim* or *InfraRisk*, and implemented here using the *CISIApro* framework [59], [60], [61]. While CISIApro provides the specific simulation environment used in our analysis, the methodological rationale is shared across the broader class of hybrid models that represent infrastructures as interacting agents linked by functional and informational dependencies. These frameworks collectively support the study of cascading failures, resilience metrics, and mitigation strategies in increasingly complex and digitalized critical infrastructure ecosystems.

14.1.4 Advanced Research Directions

Contemporary research on interdependent Critical Infrastructures (CIs) is increasingly expanding beyond traditional engineering boundaries, embracing multidisciplinary perspectives that integrate technological, informational, and human dimensions. This evolution reflects the recognition, already highlighted by [6], [120], [138], that resilience is not merely a matter of physical redundancy or cyber protection, but rather an *emergent property* arising from the coordinated behavior of complex socio–technical systems. Modern infrastructures operate at the intersection of cyber, physical, and social domains, where systemic stability depends simultaneously on technological robustness, adaptive control, and organizational governance.

A first prominent research direction concerns the integration of *artificial intelligence* , *blockchain*, and *big data analytics* within CI modeling and management frameworks. The use of AI-driven techniques—such as machine learning, deep reinforcement learning, and anomaly detection—has rapidly expanded to applications including predictive maintenance, degradation forecasting, and dynamic resource allocation in cyber–physical networks [126]. These methods enable continuous monitoring and early identification of anomalies, complementing classical observer-based and state-estimation strategies discussed in previous chapters. Blockchain architectures, as noted in [126], are increasingly employed to secure distributed control systems and ensure data integrity across heterogeneous operators, reducing the risk of tampering or misinformation. When combined with large-scale IoT sensing and data fusion techniques, these digital technologies facilitate the construction of *data-informed resilience models*, capable of performing real-time risk assessment and adaptive control. Furthermore, hybrid analytical environments that merge *geographical information systems (GIS)*, *semantic data integration*, and *physics-based models* have advanced geographically explicit vulnerability assessments [138], [334], enabling analysts to spatially identify cross-sector interdependencies and potential points of failure. Through these converging developments, the field is progressively shifting from static modeling toward a paradigm of *intelligent and adaptive resilience*, where continuous learning and automated reasoning form the backbone of decision support.

A second major line of inquiry focuses on understanding the *human and organizational dimensions* of cascading failures and recovery processes. Empirical analyses and behavioral

reliability models [6] demonstrate that operator behavior, maintenance practices, and inter-organizational coordination are often decisive in determining the scale and duration of disruptions. Human factors—such as delayed responses, communication breakdowns, or conflicting operational priorities—can act as catalysts for systemic escalation even when physical systems remain functionally intact. This awareness has driven the emergence of *Cyber–Physical–Social Systems* frameworks [335], which explicitly integrate human cognition, social interactions, and institutional dynamics into cyber–physical models. By embedding governance and decision-making processes within technical simulations, CPSS approaches bridge the gap between engineering resilience and organizational adaptability. Such socio-technical modeling is now essential to capture how collective learning, collaboration, and institutional flexibility shape systemic recovery, reinforcing the notion that resilience must be treated as a distributed and adaptive capability rather than a static design feature.

A third, and increasingly critical, research frontier addresses the *representation and quantification of uncertainty* in interdependency modeling. As infrastructures operate in highly dynamic and data-scarce environments, complete information about dependencies, fault probabilities, or recovery functions is rarely available [5], [139]. Accordingly, recent studies distinguish between *aleatoric uncertainty*, arising from intrinsic randomness in system behavior or environmental variability, and *epistemic uncertainty*, stemming from incomplete or imprecise knowledge [336], [337]. To accommodate both types, probabilistic and non-probabilistic reasoning frameworks—such as *possibility theory* and *fuzzy logic*—have been increasingly integrated into CI simulation environments [338], [339], [340]. These methods enable dependencies to be modeled as graded relationships, allowing analysts to distinguish between plausible and necessary outcomes under partial information.

Overall, these emerging research directions converge toward a unified vision of *next-generation interdependency modeling*—one that integrates intelligent data processing, human-in-the-loop decision-making, and rigorous uncertainty quantification [6], [138]. By combining AI-enhanced prediction, socio-technical representation, and hybrid probabilistic reasoning, the field is evolving toward the realization of *adaptive, data-informed, and uncertainty-aware resilience frameworks*. These innovations are essential for managing the increasing complexity of cyber–physical–social infrastructures, where systemic resilience depends as much on human cognition and institutional design as on engineering robustness and digital security.

14.2 System Model and CISIApro 2.0

The modeling framework adopted in this thesis builds upon the *CISIApro 2.0* platform, an agent-based, multilayer simulation environment specifically designed for analyzing interdependencies, cascading effects, and resilience within Critical Infrastructures [59], [60], [61], [62]. In CISIApro, each infrastructure is represented as an autonomous *agent* that exchanges resources, information, or services with other agents according to predefined dependency rules. These dependencies can represent physical flows (e.g., power, water, or fuel), cyber connections (e.g., data or control signals), geographic proximity, or logical and organizational relationships. By encoding such links as directed relations, the simulator reproduces how perturbations—whether accidental or intentional—propagate across sectors, potentially triggering cascading failures.

Each agent is characterized by a set of *state variables* that quantify its operational condition, internal resource levels, and exposure to cyber risks. Cybersecurity is modeled through the three classical dimensions of the CIA triad—*confidentiality*, *integrity*, and *availability*—which together determine the effective functionality of the infrastructure node. These variables evolve dynamically according to both endogenous behaviors (e.g., internal degradation, recovery, or control policies) and exogenous stimuli (e.g., failures or attacks originating from connected agents). This representation enables the analysis

of complex cyber–physical–organizational interactions in a consistent system-of-systems framework.

The simulator follows a *Mixed Holistic–Reductionist (MHR)* modeling philosophy, depicted in Figure 14.1. This architecture allows analysts to combine different levels of abstraction within a single simulation: a holistic layer provides a high-level, aggregated view of interdependencies when detailed data are unavailable, while reductionist components introduce fine-grained, process-level details where sufficient information exists. The result is a flexible and scalable structure capable of balancing representational detail with computational efficiency—an essential property for modeling large-scale, multi-sector networks.

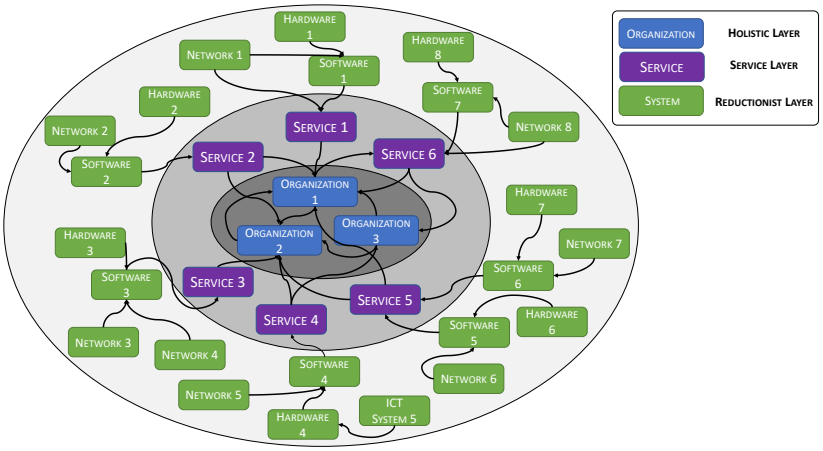


Figure 14.1: The Mixed Holistic–Reductionist framework for modeling interdependent Critical Infrastructures.

Within this architecture, each infrastructure agent (Figure 14.2) includes modules representing: (i) its internal state dynamics, (ii) its dependency relations with other agents, and (iii) its response and recovery mechanisms. These modules interact through a simulation kernel that synchronizes the evolution of all agents and manages cascading events across interconnected layers. The simulator supports both stochastic and scenario-based analyses, allowing users to explore the outcomes of accidental failures, natural hazards, or deliberate cyberattacks. Resilience metrics—such as functionality loss, recovery time, and impact duration—are computed to evaluate the system's ability to absorb, adapt, and recover from disruptions.

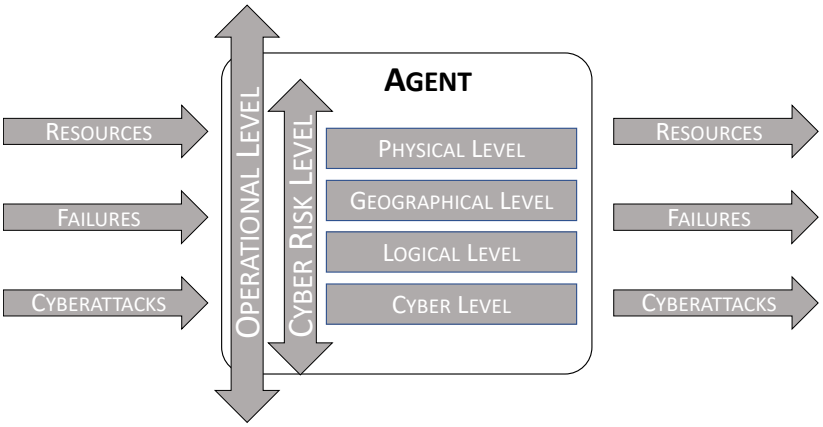


Figure 14.2: Representation of an infrastructure as an agent in CISIApro 2.0.

14.3 Thesis Contribution

Within this framework, the dissertation contributes to the advancement of interdependency modelling and cyber–physical risk assessment in three complementary directions:

1. **Integration of regulatory frameworks into systemic modelling.** The thesis operationalises the *Italian National Cybersecurity Perimeter (PSNC)* as a systemic boundary for simulation, demonstrating how regulatory data and sectoral mappings can be translated into a multilayer, agent-based representation of interdependent infrastructures. This integration bridges the gap between policy and modelling, enabling quantitative assessment of cascading cyber risks within a nationally defined governance context.
2. **Incorporation of epistemic uncertainty through possibilistic reasoning.** Building upon *CISIApro 2.0*, the work introduces a *possibilistic extension* that explicitly represents incomplete or ambiguous information using necessity and possibility measures. This enhancement allows the propagation of uncertainty across dependency networks without requiring arbitrary probabilistic assumptions, providing analysts with an interpretable confidence interval for each simulated outcome.
3. **Development of a hybrid fuzzy–possibilistic modelling framework.** The research further extends CISIApro to support fuzzy logic and fuzzy set representations, enabling the use of qualitative and linguistic inputs (e.g., "high dependency," "limited redundancy") within the same computational environment. This hybrid formulation merges quantitative data with expert judgment, offering a more nuanced and human-consistent approach to uncertainty modelling and systemic resilience evaluation.

Chapter 15

National Cybersecurity Perimeter: Modelling Systemic Risk with CISIApro 2.0

The protection of national digital infrastructures requires not only effective technical safeguards but also a systemic understanding of how cyber threats can propagate across interdependent sectors. In Italy, this systemic perspective has been institutionalized through the *Perimetro di Sicurezza Nazionale Cibernetica* (PSNC, National Cybersecurity Perimeter), established by *Law Decree n.133/2019* and its subsequent implementing decrees [326], [327], [328], [329], [330]. The Perimeter identifies operators and digital services essential to national security, prescribing obligations in risk management, certification, and incident reporting, as formalized by the *Agenzia per la Cybersicurezza Nazionale (ACN)* within the *Italian Cybersecurity Strategy 2022–2026* [331].

Although originally conceived as a regulatory and governance instrument, the PSNC implicitly delineates a *systemic boundary* for the national digital ecosystem. Within this boundary, infrastructures such as energy, telecommunications, finance, and transport form an interdependent network whose stability depends on both technological robustness and cyber integrity. Understanding these relationships is crucial for quantifying *systemic cyber risk*—the potential for localized incidents to generate cross-sector cascading effects.

This chapter presents the first original contribution of the dissertation, demonstrating how policy-driven data from the National Cybersecurity Perimeter can be operationalized within the *CISIApro 2.0* simulation environment to perform national-scale systemic risk analysis. Building on the methodological foundation developed in Chapter 14, the Perimeter's information is mapped onto a multilayer, agent-based model where each operator is represented as an autonomous entity interacting through functional, cyber, and organizational dependencies. This translation from regulation to simulation enables the quantitative exploration of cascading cyber events, revealing how different attack typologies may propagate through the national ecosystem and affect essential services.

Two representative threat scenarios are investigated to illustrate this approach: (i) a ransomware campaign targeting the financial sector, and (ii) a distributed denial-of-service (DDoS) attack directed at telecommunications operators. These attack classes are identified by ENISA as among the most pervasive and disruptive across European infrastructures [341], [342]. Their combined analysis allows us to compare integrity-driven versus availability-driven propagation mechanisms and to assess the asymmetric impact of cyber events across interconnected sectors.

The findings show that telecommunications and energy infrastructures act as critical propagation hubs within the Italian cyber–physical ecosystem, amplifying disturbances across dependent services. Moreover, the study highlights the broader value of the PSNC: beyond its regulatory function, it provides a structured data layer that can be repurposed for quantitative resilience assessment, supporting national risk governance and cross-sector coordination.

Organization of the Chapter. The remainder of this chapter is organized as follows:

- **Section 15.1** introduces the main regulatory, institutional, and systemic characteristics of the Italian National Cybersecurity Perimeter, outlining its role as both a governance and data framework.

- **Section 15.2** describes how PSNC-derived information is translated into the *CISIApro 2.0* simulation model, detailing the agent-based architecture and dependency structure adopted in the analysis.
- **Section 15.3** presents the ransomware and DDoS case studies, analyzing their propagation mechanisms, sectoral impacts, and cascading effects within the national ecosystem.
- **Section 15.4** discusses the main results, their policy implications for national resilience planning, and the methodological challenges of translating regulatory data into quantitative systemic risk models.

15.1 Regulatory and Systemic Framework of the Italian National Cybersecurity Perimeter

The *Perimetro di Sicurezza Nazionale Cibernetica* (PSNC, National Cybersecurity Perimeter) represents the cornerstone of Italy's institutional architecture for protecting critical digital infrastructures. Established by *Law Decree No. 105/2019* (converted into Law No. 133/2019) and subsequently refined through implementing decrees [326], [327], [328], [329], [330], the Perimeter identifies the public and private entities whose information systems are essential to the national interest. It defines their obligations in terms of cybersecurity governance, certification, and incident reporting, thereby creating a unified regulatory ecosystem for the protection of the country's most sensitive digital assets.

From a regulatory perspective, the PSNC establishes a multi-layered framework that links operators, supervisory bodies, and national authorities. The *Agenzia per la Cybersicurezza Nazionale* (ACN) acts as the technical and operational authority responsible for implementing the Perimeter, defining security standards, coordinating incident responses, and supervising compliance. Other key actors include the *Presidenza del Consiglio dei Ministri*, which maintains strategic oversight, and the sectoral ministries that contribute domain-specific expertise (e.g., energy, transport, finance, telecommunications). This governance structure ensures that cybersecurity policies are aligned across institutional, industrial, and technological dimensions, creating a coherent interface between regulation and national resilience planning.

At the systemic level, the PSNC delineates what can be understood as the *national digital ecosystem boundary*. Within this boundary, critical infrastructures—such as electricity, telecommunications, finance, transport, and public administration—are formally recognised as mutually interdependent components of a single cyber–physical system. This interconnection reflects both operational and informational dependencies: power sustains ICT systems; telecommunications enable control and coordination; financial and administrative services depend on secure digital platforms. Consequently, the Perimeter not only enumerates critical operators but also implicitly defines a topological map of interdependencies that govern national service continuity.

Beyond its legal and institutional function, the PSNC also constitutes a valuable *data framework*. The information collected through certification processes, incident reporting, and compliance audits provides a structured and periodically updated dataset describing the architecture of national critical infrastructures. When interpreted from a systemic perspective, these data can be repurposed to support quantitative analyses—such as the modelling of cascading failures, inter-sectoral risk propagation, and resilience evaluation. In this sense, the Perimeter bridges the gap between regulatory policy and technical modelling: it transforms legal definitions of criticality into empirical parameters that can inform simulation-based assessments.

This chapter leverages the PSNC as both a *governance framework* and a *systemic data source*. The following section translates the structural information derived from

the Perimeter into the *CISIApro 2.0* simulation environment, providing a quantitative representation of interdependencies among Italian critical infrastructures. This integration makes it possible to evaluate how cyber incidents can propagate through the national ecosystem and to identify the infrastructures that play the most critical roles in maintaining systemic stability.

15.2 System Model

The simulation framework adopted in this chapter builds upon the *CISIApro 2.0* agent-based environment introduced in Section 14.2 and detailed in [59], [60], [61], [62], [128]. It is instantiated on a representative subset of infrastructures included within the Italian *Perimetro di Sicurezza Nazionale Cibernetica* (PSNC), as defined by [326], [327], [328], [329], [330]. The goal of this model is to reproduce the main cyber–physical and organizational dependencies among critical sectors operating within the national cybersecurity boundary, thereby enabling a quantitative evaluation of cascading risk and resilience.

The selected configuration includes eight key operators, chosen to reflect the diversity of essential services and the heterogeneity of their interconnections:

- two **telecommunications providers**, representing the backbone of ICT connectivity and data exchange;
- two **electricity distribution companies**, ensuring energy supply to both industrial and ICT infrastructures;
- one **railway operator** and one **airline company**, representing the transport domain;
- one **bank**, representing the financial sector and its role in digital payment and transaction systems;
- and one **government department**, symbolizing the institutional and regulatory coordination layer.

This ensemble captures the main inter-sectoral dependencies that characterize the Italian digital ecosystem, as described by national and European frameworks on critical infrastructure protection [5], [6], [121].

Functional dependencies among these entities are formalized as *directed links*, as illustrated in Figure 15.1. These links represent the operational and informational relationships through which resources, data, or control signals flow between agents. Electricity distribution provides the physical energy necessary for ICT and operational technology components across all domains. Telecommunications ensure the cyber backbone enabling coordination, control, and data sharing. The banking sector supports financial transactions and payment services critical for transport operators and public administration. Finally, the government department represents the policy and governance node, establishing regulatory, supervisory, and coordination links with the other infrastructures. Together, these dependencies form a compact but realistic *interdependency graph*, reproducing national-scale service chains and systemic coupling mechanisms.

Each infrastructure is represented as an autonomous *agent* characterized by a set of internal state variables that describe both its operational level and cyber posture. Following the *Mixed Holistic–Reductionist* principle introduced in Chapter 14, agents can be modeled at different levels of granularity depending on the availability of information and the analytical objective. At the holistic level, operators are treated as aggregate entities exchanging high-level services (e.g., power, connectivity, finance). At the reductionist level, each operator can be decomposed into functional units (e.g., ICT systems, physical assets, management layers), allowing for finer-grained propagation dynamics and local resilience evaluation.

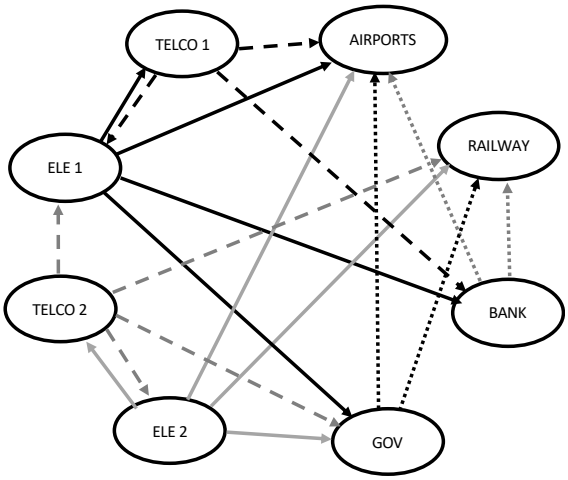


Figure 15.1: Interdependency graph across the eight infrastructures considered in the case study. Directed links denote functional and informational dependencies among sectors.

The operational state of each agent evolves over time according to the status of its dependencies and its internal resilience parameters. State transitions are governed by a set of equations describing the propagation of perturbations along dependency links, following the formalism of [61]. Disturbances—such as cyber incidents, technical failures, or natural hazards—are injected into the system as variations of the state variables associated with confidentiality, integrity, and availability (). These perturbations then propagate through the network based on dependency weights and propagation coefficients, producing cascading effects across domains. Recovery dynamics, redundancy mechanisms, and cyber risk indicators are also modeled to evaluate not only the extent of service degradation but also the capacity for restoration and containment.

By analysing the temporal evolution of the system-wide degradation patterns, the model enables a quantitative assessment of how localized cyber events (e.g., ransomware or DDoS attacks) may escalate into national-scale service disruptions. This approach allows decision-makers to identify critical propagation hubs, evaluate the systemic consequences of specific attack typologies, and test the effectiveness of redundancy or diversification strategies within the Italian National Cybersecurity Perimeter.

15.3 Cyber scenarios and results

To evaluate the systemic behaviour of the national infrastructure network under cyber stress, two representative threat classes are analysed: *ransomware* and *distributed denial-of-service* attacks. According to the ENISA Threat Landscape reports [341], [342], these categories consistently rank among the most impactful across sectors due to their disruptive potential and high frequency. Both scenarios are executed on the same dependency structure introduced in Section 15.2, allowing a direct comparison of propagation mechanisms and systemic sensitivity. While ransomware primarily targets data integrity and internal ICT availability, DDoS attacks propagate through service-level dependencies by impairing network connectivity and communication functions.

Ransomware on financial services

The first scenario models a ransomware attack against the banking sector, leading to the unavailability of payment services and digital transactions. The system-wide synoptic view in Figure 15.2 shows the immediate systemic footprint: the bank experiences a drastic drop in confidentiality, integrity, and availability (CIA) levels, which translates into a severe loss of operativity. Transport operators—specifically the airline and railway

companies—exhibit partial operational degradation, as their ticketing and payment systems rely on banking services. Telecommunications and energy infrastructures remain functionally available, yet they experience an increase in cyber-risk indicators due to potential reputational effects and latent exposure to lateral movement originating from the compromised financial ICT network.

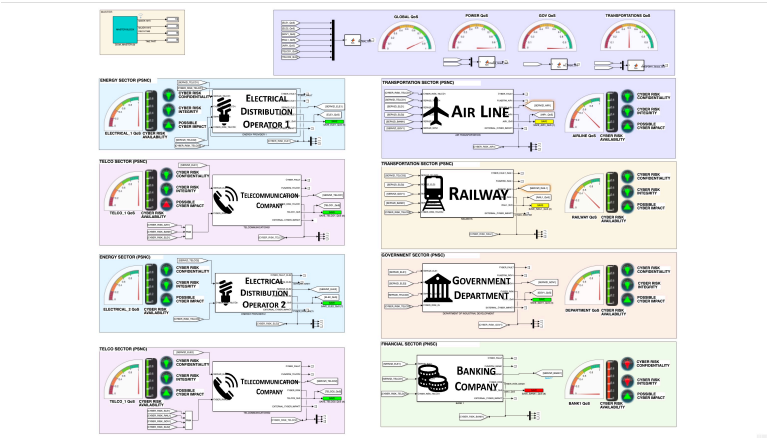


Figure 15.2: System-wide synoptic under ransomware on the banking sector. Icon backgrounds indicate operational levels, vertical CIA bars represent confidentiality, integrity, and availability states, and overlay markers denote incoming cyber-risk indicators from dependent sectors.

A closer inspection of the affected bank (Figure 15.3) highlights the central role of shared ICT assets: the compromise of a single internal service rapidly cascades across multiple business functions, creating a single point of failure that undermines the entire company's operational capacity. The impacts propagate asymmetrically to dependent sectors. In the airline domain (Figure 15.4), degradation concentrates on sales and ticketing systems, whereas in the railway sector (Figure 15.5), payment-related services and customer interfaces are primarily affected. Telecommunications (Figure 15.6) maintain normal functionality but display heightened cyber-risk due to the potential for cross-sectoral infection vectors.

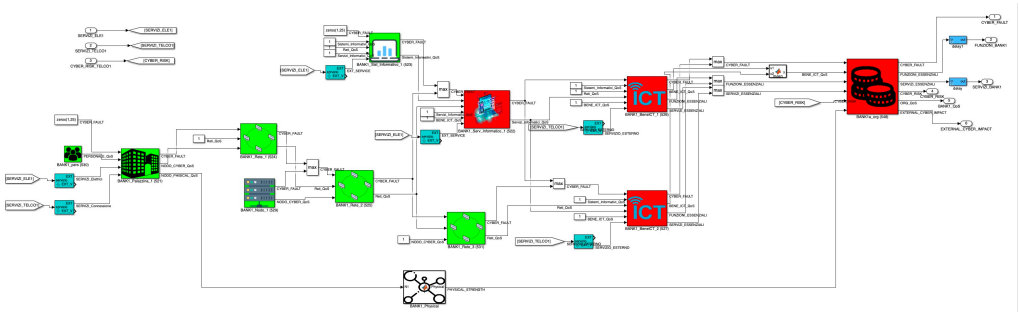


Figure 15.3: Bank view under ransomware. The compromise of a shared ICT service acts as a single point of failure, triggering cascading degradation across multiple internal functions.

DDoS on telecommunications

The second scenario simulates a large-scale DDoS campaign targeting one of the two telecom operators. This class of attack primarily compromises service availability by exhausting network and server resources, with immediate repercussions on all dependent sectors. The system-wide synoptic in Figure 15.7 reveals a markedly different propagation

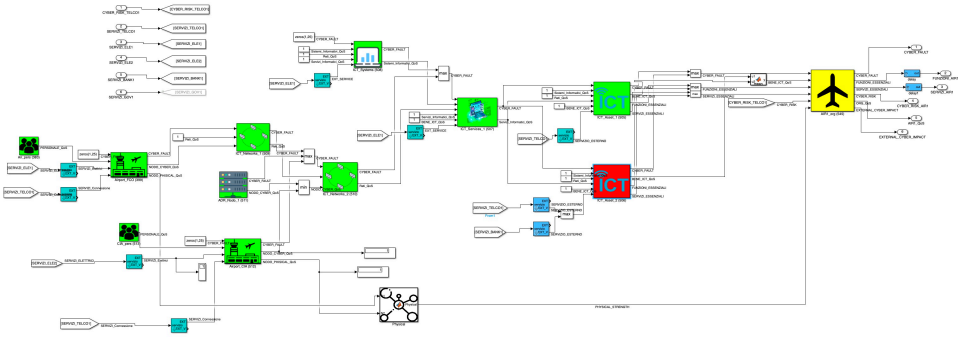


Figure 15.4: Airline operator under ransomware. Service degradation concentrates on sales and ticketing functions that depend on disrupted payment systems.

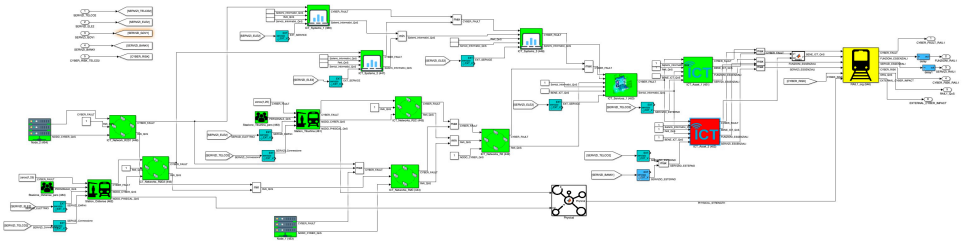


Figure 15.5: Railway operator under ransomware. Degradation affects customer and payment services that depend on disrupted financial transactions.

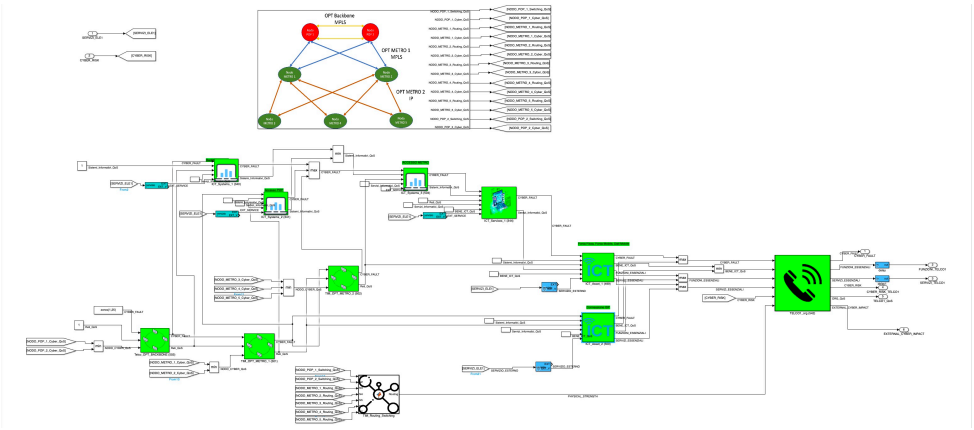


Figure 15.6: Telecommunications operator under ransomware. No direct service outage occurs, but cyber-risk exposure increases due to potential lateral movement across interconnected networks.

pattern compared to the ransomware case: degradation spreads rapidly through ICT connectivity links, affecting every operator that depends on telecom services. The severity of the impact varies according to redundancy and multi-homing configurations, illustrating the protective role of diversified connectivity.

At the operator level, the attacked telecom agent (Figure 15.8) experiences widespread degradation across its service portfolio, consistent with volumetric attacks affecting both backbone and customer-facing functions. The railway operator (Figure 15.9) suffers the sharpest degradation in stations relying on a single telecom provider, particularly where ICT coordination bottlenecks exist. Government services (Figure 15.10) display a reduction in availability for online functions and ICT assets, while energy distribution operators (Figure 15.11) show asymmetric outcomes: the company with dual telecom links maintains operativity, whereas the single-homed one faces severe service disruption.

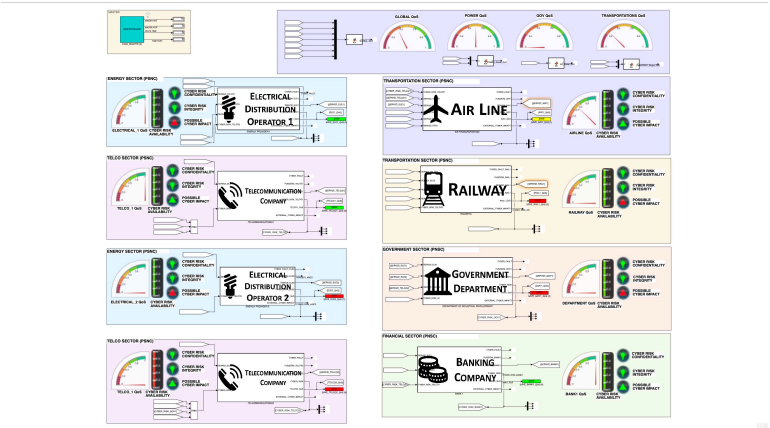


Figure 15.7: System-wide synoptic under DDoS on telecommunications. Availability impairments propagate through ICT dependencies, affecting all interconnected sectors.

Airline operations (Figure 15.12) also experience degradation, especially in customer service and airport ICT domains dependent on continuous network availability.

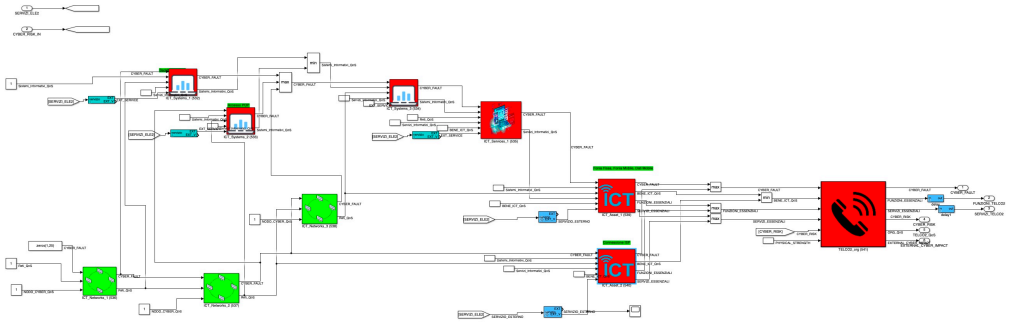


Figure 15.8: Telecommunications operator under DDoS. Widespread degradation occurs across multiple service lines due to bandwidth exhaustion and network congestion.

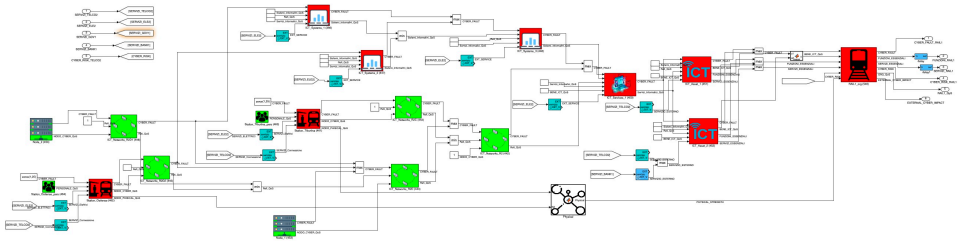


Figure 15.9: Railway operator: significant degradation at stations with single-provider connectivity and ICT coordination bottlenecks.

15.4 Discussion

The Italian National Cybersecurity Perimeter provides a unique bridge between regulation and modelling. Although originally conceived as a compliance and oversight instrument, its structure, taxonomy, and reporting obligations implicitly define a systemic map of the national digital ecosystem. By translating this information into a simulation framework, it becomes possible to move from qualitative policy requirements to quantitative analysis of interdependencies, cascading effects, and sectoral vulnerabilities.

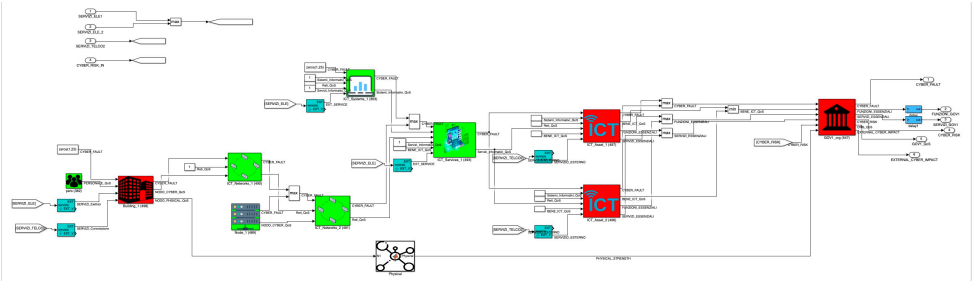


Figure 15.10: Government department: online functions and ICT assets degrade due to telecom unavailability.

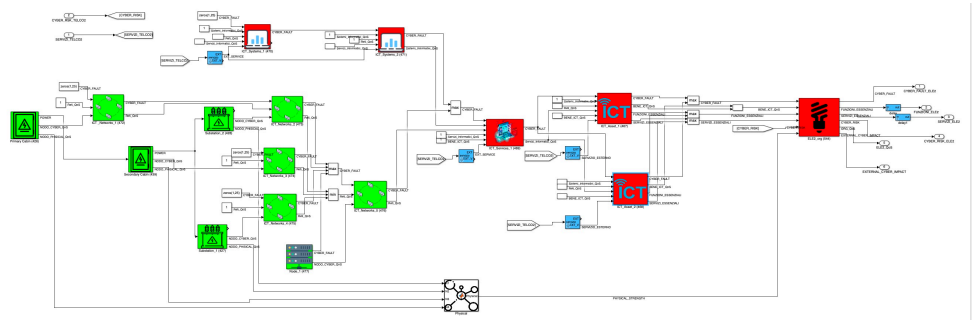


Figure 15.11: Electricity distribution: asymmetric impact driven by single-homing versus dual-provider redundancy.

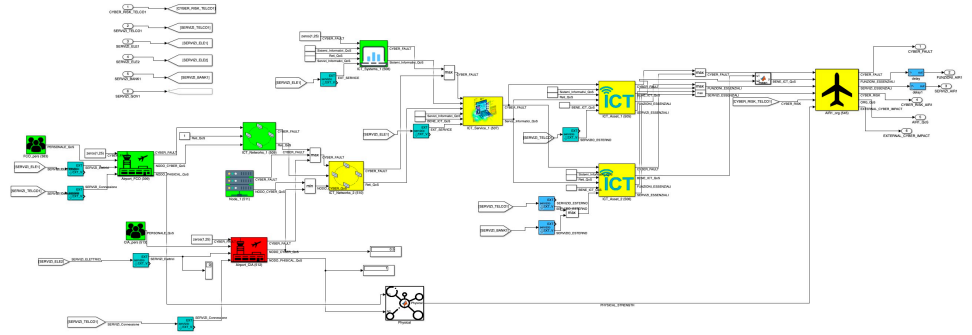


Figure 15.12: Airline operator: ICT disruption propagates to customer services and airport operations through network dependency.

The scenarios developed in this chapter demonstrate the added value of this approach. The ransomware attack on the financial sector produces localised but intense cascades, primarily affecting digital payment chains and customer-facing functions. Conversely, the DDoS scenario on telecommunications generates rapid, wide-ranging availability degradation across multiple sectors, reflecting the pervasive dependence of modern infrastructures on ICT connectivity. Together, these results highlight the systemic centrality of telecommunications and energy as propagation hubs within the Perimeter, and the asymmetric nature of cyber risk across domains.

From a governance perspective, such simulation-based analysis offers practical support for national resilience planning. It enables the identification of critical operators whose compromise would have disproportionate systemic consequences, as well as the assessment of redundancy and diversification strategies that can mitigate these effects. In this sense, the Perimeter's regulatory data—when interpreted through a systemic lens—can inform investment priorities, risk communication, and coordinated response planning across government and industry.

At the same time, the study exposes inherent challenges in translating policy-driven datasets into quantitative models. Incomplete information, heterogeneous data quality, and limited visibility into inter-organisational dependencies introduce significant *epistemic uncertainty*—a form of uncertainty that cannot be reduced by probabilistic means alone. Addressing this issue requires new methodological tools capable of representing and propagating imprecision within the model, an aspect that motivates the possibilistic extension developed in Chapters 16- 17.

Chapter 16

Managing Uncertainty in Interdependency Models: A Possibilistic Extension of CISIApro 2.0

Quantitative models of interdependent Critical Infrastructures have become essential tools for assessing cascading effects and systemic vulnerabilities across complex cyber–physical systems [5], [6], [343]. Their analytical power, however, depends critically on the reliability and completeness of the input data describing dependency strengths, recovery dynamics, and cross-sector couplings. In practice, such parameters are often uncertain: they derive from expert judgement, partial observations, heterogeneous datasets, or classified information with limited accessibility. This incomplete knowledge introduces a form of *epistemic uncertainty*, distinct from random variability (*aleatoric uncertainty*), as it reflects ignorance rather than intrinsic randomness [139], [336], [337]. Traditional probabilistic models struggle to represent this condition, since they require assigning precise probability distributions even when the underlying evidence is fragmentary or qualitative.

This chapter addresses this limitation by extending the interdependency modelling framework introduced in Chapter 14 with a *possibilistic representation of uncertainty*. Building upon the *CISIApro 2.0* environment [59], [60], [61], [128], the proposed approach replaces strict probabilistic assumptions with the dual formalism of *possibility* and *necessity* measures [338], [340]. This framework allows incomplete or imprecise information to be propagated through dependency networks while respecting the *least-commitment principle*—that is, without imposing artificial precision. Each simulated outcome is thus bounded by two complementary quantities: a conservative lower limit (*necessity*), indicating what is almost certainly true, and an optimistic upper limit (*possibility*), expressing what remains plausible given the available evidence. The resulting *uncertainty band* provides a transparent view of confidence levels within the model, highlighting where robust knowledge supports the analysis and where additional data or redundancy are needed.

The possibilistic extension preserves the agent-based, multilayer architecture of CISIApro while introducing an additional uncertainty layer on top of the existing propagation dynamics. Agents—representing infrastructures or organisational entities—retain their local dependency structures and resilience parameters, but their states are now described by intervals reflecting epistemic confidence. This enhancement allows decision-makers to interpret systemic risk not as a single deterministic trajectory, but as a bounded region of plausible outcomes, improving both analytical transparency and governance relevance.

The methodology is demonstrated through a national-scale multi-sector scenario encompassing energy, telecommunications, transport, finance, and public administration. By introducing uncertainty in dependency strengths and recovery rates, the simulation explores how epistemic gaps spread across domains, how redundancy mitigates uncertainty propagation, and how the necessity–possibility gap can guide resilience enhancement and data collection priorities. The proposed approach thus bridges the gap between quantitative simulation and expert-based reasoning, contributing to the development of uncertainty-aware tools for critical infrastructure protection and policy support.

Organization of the Chapter. The remainder of this chapter is structured as follows:

- **Section 16.1** introduces the system model and describes the configuration of the interdependency network used in the simulations. It explains how epistemic uncertainty is represented through possibility distributions and how it is propagated among agents.
- **Section 16.2** presents the mathematical background of possibility theory, defining possibility and necessity measures, the least-commitment principle, and their integration into the CISIApro consensus dynamics.
- **Section 16.3** reports the simulation results, illustrating how epistemic uncertainty affects the propagation of cyber-induced disturbances and how the resulting uncertainty bands can be interpreted.
- **Section 16.4** discusses the methodological and practical implications of the possibilistic extension, emphasizing its benefits for uncertainty-aware systemic risk analysis and decision support.

16.1 System Model

The possibilistic extension developed in this chapter builds directly upon the interdependency framework introduced in Chapters 14 and 15, where Critical Infrastructures are represented as autonomous *agents* exchanging resources, information, and control signals through directed dependency links [59], [60], [61], [128]. Each agent represents a distinct infrastructure operator or subsystem whose operational state evolves in time according to both internal dynamics and the condition of its dependencies. This agent-based formalism—central to the *CISIApro 2.0* simulation environment—enables the explicit representation of cross-sector interactions and the analysis of cascading failures within complex, interdependent systems.

In its original formulation, CISIApro uses deterministic or probabilistic parameters to describe key aspects of interdependency, such as dependency strength, recovery time, or propagation delay. However, many of these parameters are affected by *epistemic uncertainty*, arising from incomplete knowledge, heterogeneous data quality, or confidentiality constraints. To account for this, the present extension replaces single-valued parameters with *possibility distributions*, which represent ranges of plausible values rather than precise numerical estimates. This approach allows the model to propagate incomplete or qualitative information through the network without imposing arbitrary probability assignments, thereby offering a more transparent and flexible treatment of uncertainty [338], [340].

At each simulation step, the state of every agent is characterised by a pair of complementary measures: *necessity* and *possibility*. The necessity measure provides a conservative lower bound of operativity—what is almost certainly true given available evidence—while the possibility measure represents an optimistic upper bound, denoting what remains plausible. Together, they form an *uncertainty interval* that captures the confidence level associated with the agent's operational status. When multiple dependencies converge on the same node, these measures are aggregated according to the *least-commitment principle*: plausibility accumulates through the *maximum* operator (reflecting the union of possible evidence), whereas necessity is limited by the *minimum* operator (representing the constraint imposed by the least certain input). This formulation ensures that uncertainty propagates coherently across interdependent infrastructures while preserving analytical conservatism.

The interdependency network adopted for the simulations mirrors the one defined in Chapter 15 (see Figure 15.1). It comprises eight representative operators drawn from the national Cybersecurity Perimeter: two telecommunications providers, two electricity distribution companies, a railway operator, an airline, a bank, and a government department. These entities collectively represent the essential pillars of the national

cyber–physical ecosystem—energy, communications, transport, finance, and public administration—and the functional dependencies that connect them. As previously described, electricity ensures the power supply to ICT and operational technology systems; telecommunications provide the data backbone enabling coordination and control; financial services sustain payment and transaction chains; and the government acts as a supervisory and coordination node. This network thus provides a realistic and compact reference structure for analysing the propagation of uncertainty across multiple interdependent domains.

Each simulation run begins with a triggering event—such as a degradation in power supply or a disruption in communications—and traces how its effects propagate through the dependency network under uncertain conditions. Unlike deterministic or probabilistic approaches, the possibilistic formulation produces a range of outcomes for each infrastructure: the *necessity* trajectory corresponds to robustly supported degradation, while the *possibility* trajectory captures degradation that is plausible but not certain. The distance between the two represents the *necessity–possibility band*, which quantifies both the expected impact and the degree of epistemic confidence in that estimate. This dual-layer view allows analysts to identify which infrastructures are certainly affected and where uncertainty dominates, thus guiding data improvement, redundancy design, and targeted resilience measures.

To visualise the overall system state, the CISIApro 2.0 environment provides an *uncertainty-aware synoptic dashboard*. As shown in Figure 16.1, each infrastructure agent is represented by an icon whose shading reflects its level of operativity, while the associated CIA (Confidentiality–Integrity–Availability) bars display the spread between necessity and possibility. Darker tones correspond to higher operativity, whereas wider CIA bars indicate lower confidence or greater uncertainty in the assessment. This graphical interface supports a rapid and intuitive understanding of both systemic degradation and the epistemic reliability of the underlying results.

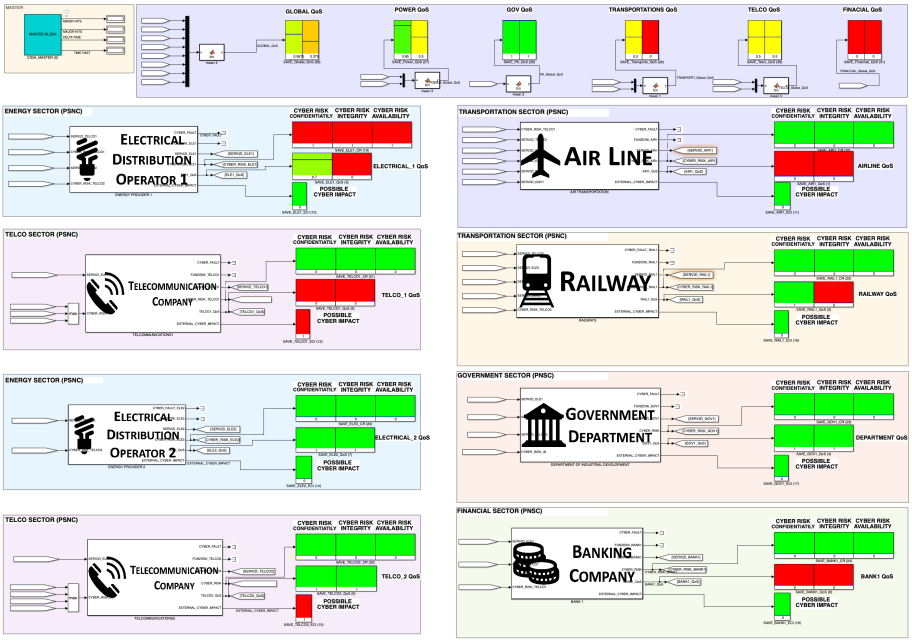


Figure 16.1: Uncertainty-aware synoptic dashboard in CISIApro 2.0. Darker shades indicate higher operativity levels, while the width of the CIA bars represents the necessity–possibility confidence interval associated with each organisation.

16.2 Mathematical background

The possibilistic extension of CISIApro relies on the formal apparatus of *possibility theory*, a mathematical framework designed to represent incomplete or imprecise knowledge. Unlike probability theory, which quantifies the frequency or likelihood of events, possibility theory captures *degrees of plausibility* and *degrees of certainty* based on the information available. This makes it particularly suitable for modelling epistemic uncertainty—situations where data are scarce, qualitative, or conflicting, as often occurs in cyber–physical risk assessment.

Possibility and necessity measures

Let X denote the universal set of possible system states, and let $\mathcal{P}(X)$ be its powerset. A *possibility measure* $Pos : \mathcal{P}(X) \to [0, 1]$ assigns to each event $A \subseteq X$ the degree to which it is *compatible* with the available knowledge. Its definition follows a set of axioms that ensure internal consistency:

$$Pos(\emptyset) = 0, \qquad Pos(X) = 1, \tag{16.1}$$

together with monotonicity ($A \subseteq B \Rightarrow Pos(A) \leq Pos(B)$), continuity for increasing sequences of sets, and the *maxitivity* property:

$$Pos\left(\bigcup_{i \in I} A_i\right) = \sup_{i \in I} Pos(A_i). \tag{16.2}$$

Intuitively, if any of the events A_i is highly possible, their union is at least equally possible.

Dually, the *necessity measure* $Nec : \mathcal{P}(X) \to [0, 1]$ quantifies the degree to which an event is *certainly true* given the available evidence. It satisfies the dual axioms:

$$Nec(\emptyset) = 0, \qquad Nec(X) = 1, \tag{16.3}$$

and the properties of monotonicity, continuity for decreasing sequences, and *minitivity*:

$$Nec\left(\bigcap_{i \in I} A_i\right) = \inf_{i \in I} Nec(A_i). \tag{16.4}$$

The two measures are intrinsically related. For any event A, $Nec(A)$ represents a conservative lower bound on the confidence in A, while $Pos(A)$ provides an optimistic upper bound:

$$Nec(A) \leq Pos(A), \tag{16.5}$$

$$Pos(A \cap B) \leq \min\{Pos(A), Pos(B)\}, \tag{16.6}$$

$$Nec(A \cap B) \geq \max\{0,\ Nec(A) + Nec(B) - 1\}. \tag{16.7}$$

Operationally, $Nec(\cdot)$ corresponds to the *robustness* or lower-probability bound of a statement—what is almost certainly true—while $Pos(\cdot)$ captures its *plausibility*—what could reasonably hold given incomplete information. The interval $[Nec(A),\ Pos(A)]$ thus encodes both the range and the confidence of uncertain statements, distinguishing reliable outcomes from merely plausible ones.

Least-commitment combination and agent consensus

When several uncertain pieces of evidence refer to the same event, their combination must preserve consistency without introducing unwarranted precision. This idea is formalised by the *least-commitment principle*, which prescribes using the most conservative operators compatible with available knowledge. Accordingly, conjunctive information (common

evidence supporting the same conclusion) is combined using the pointwise minimum, while disjunctive information (alternative sources) uses the pointwise maximum.

For two uncertain statements A_1 and A_2 with associated measures (Pos_1, Nec_1) and (Pos_2, Nec_2), the combined possibility and necessity are given by:

$$Pos_{12}(A) = \sup_{x \in A} \min\{Pos_1(x), Pos_2(x)\}, \qquad Nec_{12}(A) = \inf_{x \in A} \min\{Nec_1(x), Nec_2(x)\}. \tag{16.8}$$

At the network level, this translates into an *uncertainty-aware consensus mechanism* for agent-based systems. Let $x_i(k)$ denote the uncertain state of agent i at communication step k, and $\mathcal{J}_i$ its set of incoming neighbours. The update rules implemented in CISIApro are:

$$Pos\big(x_i(k{+}1)\big) = \max_{j \in \mathcal{J}_i} Pos\big(x_j(k)\big), \tag{16.9}$$

$$Nec\big(x_i(k{+}1)\big) = \min_{j \in \mathcal{J}_i} Nec\big(x_j(k)\big). \tag{16.10}$$

These operators convey an intuitive interpretation: plausibility accumulates through multiple incoming dependencies (optimistic aggregation), while necessity is constrained by the least certain input (pessimistic aggregation). This mechanism ensures that uncertainty propagates coherently through the dependency network, maintaining both upper and lower confidence bounds on each agent's state.

Multilayer graph and node dynamics

The resulting system can be represented as a multilayer directed graph

$$M = (\mathcal{V}_M, \mathcal{E}_M, \mathcal{V}, \mathbf{L}),$$

where each layer $\mathcal{L}_a \in \mathbf{L}$ corresponds to a domain (e.g., energy, telecommunications, finance), and each node-layer pair $(u, \boldsymbol{\alpha})$ represents an active infrastructure component or organisational unit. Each node maintains a status vector $x_u(t)$, which evolves over discrete time steps according to its internal dynamics, the states of its neighbours, and exogenous disturbances:

$$\begin{cases} x_u(t{+}1) = g_u\big(x_u(t),\, y_{\Gamma^+(u,\mathbf{L})}(t),\, z_u(t)\big), \\ \quad y_u(t) = h_u\big(x_u(t),\, z_u(t)\big), \end{cases} \tag{16.11}$$

where the incoming neighbourhood $\Gamma^+(u, \mathbf{L})$ is defined as

$$\Gamma^+(u, \mathbf{L}) = \big\{\, v \in \mathcal{V}_M \;\big|\; \exists\, \boldsymbol{\beta}, \boldsymbol{\alpha} \in \mathbf{L} : ((v, \boldsymbol{\beta}), (u, \boldsymbol{\alpha})) \in \mathcal{E}_M \,\big\}. \tag{16.12}$$

In the possibilistic extension, all variables and inputs appearing in g_u and h_u are lifted to (Pos, Nec) pairs and propagated according to the consensus rules (16.9)–(16.10). This allows the model to capture not only how disruptions spread through physical and logical links, but also how the degree of epistemic confidence associated with each component evolves over time. As a result, the simulation produces a dual-layer view of systemic risk—quantifying both the extent of degradation and the reliability of the underlying knowledge.

16.3 Results

The results of the possibilistic simulations illustrate how epistemic uncertainty affects the propagation of disruptions across interdependent infrastructures. Two representative triggering events are analysed. The first corresponds to an availability-oriented disturbance, modelled as a synthetic DDoS-like degradation in telecommunications services. The second represents a ransomware-like impairment affecting a key software component in

the power domain. For each organisation, the outcomes are reported in terms of *necessity* and *possibility* of operativity, visualised through dark (high) and light (low) shading, respectively. The gap between the two surfaces—the *uncertainty band*—quantifies how incomplete knowledge about redundancy, failover strategies, or inter-domain coupling translates into confidence variability in the predicted impact.

Electrical distribution operator (2). Figures 16.2–16.3 present the results for the second power distribution company under a cyber-induced impairment in its supervisory software. The necessity map (Figure 16.2) shows that most internal components retain a robust level of operativity, except for those directly connected to the affected control layer. The possibility map (Figure 16.3), on the other hand, reveals a wider range of plausible degradation, extending toward downstream nodes and communication interfaces with the telecommunications domain. This divergence reflects uncertainty in cross-domain dependencies—specifically, the limited knowledge about the reliability of communication channels linking control centres and substations.

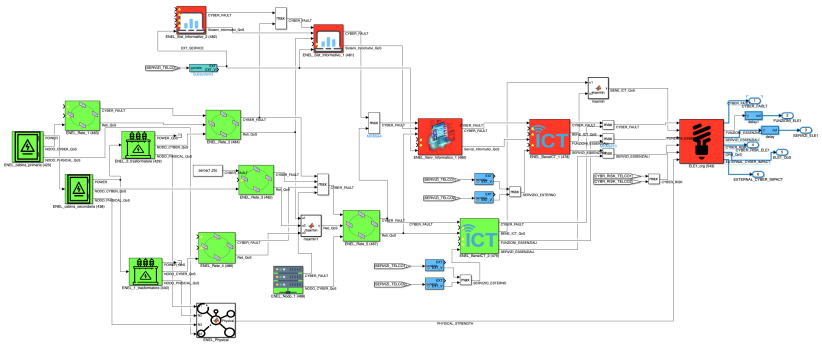


Figure 16.2: Necessity measure for agents in Electricity Distribution Company 2. Darker areas indicate high operativity levels that are *necessarily* maintained, i.e., robust under epistemic uncertainty.

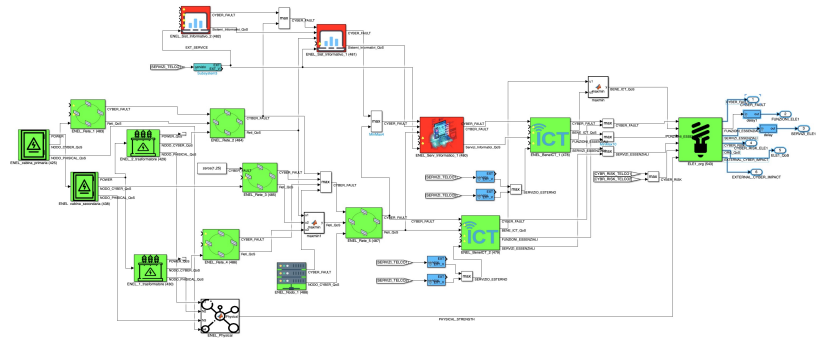


Figure 16.3: Possibility measure for agents in Electricity Distribution Company 2. Darker areas indicate high operativity levels that are *possibly* maintained, i.e., plausible under incomplete or uncertain knowledge.

Telecommunications operator (1). Figures 16.4–16.5 illustrate the effect of an upstream power disturbance on the telecommunications network. Here, the necessity and possibility maps are relatively close, indicating limited epistemic uncertainty for most nodes. This tight band suggests that a subset of telecom agents—particularly those with single power feeds—are almost certainly affected, whereas others benefit from known redundancy and autonomous backup power. The small remaining discrepancies are associated with uncertain failover conditions and incomplete data on energy storage capacity and switching delays.

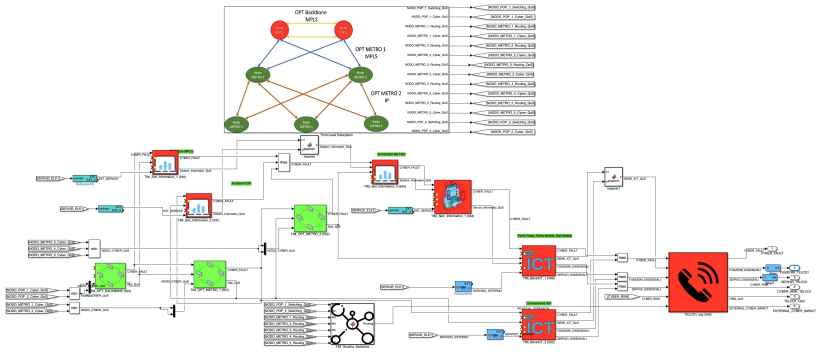


Figure 16.4: Necessity measure for the telecommunications network. Darker areas represent degradation that is *robustly* expected under upstream power impairment, i.e., necessarily implied by the interdependency structure.

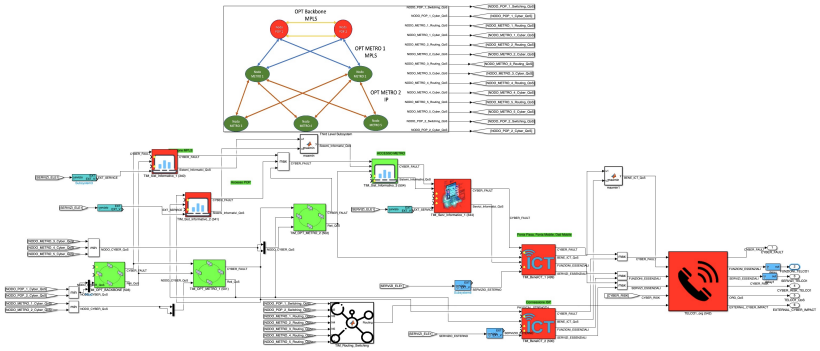


Figure 16.5: Possibility measure for the telecommunications network. Darker areas represent degradation that is *plausible* when uncertain spillovers and failover conditions are considered, i.e., potentially induced under incomplete or ambiguous information.

Airline company. The airline case (Figures 16.6–16.7) exhibits one of the widest uncertainty bands in the simulation. Here, operativity depends jointly on airport ICT services and external connectivity to telecommunication and financial systems. While several components show high plausibility of degradation (light zones in Figure 16.7), only a subset is necessarily affected (Figure 16.6), primarily those tied to ground operations and flight coordination. The wide spread between the two measures highlights the role of partial information on redundancy and unclear failover behaviour in complex multi-stakeholder environments such as aviation networks.

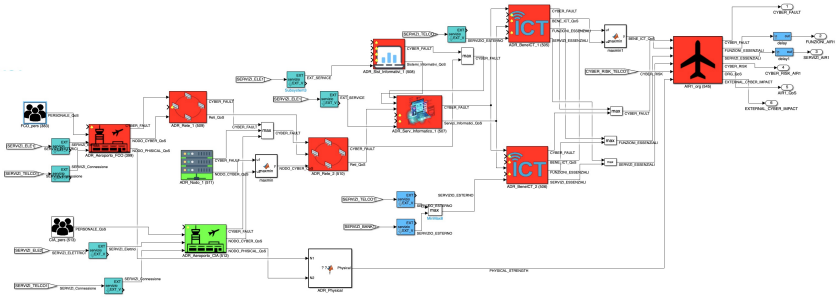


Figure 16.6: Necessity measure for the airline sector. Darker areas highlight agents whose degradation is *robustly* supported by available evidence, i.e., necessarily implied by upstream or structural dependencies.

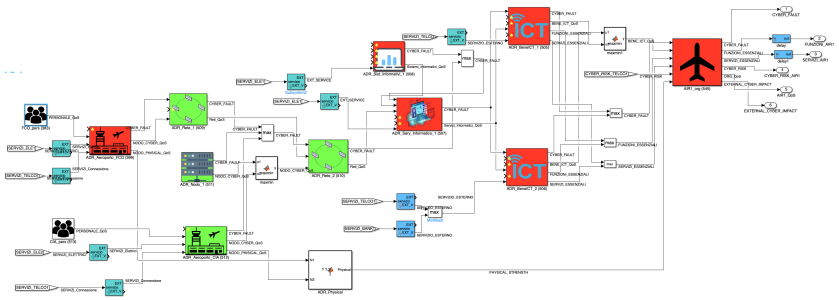


Figure 16.7: Airline: possibility measure indicating a broader set of *plausible* degradations driven by uncertain cross-sector dependencies.

Railway company. Figures 16.8–16.9 report the results for the railway sector. Degradations concentrate on coordination centres and stations, where dependency on telecommunications and power is most critical. The necessity map identifies a limited number of nodes as certainly impacted, while the possibility map shows a wider pattern of plausible disruptions, particularly in peripheral stations. This outcome aligns with operational heterogeneity: some sites are dual-homed to multiple providers, while others depend on single connections, creating uncertainty in continuity of service.

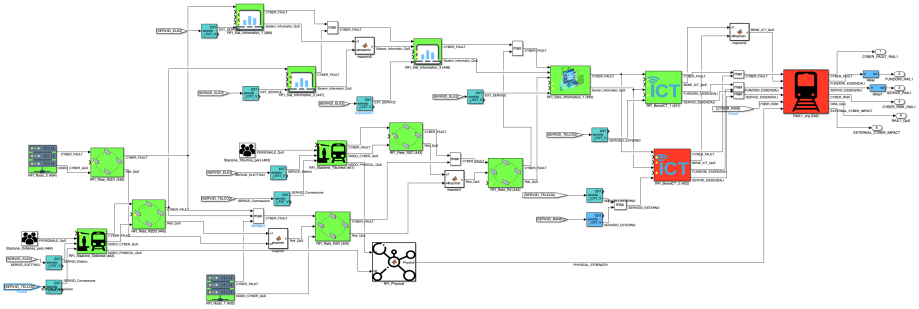


Figure 16.8: Railway: necessity of degradation concentrated on coordination assets and single-homed stations.

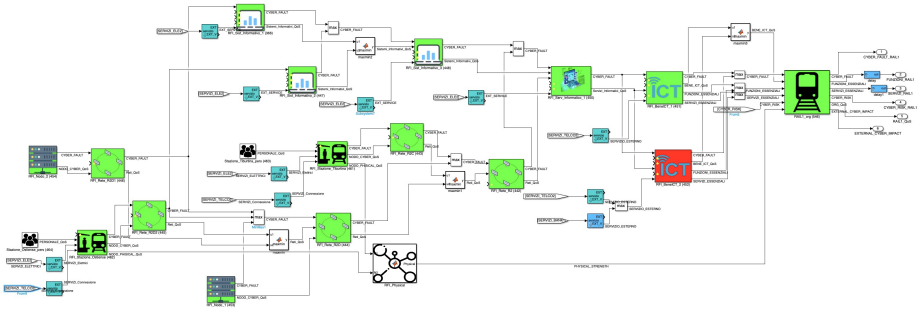


Figure 16.9: Railway: possibility map showing wider plausible degradation in areas with uncertain redundancy.

Bank and Government department. Finally, Figures 16.10–16.11 and 16.12 summarise the outcomes for the financial and public administration sectors. In the banking system, core back-end services display narrow uncertainty bands—consistent with well-instrumented failover and high monitoring coverage—whereas customer-facing services show a wider band, reflecting incomplete visibility of third-party dependencies. The

government department exhibits high overall robustness, with necessity and possibility values both near unity across most functions. Only online public services reveal noticeable uncertainty, linked to partial knowledge about their hosting and connectivity arrangements.

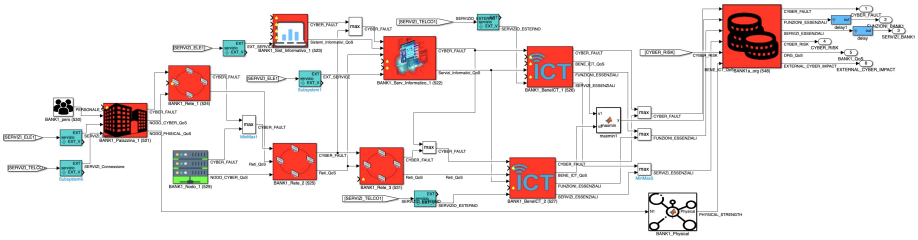


Figure 16.10: Bank: necessity of operativity across sub-services; uncertainty appears primarily at customer-facing interfaces.

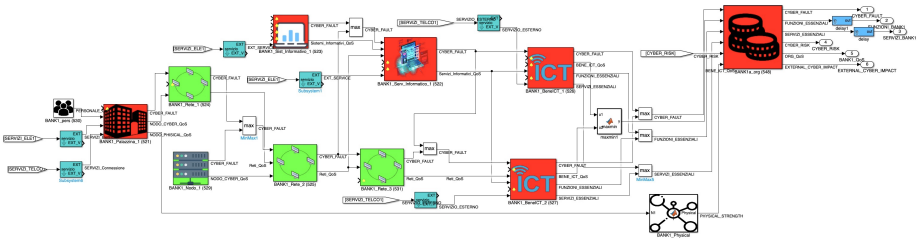


Figure 16.11: Bank: possibility measure revealing a broader range of plausible degradations where telemetry or dependency data are incomplete.

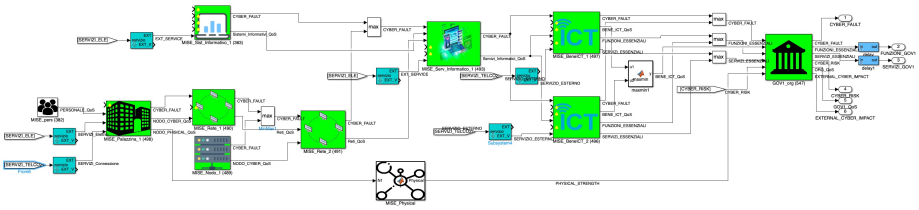


Figure 16.12: Government department: high necessity of operativity due to dual energy and ICT feeds; uncertainty is limited to online services.

16.4 Discussion

The possibilistic extension developed in this chapter equips CISIApro with an explicit treatment of epistemic uncertainty, aligning the simulation framework with the practical realities of cyber–physical risk assessment. Unlike probabilistic methods, which require precise statistical inputs, the possibilistic formulation accommodates incomplete, qualitative, or conflicting information—conditions that frequently arise due to scarce data, confidentiality constraints, and heterogeneous sources of evidence. By representing each variable through a pair of complementary measures—*necessity* and *possibility*—the model distinguishes between what is *almost certain* and what is merely *plausible*, thereby providing a transparent confidence interval around each simulated outcome.

The simulation results demonstrate that the proposed approach captures meaningful patterns of epistemic uncertainty across interdependent infrastructures. Highly central dependencies—such as electricity supply and telecommunications connectivity—tend to exhibit *tight necessity–possibility bands*, reflecting well-understood and consistently modelled interactions. Conversely, cross-domain couplings that involve loosely specified

failover mechanisms or limited monitoring information produce *wider bands*, revealing where knowledge gaps dominate system behaviour. These findings highlight that uncertainty propagation does not simply mirror the structure of interdependencies, but also depends critically on the *quality and completeness of information* available for each link.

From a methodological standpoint, the extension preserves the semantics and modularity of the original CISIApro framework. The core agent-based dynamics and propagation functions remain unchanged; the enhancement consists in lifting each state variable to a (Pos, Nec) pair and applying the max/min consensus rules derived from the least-commitment principle. This design choice ensures full backward compatibility with existing models, while endowing the simulator with greater expressive power and interpretability. Moreover, the computational cost remains modest, as the operations involved are algebraically simple and parallelisable across agents.

From a governance and decision-making perspective, the uncertainty-aware outputs offer actionable insights for both operators and policymakers. Assets or functions exhibiting high *Nec*-level impacts correspond to effects that are robustly supported by available evidence and therefore call for immediate mitigation or protection measures. Conversely, assets characterised by a wide gap between *Pos* and *Nec* values indicate situations dominated by epistemic uncertainty—where additional data collection, audits, or exercises would most effectively improve resilience. In this sense, the possibilistic framework not only quantifies risk but also guides the prioritisation of information improvement and redundancy design efforts.

Overall, the possibilistic extension of CISIApro bridges the gap between data-driven simulation and expert-based reasoning, providing a coherent formalism to reason about uncertainty in complex interdependent systems. It enhances analytical transparency, prevents overconfidence in single-point predictions, and helps decision-makers distinguish between system behaviours that are well-supported and those that remain conjectural. This enriched perspective constitutes a crucial step toward more robust and explainable modelling of critical infrastructure interdependencies under incomplete knowledge.

Chapter 17

Fuzzy Modelling of Cyber–Physical Interdependencies: Toward Hybrid Possibilistic Reasoning in CISIApro 2.0

While the possibilistic extension presented in Chapter 16 introduced the explicit representation of epistemic uncertainty in interdependency models, real-world decision processes often involve an even broader spectrum of vagueness and subjectivity. In particular, many parameters describing cross-sector dependencies, resilience thresholds, or recovery behaviours cannot be easily quantified or bounded within deterministic intervals. They arise from expert assessments, linguistic evaluations (e.g., "high dependency," "moderate redundancy"), or partially conflicting evidence gathered from multiple sources. To address such cases, it is necessary to move beyond pure possibility theory and adopt a formalism that can capture the gradual and context-dependent nature of human reasoning.

This chapter therefore extends the CISIApro 2.0 framework toward a *fuzzy representation of interdependencies*, providing a unified environment that combines deterministic, possibilistic, and fuzzy reasoning. The fuzzy formalism enables analysts to represent imprecise information as *graded memberships* rather than binary or interval-based values. For example, instead of defining a dependency strength as either "strong" or "weak," it can be expressed through a continuous membership function that reflects partial degrees of belief. This approach preserves the interpretability of qualitative expert knowledge while allowing its consistent integration into the mathematical machinery of the simulation. The resulting framework thus supports a more nuanced and human-consistent interpretation of uncertainty in large-scale cyber–physical systems.

Building upon the possibilistic logic introduced in the previous chapter, this fuzzy extension incorporates the theory of fuzzy sets and fuzzy measures [344], [345], [346], [347], [348]. It employs triangular or trapezoidal membership functions to represent dependency parameters, recovery rates, and confidence levels, and uses *fuzzy aggregation operators* to combine them across the interdependency network. By doing so, CISIApro becomes capable of simulating both quantitative data and qualitative expert inputs within the same model, enabling a richer exploration of resilience, cascading effects, and systemic uncertainty.

From a methodological perspective, this hybrid possibilistic–fuzzy formulation aligns CISIApro with current trends in uncertainty modelling and decision support [339], [340]. It provides the theoretical foundation for integrating human expertise, incomplete data, and computational simulation within a single, coherent analytical framework. In practical terms, this enhancement allows national and sectoral authorities to conduct scenario analyses even when empirical data are scarce or when critical information is available only in qualitative form—such as "high likelihood of propagation," or "limited redundancy."

Organization of the Chapter. The remainder of this chapter is organized as follows:

- **Section 17.1** introduces the modelling approach adopted for the fuzzy extension of *CISIApro 2.0*. It explains how fuzzy logic and fuzzy set theory are embedded within the agent-based structure of the simulator, detailing the use of triangular fuzzy numbers to represent uncertainty in confidentiality, integrity, and availability (CIA) dimensions.

- **Section 17.2** presents the mathematical background of fuzzy sets and triangular fuzzy numbers (TFNs). It formalises the membership functions, aggregation, and defuzzification operators used to model and propagate qualitative uncertainty across interdependent infrastructures.
- **Section 17.3** applies the fuzzy framework to the same simulated configuration introduced in Chapter 15. It analyses two representative use cases—an availability-oriented disturbance in the energy sector and a confidentiality breach in public administration—demonstrating how fuzzy reasoning enhances the interpretation of cascading effects and uncertainty propagation.
- **Section 17.4** discusses the main insights derived from the fuzzy simulations, evaluating their methodological relevance and practical implications for uncertainty-aware systemic risk analysis within *CISIApro 2.0*.

17.1 Modelling Approach

The fuzzy extension proposed in this chapter builds directly upon the CISIApro 2.0 framework, which represents interdependent infrastructures as a network of autonomous *agents*. Each agent corresponds to a system component or organisational function and maintains a set of state variables that quantify its levels of *confidentiality*, *integrity*, and *availability* (CIA). Directed links encode functional dependencies—such as electricity supply, ICT connectivity, or service provision—and determine how perturbations propagate across domains.

In the baseline CISIApro model, CIA dimensions are expressed as crisp numerical values in $[0, 1]$ and aggregated through weighted linear combinations. While suitable for deterministic analysis, this representation implicitly assumes complete and precise information. In practice, however, CIA evaluations are often qualitative or uncertain, derived from expert judgment, incomplete evidence, or heterogeneous data sources. To represent such epistemic uncertainty explicitly, we replace crisp CIA scores with *Triangular Fuzzy Numbers* and propagate them consistently through the network using fuzzy min–max calculus.

Within each agent: 1. Each CIA dimension is expressed as a TFN $\tilde{X} = (a, b, c)$, where the interval $[a, c]$ captures the plausible range of values and b represents the most likely estimate. 2. A convex fuzzy aggregation combines the three fuzzy CIA components into a single *cyber–operational risk* TFN $\tilde{R}$. 3. Dependencies among agents are treated according to their semantics: - *Series-like* (mandatory) dependencies use elementwise min (fuzzy intersection); - *Parallel-like* (redundant) dependencies use elementwise max (fuzzy union).

This design preserves the modular agent structure and semantics of CISIApro 2.0 while introducing a principled mechanism for handling qualitative or uncertain inputs without additional hyperparameters.

17.2 Mathematical Background for Fuzzy Sets

A fuzzy set $\tilde{A}$ defined on a universe U is characterised by its membership function $\mu_{\tilde{A}} : U \to [0, 1]$, which assigns to each element $x \in U$ a degree of membership. In this work, we adopt *Triangular Fuzzy Numbers* (TFNs) $\tilde{A} = (a, b, c)$, whose membership function is defined as:

$$\mu_{\tilde{A}}(x) = \begin{cases} \dfrac{x-a}{b-a}, & a \le x \le b, \\ \dfrac{c-x}{c-b}, & b \le x \le c, \\ 0, & \text{otherwise.} \end{cases}$$

TFNs offer a compact and interpretable way to encode qualitative assessments, particularly when only linguistic terms such as *low*, *medium*, and *high* are available. For CIA

evaluations, we define three overlapping TFNs over $[0, 1]$:

$$\textit{low} = (0, 0, 0.4), \qquad \textit{medium} = (0.2, 0.5, 0.8), \qquad \textit{high} = (0.6, 1, 1).$$

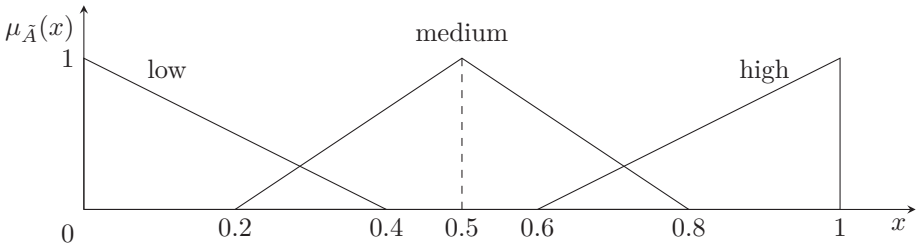


Figure 17.1: Triangular fuzzy membership functions for the linguistic scale (*low*, *medium*, *high*) over $[0, 1]$.

Fuzzy Aggregation and Defuzzification

Within each agent, confidentiality $\tilde{C}$, integrity $\tilde{I}$, and availability $\tilde{A}$ are combined into a consolidated fuzzy risk indicator:

$$\tilde{R} = \lambda_C \, \tilde{C} + \lambda_I \, \tilde{I} + \lambda_A \, \tilde{A}, \qquad \lambda_C + \lambda_I + \lambda_A = 1,$$

where the weights λ express mission priorities (e.g., availability-dominant systems in OT environments). To obtain a scalar index for visualisation or decision support, we apply *centroid defuzzification*:

$$\bar{x}(\tilde{X}) = \frac{a + b + c}{3},$$

which corresponds to the centroid of the TFN and represents its expected operability level. The support $[a, c]$ remains available to visualise the associated uncertainty band.

Fuzzy Propagation Across Dependencies

Propagation between agents follows a fuzzy min–max calculus applied to the TFN vertices. For two fuzzy inputs $A = (a_1, a_2, a_3)$ and $B = (b_1, b_2, b_3)$:

$$\min(A, B) = \big(\min(a_1, b_1),\, \min(a_2, b_2),\, \min(a_3, b_3)\big),$$

$$\max(A, B) = \big(\max(a_1, b_1),\, \max(a_2, b_2),\, \max(a_3, b_3)\big).$$

The min operator models *mandatory* dependencies—where degradation in any input directly limits the output—while the max operator captures *redundant* dependencies—where alternative paths or backup resources maintain operability. This rule set ensures consistent and conservative uncertainty propagation throughout the network without requiring additional tuning parameters.

17.3 Case Study and Results

The fuzzy extension of CISIApro 2.0 is evaluated using the same simulated setup introduced in Chapter 15. This configuration reproduces the interconnection structure among eight representative infrastructures included in the Italian National Cybersecurity Perimeter: two telecommunications operators, two electricity distribution companies, a railway operator, an airline, a bank, and a government department. The topology of this multi-sector network mirrors realistic national-scale dependencies, in which electricity powers ICT and transport systems, telecommunications provide connectivity and data exchange to all sectors, the banking system enables financial transactions, and public administration

fulfils regulatory and coordination roles. The same dependency graph, previously shown in Figure 15.1, is adopted here as the reference structure to ensure full consistency and comparability of results across modelling approaches.

The results are presented through two complementary visualisations: (i) *Triangular Fuzzy Number (TFN) glyphs* for selected agents, showing the fuzzy composition of confidentiality, integrity, and availability (CIA); and (ii) *system synoptics*, where colour intensity encodes the centroid (defuzzified operability), while the width of the TFN indicates the associated uncertainty.

Two representative use cases are analysed. The first models an availability-centric disturbance originating in the energy sector; the second simulates a confidentiality breach in public administration. Together, they demonstrate the expressiveness of the fuzzy approach in capturing heterogeneous impact patterns and uncertainty propagation.

Use Case 1: Vulnerability Disclosure (Availability-Centric)

The first scenario simulates the disclosure of vulnerability CVE-2024-3493, affecting a network device used in electricity distribution. A malformed fragmented packet can cause a critical non-recoverable fault, leading to service interruption and requiring manual recovery. CIA dimensions are parameterised as TFNs, with availability assigned the highest weight to reflect the operational priorities typical of critical OT systems.

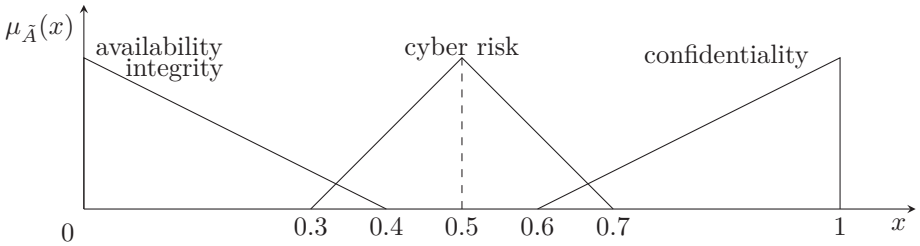


Figure 17.2: Cyber–operational risk TFN at an electrical substation (grid 2), with availability-dominant weights $(0.1, 0.4, 0.5)$ for (C, I, A).

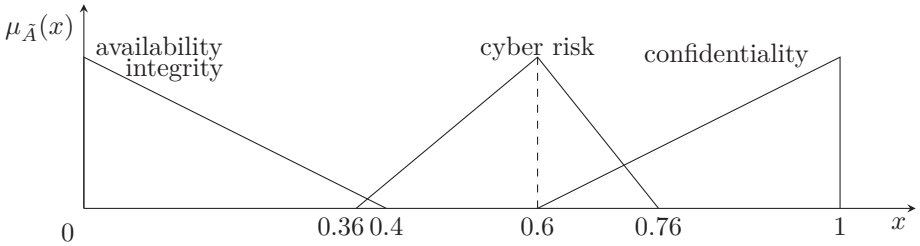


Figure 17.3: Cyber–operational risk TFN at another electric node (grid 2), with $(0.2, 0.2, 0.6)$ CIA weights.

Figures 17.2–17.3 illustrate how availability-centric weighting shifts the consolidated risk TFN towards lower operability values. The centroid of each TFN provides a scalar measure of system performance, while the triangle width represents epistemic uncertainty—arising, for instance, from incomplete device metadata or uncertain exploit conditions.

In the synoptic view of grid 2 (Figure 17.4), degradation originates at substations and propagates through dependent nodes. Because several telecommunications assets rely on these feeders, secondary effects appear in ICT networks where redundancy is limited.

Figure 17.5 confirms that telecom 2 inherits availability losses from grid 2. Where backup power or dual feeds are available, the max (fuzzy union) operation preserves higher operability levels, demonstrating how redundancy mitigates uncertainty.

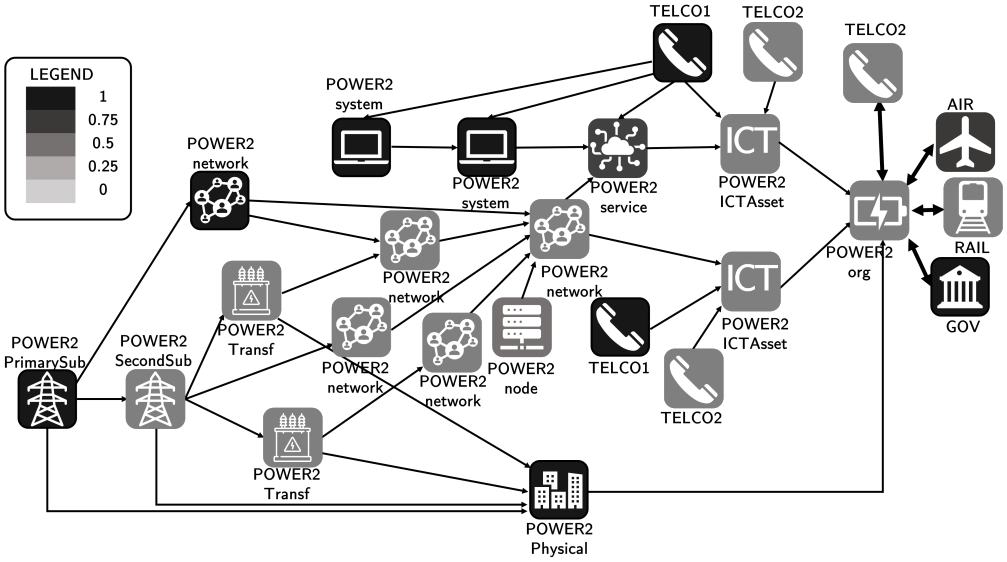


Figure 17.4: Electric distribution grid 2 in use case 1: agent operability (colour by centroid), with annotated local impacts.

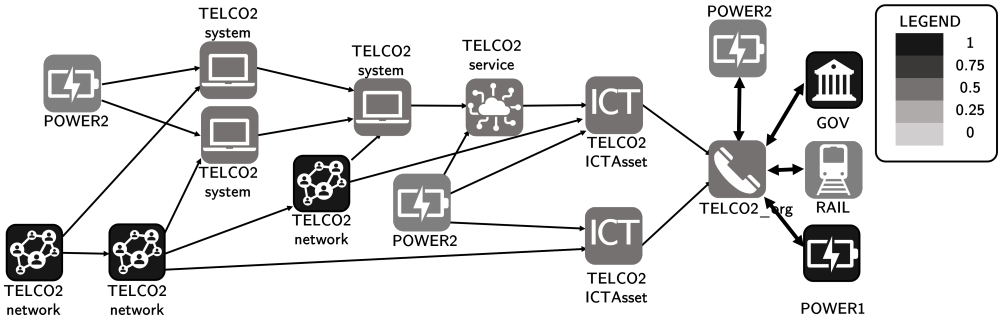


Figure 17.5: Telecommunication network 2 in use case 1: degradation driven by upstream power feeders; nodes without adequate redundancy show reduced operability.

Railway operations (Figure 17.6) decline at nodes requiring simultaneous availability of telecom and power inputs. The min (fuzzy intersection) operator models this series-like dependency, amplifying degradation in signaling and coordination subsystems.

Airline and banking systems (Figures 17.7–17.8) remain largely operational. Their architectures exhibit parallel dependencies, captured by the max fusion, which preserves higher operability and narrower TFN supports in well-instrumented subsystems.

The cross-sector dashboard in Figure 17.9 synthesises these effects: electricity grid 2 acts as the main amplifier; telecom 2 and rail propagate its losses via mandatory dependencies; airline and bank maintain acceptable performance due to redundant connections. Uncertainty remains explicit in the TFN spread, supporting transparent interpretation of model confidence.

Use Case 2: Data Breach (Confidentiality-Centric)

The second scenario focuses on a confidentiality-driven disturbance. A simulated data breach in a government department affects internal information assets and service integrity. CIA dimensions are fuzzified accordingly, with confidentiality receiving the highest weight, while availability remains essential for maintaining baseline functionality.

Figure 17.10 shows a wide confidentiality TFN, indicating significant epistemic uncertainty due to incomplete asset visibility. However, the aggregated cyber–operational

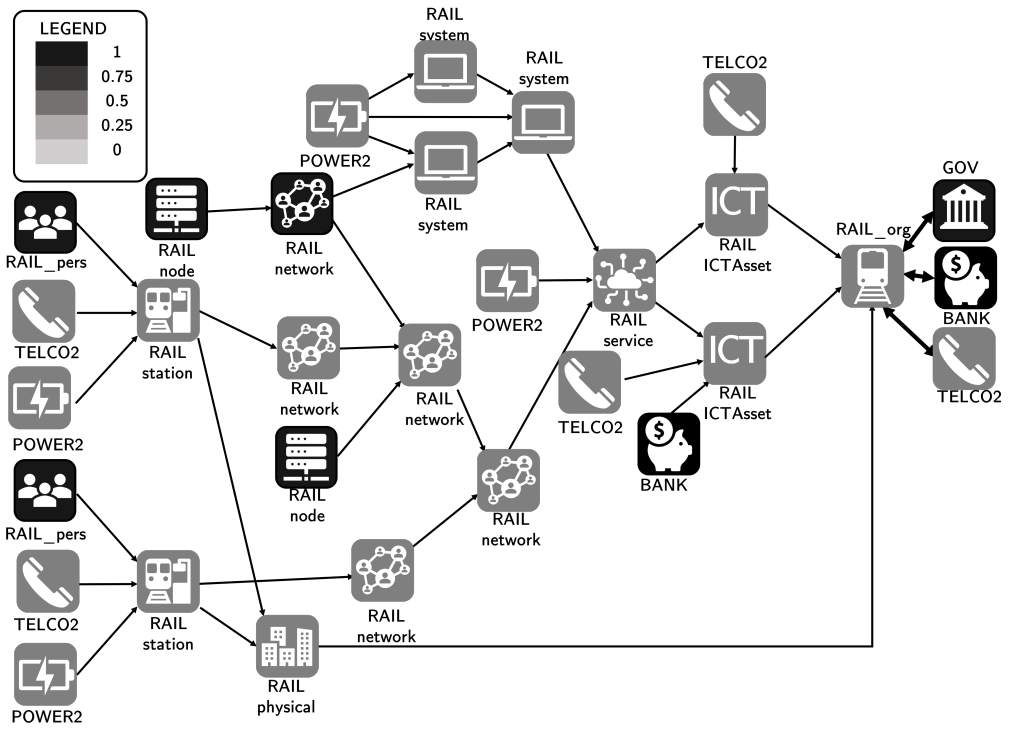


Figure 17.6: Railway network in use case 1: coupled degradation at stations relying on both telecom 2 and grid 2.

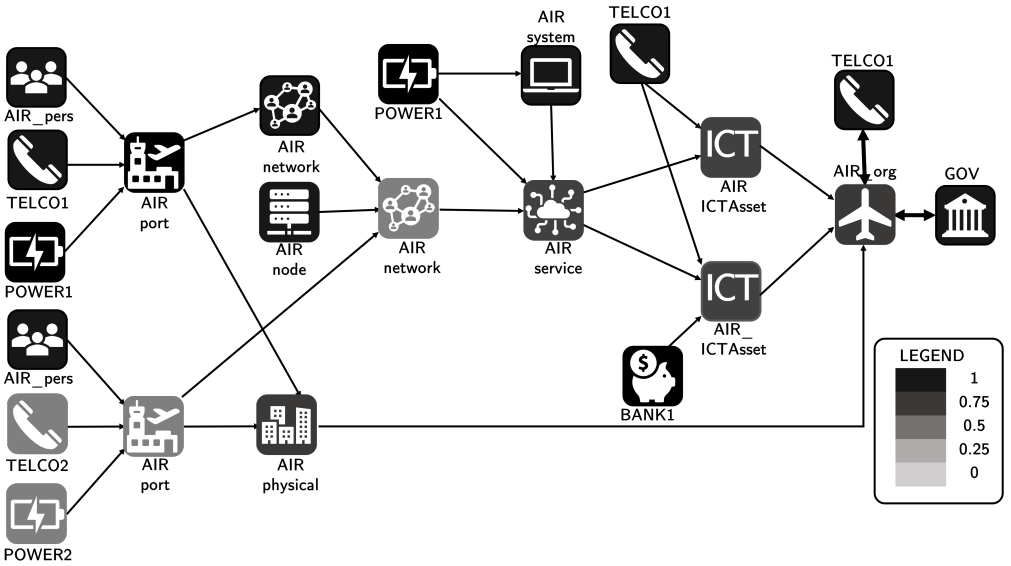


Figure 17.7: Air transport network in use case 1: limited impact due to diversified power and telecom redundancy.

risk remains moderate, as availability is less affected. The synoptic in Figure 17.11 confirms that the event remains localised and produces negligible systemic effects—consistent with the logical rather than physical nature of the dependency links.

Overall, the two case studies demonstrate the expressive power of fuzzy modelling in CISIApro 2.0. The availability-centric scenario highlights how redundancy and dependency type influence uncertainty propagation, while the confidentiality-centric case shows that fuzzy CIA representations allow qualitative assessments to be transparently integrated

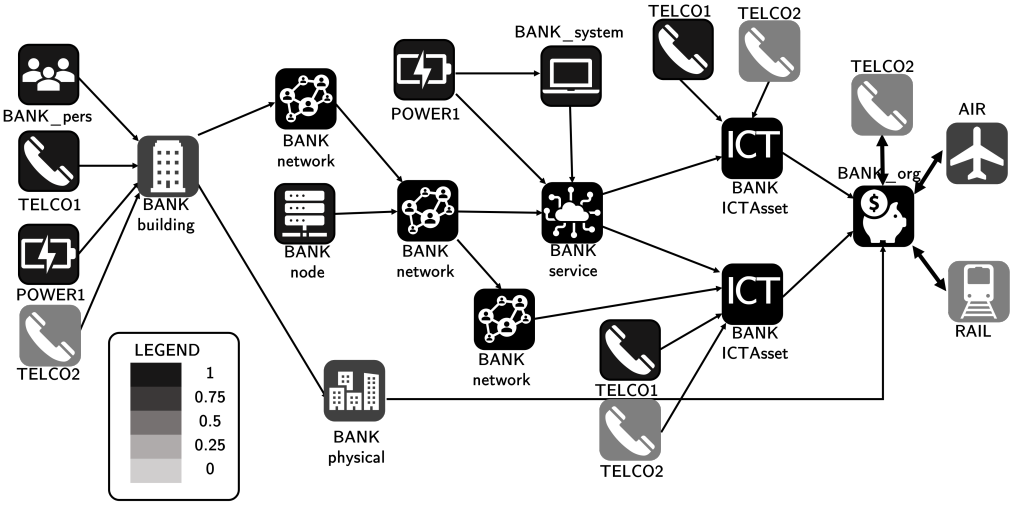


Figure 17.8: Bank network in use case 1: minor degradation thanks to parallel telecom dependencies and resilient data-centre power.

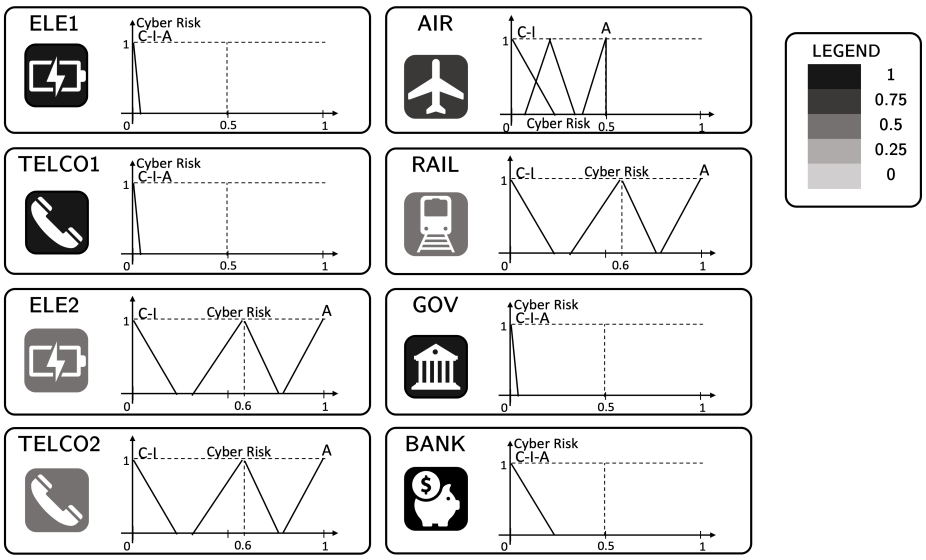


Figure 17.9: Use case 1 dashboard: background colour encodes operability (TFN centroid); triangle width expresses uncertainty. Electric 2, telecom 2, and rail are most affected.

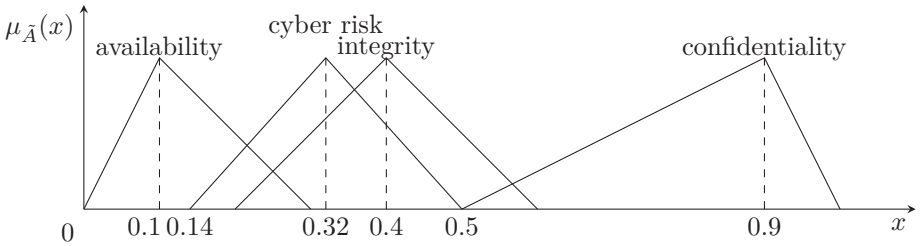


Figure 17.10: Government department node: confidentiality-dominant effect; availability-weighted aggregation maintains moderate overall risk.

into systemic analysis.

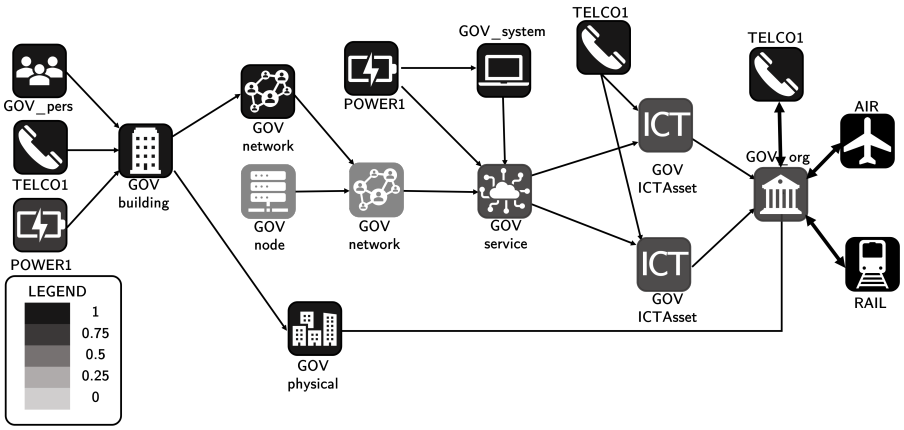


Figure 17.11: Government department in use case 2: localised impact, with minimal cross-sector propagation.

17.4 Discussion

The fuzzy CIA extension introduced in CISIApro 2.0 provides a simple yet effective way to represent uncertainty throughout the entire impact assessment process—from local evaluations of single agents to the analysis of cross-sector propagation. By replacing crisp CIA values with triangular fuzzy numbers, the model explicitly captures the imprecision and subjectivity that often characterise expert-based or incomplete data sources, while preserving full compatibility with the CISIApro agent architecture and dependency logic.

The experiments demonstrate how this addition enhances both interpretability and analytical robustness. In the availability-centric scenario, the electricity distribution operator (grid 2) emerges as the main stressor: its impairment propagates to telecommunications and rail through mandatory (series-like) dependencies, clearly visualised in the fuzzy synoptics. In contrast, the confidentiality-centric breach within public administration remains confined, producing only limited systemic effects. These results show that the fuzzy framework not only preserves the relative behaviour of interdependent infrastructures but also provides a richer representation of uncertainty, distinguishing what is *plausible* from what is *likely*. The resulting dashboards combine an intuitive scalar operability index (the TFN centroid) with a visual uncertainty band (the TFN support), making the analysis more transparent for both technical and decision-making audiences.

Despite these advantages, several limitations remain. First, the definition of fuzzy parameters still relies on expert judgement; without proper calibration, this may introduce optimistic or conservative bias. Second, the current propagation scheme is static and does not yet capture temporal aspects such as failover delays, repair dynamics, or cascading timing. Finally, the use of min–max fusion, while conservative and computationally efficient, assumes separability among dependencies and may over- or under-estimate impacts in tightly coupled systems.

Overall, the fuzzy extension represents a significant step toward more realistic and interpretable impact assessments. It bridges quantitative and qualitative reasoning, allowing CISIApro 2.0 to operate under uncertainty without sacrificing traceability or scalability—an essential requirement for decision support in complex, interdependent critical infrastructures.

References — Observering and Simulation of Critical Infrastructure Interdependencies

[5] Ouyang, M., "Review on modeling and simulation of interdependent critical infrastructure systems," *Reliability engineering & System safety*, vol. 121, pp. 43–60, 2014.

[6] Setola, R., Rosato, V., Kyriakides, E., and Rome, E., *Managing the complexity of critical infrastructures: A modelling and simulation approach*. Springer Nature, 2016.

[59] Foglietta, C. and Panzieri, S., "Resilience in critical infrastructures: The role of modelling and simulation," in *Issues on Risk Analysis for Critical Infrastructure Protection*, IntechOpen, 2020.

[60] Bernardini, E., Foglietta, C., and Panzieri, S., "Modeling telecommunications infrastructures using the cisiapro 2.0 simulator," in *International Conference on Critical Infrastructure Protection*, Springer, 2020, pp. 325–348.

[61] Foglietta, C., Palazzo, C., Santini, R., and Panzieri, S., "Assessing cyber risk using the cisiapro simulator," in *International Conference on Critical Infrastructure Protection*, Springer, 2015, pp. 315–331.

[62] Foglietta, C. et al., "The italian national security perimeter for cyber: Assessing and modeling cyber threats in interdependent critical infrastructures with the cisiapro simulator," in *Seventeenth Annual IFIP WG 11.10 International Conference on Critical Infrastructure Protection*, 2023.

[120] Rinaldi, S., Peerenboom, J., and Kelly, T., "Identifying, understanding, and analyzing critical infrastructure interdependencies," *IEEE Control Systems Magazine*, vol. 21, no. 6, pp. 11–25, 2001.

[121] Pederson, P., Dudenhoeffer, D., Hartley, S., and Permann, M., "Critical infrastructure interdependency modeling: A survey of us and international research," *Idaho National Laboratory*, vol. 25, p. 27, 2006.

[122] Satumtira, G. and Dueñas-Osorio, L., "Synthesis of modeling and simulation methods on critical infrastructure interdependencies research," *Sustainable and Resilient Critical Infrastructure Systems: Simulation, Modeling, and Intelligent Engineering*, pp. 1–51, 2010.

[123] Brown, T., Beyeler, W., and Barton, D., "Assessing infrastructure interdependencies: The challenge of risk analysis for complex adaptive systems," *International Journal of Critical Infrastructures*, vol. 1, no. 1, pp. 108–117, 2004.

[124] Liu, R. and Srivastava, A., "Integrated simulation to analyze the impact of cyber-attacks on the power grid," in *2015 Workshop on Modeling and Simulation of Cyber-Physical Energy Systems (MSCPES)*, IEEE, 2015, pp. 1–6.

[125] Drias, Z., Serhrouchni, A., and Vogel, O., "Analysis of cyber security for industrial control systems," in *2015 international conference on cyber security of smart cities, industrial control system and communications (ssic)*, IEEE, 2015, pp. 1–8.

[126] Bhamare, D., Zolanvari, M., Erbad, A., Jain, R., Khan, K., and Meskin, N., "Cybersecurity for industrial control systems: A survey," *computers & security*, vol. 89, p. 101 677, 2020.

[127] Masucci, D., Palazzo, C., Foglietta, C., and Panzieri, S., "Enhancing decision support with interdependency modeling," in *International Conference on Critical Infrastructure Protection*, Springer, 2016, pp. 169–183.

[128] Oliva, G., Panzieri, S., and Setola, R., "Modeling and simulation of critical infrastructures," *WIT Transactions on State-of-the-art in Science and Engineering*, vol. 54, 2012.

[132] Chen, J., "Cybergrid: Teaching and experiencing international teamwork through virtual environment," 2017.

[133] Ficco, M., Choraś, M., and Kozik, R., "Simulation platform for cyber-security and vulnerability analysis of critical infrastructures," *Journal of Computational Science*, vol. 22, pp. 179–186, 2017. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1877750317303605.

[134] Balakrishnan, S. and Zhang, Z., "Modeling interdependent effects of infrastructure failures using imprecise dependency information," *Sustainable and Resilient Infrastructure*, vol. 7, no. 2, pp. 153–169, 2022.

[138] Sun, W., Bocchini, P., and Davison, B. D., "Overview of interdependency models of critical infrastructure for resilience assessment," *Natural Hazards Review*, vol. 23, no. 1, p. 04 021 058, 2022.

[139] Reilly, A. C., Baroud, H., Flage, R., and Gerst, M. D., "Sources of uncertainty in interdependent infrastructure and their implications," *Reliability Engineering & System Safety*, vol. 213, p. 107 756, 2021.

[321] Xiao, N., Sharman, R., Rao, H. R., and Upadhyaya, S., "Infastructure interdependencies modeling and analysis-a review and synthesis," in *AMCIS 2008 Proceedings*, 2008.

[322] Commission, E., *NIS Directive*, https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=uriserv:OJ.L_.2016.194.01.0001.01.ENG&toc=OJ:L:2016:194:TOC, 2016.

[323] Italian Official Journal, L.D. n.65/2018, *Italian transposition law EU NIS Directive*, https://www.gazzettaufficiale.it/eli/id/2018/06/09/18G00092/sg, 2018.

[324] *ECCC*, https://cybersecurity-centre.europa.eu/index_en, Accessed: 2023-01-13.

[325] *Italian Official Journal, L.D. n.133/2019*, https://www.gazzettaufficiale.it/eli/id/2019/11/20/19G00140/sg, 2019.

[326] *Italian Official Journal, DPCM n.131/2020*, https://www.gazzettaufficiale.it/eli/id/2020/10/21/20G00150/sg, 2020.

[327] *Italian Official Journal, DPCM n.81/2021*, https://www.gazzettaufficiale.it/eli/id/2021/06/11/21G00089/sg, 2021.

[328] *Italian Official Journal, DPCM n.15/2021*, https://www.gazzettaufficiale.it/eli/id/2021/08/19/21A05087/sg, 2021.

[329] *Italian Official Journal, DPCM n.92/2022*, https://www.gazzettaufficiale.it/eli/id/2022/07/15/22G00099/sg, 2022.

[330] *Italian Official Journal, L.D. n.82/2021*, https://www.gazzettaufficiale.it/eli/id/2021/06/14/21G00098/sg, 2021.

[331] *Italian National Cybersecurity Agency (ACN), Cybersecurity strategy 2022 - 2026*, https://www.acn.gov.it/ACN_EN_Strategia.pdf, 2022.

[332] Juárez García, H., Marti, J., and Srivastava, K., "Infrastructure interdependency simulation using i2sim in a study case," Sep. 2012.

[333] Franchina, L. and Socal, A., "Innovative predictive model for smart city security risk assessment," in *2020 43rd International Convention on Information, Communication and Electronic Technology (MIPRO)*, IEEE, 2020, pp. 1831–1836.

[334] Assante, M. J. and Lee, R. M., "The industrial control system cyber kill chain," *SANS Institute InfoSec Reading Room*, vol. 1, 2015.

[335] Sobb, T., Turnbull, B., and Moustafa, N., "A holistic review of cyber–physical–social systems: New directions and opportunities," *Sensors*, vol. 23, no. 17, p. 7391, 2023.

[336] Der Kiureghian, A. and Ditlevsen, O., "Aleatory or epistemic? does it matter?" *Structural safety*, vol. 31, no. 2, pp. 105–112, 2009.

[337] Hüllermeier, E. and Waegeman, W., "Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods," *Machine Learning*, vol. 110, pp. 457–506, 2021.

[338] Dubois, D., Nguyen, H. T., and Prade, H., "Possibility theory, probability and fuzzy sets misunderstandings, bridges and gaps: Misunderstandings, bridges and gaps," *Fundamentals of fuzzy sets*, pp. 343–438, 2000.

[339] Dubois, D., "Possibility theory and statistical reasoning," *Computational statistics & data analysis*, vol. 51, no. 1, pp. 47–69, 2006.

[340] Baudrit, C., Dubois, D., and Guyonnet, D., "Joint propagation and exploitation of probabilistic and possibilistic information in risk assessment," *IEEE transactions on fuzzy systems*, vol. 14, no. 5, pp. 593–608, 2006.

[341] European Union Agency for Cybersecurity, *ENISA threat landscape 2022 : overview of current and emerging cyber-threats*. European Network and Information Security Agency, 2022.

[342] European Union Agency for Cybersecurity, *ENISA threat landscape for ransomware attacks*. European Network and Information Security Agency, 2022.

[343] Rinaldi, G., Menon, P. P., Edwards, C., Ferrara, A., and Shtessel, Y., "Adaptive dual-layer super-twisting sliding mode observers to reconstruct and mitigate disturbances and communication attacks in power networks," *Automatica*, vol. 129, p. 109 656, 2021.

[344] Klir, G. and Yuan, B., *Fuzzy sets and fuzzy logic*. Prentice hall New Jersey, 1995, vol. 4.

[345] Wang, Z. and Klir, G. J., *Fuzzy Measure Theory*. USA: Kluwer Academic Publishers, 1993.

[346] Dubois, D. and Prade, H., "Fuzzy numbers: An overview," *Readings in Fuzzy Sets for Intelligent Systems*, pp. 112–148, 1993.

[347] Bai, X. and Asgarpoor, S., "Fuzzy-based approaches to substation reliability evaluation," *Electric Power Systems Research*, vol. 69, no. 2-3, pp. 197–204, 2004.

[348] Wang, Y.-M., "Centroid defuzzification and the maximizing set and minimizing set ranking based on alpha level sets," *Computers & Industrial Engineering*, vol. 57, no. 1, pp. 228–236, 2009.

Conclusions

This dissertation has investigated the problem of resilience in cyber–physical and interdependent systems through a perspective that is both analytical and systemic. Across the different chapters, the research has sought to demonstrate that the resilience of modern infrastructures cannot be ensured solely by technical countermeasures or post-incident mitigation strategies. Instead, resilience must emerge as a property of the system's structure, of the information that circulates within it, and of the ways in which its agents—whether human or artificial—coordinate their actions under uncertainty. The guiding principle throughout this work is that resilience is not a static attribute, but a dynamic behaviour: it depends on how systems perceive, interpret, and respond to disturbances, including those intentionally caused by intelligent adversaries.

The thesis followed a logical progression that moved from the microscopic level of estimation and detection, to the mesoscopic level of distributed coordination, and finally to the macroscopic level of national-scale systemic risk. Each layer addressed a specific manifestation of the resilience problem, yet all were united by a common methodological thread: the use of models—mathematical, algorithmic, and conceptual—as instruments for understanding and improving the behaviour of complex systems.

At its foundation, the research proposed observer-based architectures capable of detecting and isolating malicious perturbations in networked control systems. The Multirate Interlaced Kalman Filter and the Distributed Sliding Mode Observer were conceived to reconcile two seemingly conflicting objectives: sensitivity to anomalies and robustness to model uncertainty. This duality reflects a recurrent theme in the dissertation—the search for balance between responsiveness and stability, precision and generality, trust and skepticism. The results demonstrated that distributed observers, when properly coordinated, can reconstruct hidden attack signals without requiring centralised supervision, paving the way for scalable and privacy-preserving monitoring architectures.

Building upon this foundation, the research introduced the temporal dimension of information as a determinant of system performance and cyber resilience. Through the concept of the *Age of Incorrect Information*, it became possible to characterise not only how frequently data are refreshed, but also how their correctness degrades under adversarial conditions. This formulation unifies communication latency, data integrity, and attack dynamics within a single metric, revealing that the erosion of information quality—if left unaddressed—can itself become a vector of systemic vulnerability. By embedding this logic within a game-theoretic framework, the thesis offered a formal means to study the strategic interplay between attackers and defenders, showing that resilience can be understood as an emergent equilibrium between competing adaptive behaviours.

The analysis then extended to the cooperative dimension of large-scale systems. Distributed consensus and coordination mechanisms form the backbone of modern cyber–physical infrastructures—from industrial automation to smart grids and transport networks—but they are also key points of fragility under cyber attack. The thesis contributed to this field by proposing secure consensus protocols that integrate trust management and privacy preservation. Within these frameworks, agents are able to evaluate the reliability of incoming data, isolate deceptive peers, and still converge toward a collective decision. This demonstrates that resilience and confidentiality—often viewed as conflicting objectives—can in fact be made complementary through careful algorithmic design.

At the macroscopic level, the research explored how interdependencies among critical infrastructures can amplify or dampen the effects of cyber incidents. Using the CISIApro 2.0 platform as a foundation, the institutional structure of the Italian National Cybersecurity Perimeter was translated into a living quantitative model. In this way, a policy instrument—originally conceived for regulation and oversight—was transformed into an analytical tool for systemic risk assessment. The simulations revealed that sectors such as telecommunications and energy act as propagation hubs, while finance and public administration function as coupling points between digital and physical domains. This result illustrates how systemic cyber risk can only be understood through

an integration of technical modelling and institutional insight.

A methodological turning point emerged when confronting the limits of available information. Many of the data required for systemic modelling are incomplete, qualitative, or classified; parameters such as dependency strengths, recovery times, or redundancy levels are often known only approximately. To address this gap, the thesis introduced two major extensions to the CISIApro framework—one based on possibility theory and one on fuzzy logic—that allow the explicit representation of epistemic uncertainty. Rather than forcing spurious precision through probabilistic assumptions, these extensions make it possible to work with partial knowledge, expressing outcomes as ranges of plausibility and necessity. This represents not only a technical improvement but also a conceptual shift: uncertainty is reframed as a legitimate form of knowledge—something that must be made explicit, interpretable, and traceable if decision-making is to be both rigorous and trustworthy.

From a critical and forward-looking standpoint, the research presented in this dissertation naturally opens several avenues for further development. These are not limitations in the negative sense, but rather the natural continuation of an ambitious methodological trajectory that sought to link analytical rigour with systemic realism.

A first line of advancement concerns the *model-based detection* framework. The proposed observer architectures rely on an accurate characterisation of system dynamics and interconnections. In real industrial environments, however, such models are rarely complete and evolve continuously with updates and reconfigurations. This points toward the development of *hybrid learning–model approaches*, where data-driven inference complements physical modelling while preserving interpretability. Machine learning could thus be used to adapt to non-stationary conditions, while the model-based core provides the structure and guarantees necessary for trust and certification.

A second direction arises from the *information-centric and game-theoretic* formulation. While the AoII metric captures the essence of information decay and adversarial timing, it still assumes rational and fully informed players. Future research could extend this framework toward *behavioural and adaptive game models*, capturing bounded rationality, learning dynamics, and the cognitive aspects of human decision-making under stress. Such an extension would bridge formal resilience theory with the socio-technical reality of cyber conflict.

A third research path relates to the *systemic modelling of interdependent infrastructures*. Although the CISIApro 2.0 simulations capture structural dependencies effectively, the current framework remains static. Extending it to a *dynamic, time-resolved interdependency model*—incorporating repair functions, temporal feedbacks, and coordination mechanisms—would significantly enhance its realism and policy relevance. This would enable not only the analysis of cascading failures but also of recovery and adaptive cooperation processes over time.

A fourth frontier concerns the *representation and calibration of uncertainty*. The possibilistic and fuzzy approaches introduced here mark a significant step toward the explicit inclusion of epistemic uncertainty, but they still rely heavily on expert elicitation. Future work should focus on *data-informed calibration*, integrating expert judgment with empirical datasets, monitoring systems, and Bayesian updating. This would provide a more statistically grounded and transparent quantification of uncertainty, strengthening the interpretability and reproducibility of systemic analyses.

Finally, an overarching challenge concerns the interface between humans and uncertainty-aware systems. As models become more sophisticated, the bottleneck shifts from computation to comprehension. How can analysts interpret fuzzy or possibilistic outcomes? How can decision-makers translate uncertainty intervals into policy or operational choices? These questions call for interdisciplinary work at the intersection of cognitive science, human–computer interaction, and resilience engineering. Developing *explainable and participatory modelling environments* will be essential to ensure that complex systemic analyses remain transparent, communicable, and ethically grounded.

Taken together, these directions outline a coherent roadmap for future research. They do not contradict the findings of this dissertation but extend them, reinforcing its central message: resilience is not a fixed property, but an evolving capability. Each methodological layer—observer design, information dynamics, systemic modelling, and uncertainty representation—offers both a concrete contribution and a challenge for further refinement. The next step lies in unifying these elements through adaptive, data-driven, and human-centred approaches that preserve interpretability while embracing the complexity of real-world systems.

Ultimately, this dissertation argues that resilience in cyber–physical systems emerges from the continuous interaction between structure, knowledge, and governance. Technological safeguards alone are insufficient if they are not embedded within architectures capable of learning, adapting, and communicating uncertainty effectively. The vision proposed here is therefore one of *living resilience*: a property that grows through feedback, evolves through experience, and depends as much on trust and collaboration as on algorithms and models. If the work presented in these pages contributes, even partially, to advancing this vision—by offering both formal tools and conceptual bridges—then it will have achieved its fundamental purpose.

Appendices

Appendix A

Proof of Theorems in Chap. 3

A.1 Proof of Theorem 3.2.1

We begin with the weak-SED case and for $i \neq j$ we write

$$\begin{aligned} |X_{i,j}| &= \left| x\beta \sum_{k\geq 1} \lambda^k [U^k]_{i,j} \right| = \left| x\beta \sum_{k\geq d(i,j)} \lambda^k [U^k]_{i,j} \right| \\ &\leq x\beta \sum_{k\geq d(i,j)} \lambda^k |[U^k]_{i,j}| \\ &\leq x\beta \sum_{k\geq d(i,j)} \lambda^k = \tfrac{x\beta}{1-\lambda} \lambda^{d(i,j)} \ , \end{aligned}$$

where in the first row we exploited the fact that i and j are at distance $d(i,j)$, hence they are defined by the contributions U^k starting from $k = d(i,j)$, and in the third row we used $|[U^k]_{i,j}| \leq 1$ since the norm of U^k is upper bounded by 1. This proves the result for the weak-SED case.

For strong-SED properties, the steps are similar, with the addition of a factorial, so that

$$|X_{i,j}| = \left| x\beta \sum_{k\geq d(i,j)} \tfrac{1}{k!} \lambda^k [U^k]_{i,j} \right| \leq x\beta \sum_{k\geq d(i,j)} \tfrac{1}{k!} \lambda^k \ ,$$

where we also note that

$$\bar{\gamma}_d(\lambda) = \sum_{k\geq d} \tfrac{1}{k!} \lambda^k = \lambda^d \sum_{n\geq 0} \tfrac{1}{(n+d)!} \lambda^n < \lambda^d \, e^\lambda \ .$$

A.2 Proof of Theorem 3.2.2

The result for F_r is a straightforward consequence of the fact that $F_r = e^{\Delta t L_r}$ can be written in terms of the exponential series, that is

$$F_r = e^{\Delta t L_r} = \sum_{k\geq 0} \tfrac{1}{k!} \left(\Delta t L_r\right)^k = \sum_{k\geq 0} \tfrac{1}{k!} \left(\Delta t \|L_r\|\right)^k U^k \ ,$$

where $U^k = (L_r/\|L_r\|)^k$ is active, by definition, up to distance k. The transpose is equivalent, since $L_r^\top$ is still active only at distance 1.

For $Z_r = I + G_r B_r G_r^\top$, instead, we need to identify a strict expression which can be obtained through close inspection of products. We first recall that by definition B_r is constrained to distance one, that is $B_r = \|B_r\| \, U$, and it is also positive semidefinite, so that we can also write

$$B_r = \tfrac{1}{2} \|B_r\| \left(I + U\right) ,$$

in the U notation. We take this second expression. Then, for $G_r = (e^{\Delta t L_r} - I) \, L_r^{-1}$ and $G_r^\top$, it is sufficient to recall the power series

$$\tfrac{e^x - 1}{x} = \sum_{k\geq 1} \tfrac{1}{k!} x^{k-1} = \sum_{m\geq 0} \tfrac{1}{(m+1)!} x^m \ .$$

This provides

$$G_r = \Delta t \sum_{k\geq 0} \tfrac{1}{(k+1)!}\lambda_f^k\, U^k$$
$$G_r U = \Delta t \sum_{k\geq 1} \tfrac{1}{k!}\lambda_f^{k-1}\, U^k$$
$$G_r(I+U) = \Delta t \left(I + \sum_{k\geq 1} \tfrac{1}{k!}\Big(\tfrac{1}{k+1} + \tfrac{1}{\lambda_f}\Big)\lambda_f^k\, U^k \right) ,$$

and by further adding the multiplication with $G_r^\top$ we obtain

$$\begin{aligned}
\frac{G_r(I+U)G_r^\top}{\Delta t^2} &= \sum_{k\geq 0} \tfrac{1}{(k+1)!}\lambda_f^k\, U^k \\
&\quad + \sum_{k\geq 1} \tfrac{1}{k!}\lambda_f^k\, U^k \sum_{m=1}^{k} \tfrac{k!}{m!(k-m+1)!}\Big(\tfrac{1}{m+1} + \tfrac{1}{\lambda_f}\Big) \\
&= I + \sum_{k\geq 1} \tfrac{1}{k!}\lambda_f^k\, U^k \left(\tfrac{1}{k+1} + \tfrac{2^{k+2}-k-4}{(k+1)(k+2)} + \tfrac{2^{k+1}-2}{\lambda_f(k+1)}\right) \\
&= I + \sum_{k\geq 1} \tfrac{1}{k!}\lambda_f^k\, U^k \Big(\underbrace{\tfrac{2^{k+2}-2}{(k+1)(k+2)}}_{\leq 2^{k-1}} + \tfrac{1}{\lambda_f} \underbrace{\tfrac{2^{k+1}-2}{k+1}}_{\leq 2^{k-1}} \Big) ,
\end{aligned}$$

where we used m for $G_r(I+U)$ and $\ell = k-m$ for G_r. This proves

$$Z_r = (1+q) + q\tfrac{1}{2}(1+\tfrac{1}{\lambda_f}) \sum_{k\geq 1} \tfrac{1}{k!}\alpha_k\, (2\lambda_f)^k\, U^k$$

where

$$\alpha_k = (1+\tfrac{1}{\lambda_f})^{-1}\Big(\tfrac{4(2-2^{-k})}{(k+1)(k+2)} + \tfrac{1}{\lambda_f}\tfrac{4(1-2^{-k})}{(k+1)}\Big) \leq 1$$

is also decreasing in k. This decreasing behaviour allows to redistribute weights in such a to obtain a pure strong-SED expression, since a weight at U^k can be transferred to any U^n with $n > k$, which implies a larger distance. What needs to be taken into account is simply that the coefficients sum does not modify (i.e., the sum when replacing U^k with 1). The result is then obtained from

$$G_r\big|_{U^k=1} = \Delta t\, \tfrac{e^{\lambda_f}-1}{\lambda_f} , \quad B_r\big|_{U^k=1} = 2\,\tfrac{1}{2}\|B_r\| .$$

A.3 Proof of Theorem 3.2.3

Let us investigate the product of $X_1 = \mathcal{S}(\lambda_1)$ and $X_2 = \mathcal{S}(\lambda_2)$, for which we can write

$$\begin{aligned}
X_1 X_2 &= \sum_{k,\ell\geq 0} \tfrac{\lambda_1^k}{k!}\, U^k\, \tfrac{\lambda_2^\ell}{\ell!}\, U^\ell \\
&= \sum_{i\geq 0} \tfrac{1}{i!}\, U^i \sum_{k=0}^{i} \tfrac{i!}{k!(i-k)!}\, \lambda_1^k\, \lambda_2^{i-k} \\
&= \sum_{i\geq 0} \tfrac{1}{i!}\, U^i\, (\lambda_1+\lambda_2)^i ,
\end{aligned}$$

where we reordered k and ℓ in such a way that $i = k+\ell$. Hence the result.

A.4 Proof of Theorem 3.2.4

Let us investigate the product between $X_s = \mathcal{S}(\lambda_s)$ and $X_w = \mathcal{W}_0(\lambda_w)$, for which we have

$$\begin{aligned} X_s X_w &= \sum_{n\geq 0}\sum_{\ell\geq 1} \tfrac{1}{n!}\lambda_s^n U^n \, \lambda_w^\ell U^\ell \\ &= \sum_{k\geq 1} U^k \sum_{n=0}^{k-1} \tfrac{1}{n!}\, \lambda_s^n \, \lambda_w^{k-n} \, , \end{aligned}$$

where we reordered n and ℓ in such a way that $k = n + \ell$. This explains the result.

A.5 Proof of Theorem 3.2.5

We give direct proof by induction. Consider

$$Y = I + \beta \sum_{k\geq 1} \lambda^k U^k \, ,$$

for some $\|U^k\| \leq 1$, and assume that

$$Y^{-1} = I + \beta \sum_{\ell\geq 1} (1+\beta)^{\ell-1} \lambda^\ell V^\ell \, ,$$

where we want to show that the equality holds for $\|V^k\| \leq 1$. This can be verified by investigating

$$\begin{aligned} I &= Y\, Y^{-1} \\ &= \Big(I + \beta \sum_{k\geq 1} \lambda^k U^k\Big) \Big(I + \beta \sum_{\ell\geq 1} (1+\beta)^{\ell-1} \lambda^\ell V^\ell\Big) \\ &= I + \sum_{i\geq 1} \Bigg(\beta U^i + \beta (1+\beta)^{i-1} V^i \\ &\qquad\qquad + \beta^2 \sum_{\ell=1}^{i-1} (1+\beta)^{\ell-1} U^{i-\ell} V^\ell \Bigg) \, \lambda^i \, . \end{aligned}$$

For $k = 1$ the above ensures

$$V^1 = -U^1 \, ,$$

and therefore $\|V^1\| = \|U^1\| \leq 1$. For $i \geq 2$, we proceed by induction and assume $\|V^k\| \leq 1$ for $k < i$. From the above we can then write

$$V^i = -\frac{U^i + \beta \sum_{\ell=1}^{i-1} (1+\beta)^{\ell-1} U^{i-\ell} V^\ell}{(1+\beta)^{i-1}} \, ,$$

whose norm satisfies (by induction)

$$\|V^i\| \leq \tfrac{1}{(1+\beta)^{i-1}} \left(1 + \beta \sum_{\ell=1}^{i-1} (1+\beta)^{\ell-1}\right) = 1 \, .$$

thus proving the result.

A.6 Proof of Theorem 3.2.6

Under the assumption that X is symmetric, then all matrices U^k need to be symmetric. If not, their anti-symmetric parts can be dropped without modifying their sum. By

exploiting the outcomes of the proof of Theorem 3.2.5, for the first orders of V^i we have

$$\begin{aligned}
V^1 &= -U^1 \\
V^2 &= \frac{\beta(U^1)^2 - U^2}{1+\beta} \\
V^3 &= \frac{-\beta^2(U^1)^3 + \beta(U^2U^1 + U^1U^2) - U^3}{(1+\beta)^2} \\
V^4 &= \frac{\beta^3(U^1)^4 - \beta^2(U^2(U^1)^2 + U^1U^2U^1 + (U^1)^2U^2)}{(1+\beta)^3} \\
&\quad + \frac{\beta(U^3U^1 + (U^2)^2 + U^1U^3) - U^4}{(1+\beta)^3} \,,
\end{aligned}$$

where each V^i is built as a sum of components where β^n, $n = 0, \ldots, i-1$, multiplies factors containing all possible combinations of $n+1$ elements U^ℓ, with ℓs summing up to the reference distance i, and with positive signs for odd n, and negative for even n. The (absolute) sum of such factors was previously proved to be $(1+\beta)^{i-1}$, i.e., the denominator. Now, although for odd i nothing new can be said, the even i case has some peculiarities. For example, V^2 carries a positive semidefinite contribution, namely, $(U^1)^2$, since U^1 is symmetric. Its overall strength is

$$\alpha = \tfrac{\beta}{1+\beta} \,.$$

But the same can be said for V^4, when we write it in the form

$$\begin{aligned}
V^4 &= \frac{\beta(\beta(U^1)^2 - U^2)^2}{(1+\beta)^3} \\
&\quad + \frac{-\beta^2(U^1U^2U^1) + \beta(U^3U^1 + U^1U^3) - U^4}{(1+\beta)^3}
\end{aligned}$$

from which we observe that the strength of the semi-definite component is again $\alpha = \beta(1+\beta)^2/(1+\beta)^3$. The property holds for every even i. In fact in V^i we can identify a subpart given by a square expression $(\bar{V}^i)^2$ where $\bar{V}^i$ collects all possible combinations of elements U^ℓ summing up to distance $i/2$. The strength of $\bar{V}^i$, by the β formalism, need to be $(1+\beta)^{\frac{i}{2}-1}$, hence the strength of the positive semidefinite part is, once again, $\alpha = \beta((1+\beta)^{\frac{i}{2}-1})^2/(1+\beta)^{i-1}$.

The above ensures that, for even i it is

$$V^i = \alpha U^i_+ + (1-\alpha)U^i$$

for some positive definite U^i_+, which reduces the eigenvalue range of V^i to $[-(1-\alpha), 1]$. This implies that we can write

$$V^i = \tfrac{\alpha}{2} I + (1 - \tfrac{\alpha}{2})U^i$$

for some $|U^i| \leq 1$. This ensures writing

$$\begin{aligned}
X^{-1} &= I + \tfrac{\beta}{1+\beta} \sum_{n\geq 1} \tfrac{\alpha}{2} I \, (1+\beta)^{2n} \lambda^{2n} \\
&\quad + \tfrac{\beta}{1+\beta} \sum_{n\geq 1} (1+\beta)^{2n-1} \lambda^{2n-1} \, V^{2n-1} \\
&\quad + \tfrac{\beta}{1+\beta} \sum_{n\geq 1} (1 - \tfrac{\alpha}{2}) \, (1+\beta)^{2n} \lambda^{2n} \, V^{2n} \,,
\end{aligned}$$

but we can further shift a small part of the weight insisting on V^{2n-1} onto V^{2n}, since the latter is associated with a bigger distance. Hence, we obtain

$$\begin{aligned}
X^{-1} &= I + \tfrac{\alpha}{2} \, \tfrac{\beta(1+\beta)\lambda^2}{1-(1+\beta)^2\lambda^2} \, I \\
&\quad + \tfrac{\beta}{1+\beta} \sum_{\ell \geq 1} \left(1 - \tfrac{\alpha}{2} \, \tfrac{(1+\beta)\lambda}{1+(1+\beta)\lambda}\right) (1+\beta)^\ell \lambda^\ell \, V^\ell \,,
\end{aligned}$$

which proves the result.

A.7 Proof of Theorem 3.2.7

From the inversion Theorem 3.2.6 we have

$$I - X^{-1} = I - \tfrac{1}{x}\left(\kappa_i(\beta,\lambda)I + \kappa_w(\beta,\lambda)\,\mathcal{W}_0((1+\beta)\lambda)\right) ,$$

which is then multiplied by F_r and $F_r^\top$ (accounting for a global factor $\mathcal{S}(2\lambda_f)$) and then summed to Z_r. Therefore, the final result is a straightforward consequence of Theorem 3.2.4 (for the product) and Corollary 3.2.2 (for the sum).

A.8 Theorem 3.2.8

The proof is an application of the rationale explained in the text. By considering the definition of ξ_k in (3.28), we need to ensure both that the overall rearrangement is below the reference level, that is

$$\sum_{k\geq 1} g_k\lambda^k = \xi_1\lambda \leq x^+\beta^+\sum_{k\geq 1}(\lambda^+)^k = \tfrac{x^+\beta^+\lambda^+}{1-\lambda^+} ,$$

as well as that the rearrangement for $k \geq d$, the only one capable of shifting coefficients to the left, is below the reference level, that is

$$\sum_{k\geq d} g_k\lambda^k = \xi_d\lambda^d \leq x^+\beta^+\sum_{k\geq d}(\lambda^+)^k = \tfrac{x^+\beta^+(\lambda^+)^d}{1-\lambda^+} ,$$

which provides the result. Note that these two are sufficient conditions.

We then investigate the closed form expression for ξ_k. By definition we have

$$\xi_k = \frac{c_1}{\lambda^k}\underbrace{\sum_{n=k}^{\infty}\tfrac{1}{n!}(2\lambda_f)^n}_{\alpha_1} + \frac{c_2}{\lambda^k}\underbrace{\sum_{n=k}^{\infty}\gamma_n\Big(\tfrac{2\lambda_f}{(1+\beta)\lambda}\Big)\,(\lambda(1+\beta))^n}_{\alpha_2}$$

with

$$\begin{aligned} c_1 &= \beta_q + 1 - \tfrac{1}{x}\,\kappa_i(\beta,\lambda) \\ c_2 &= \tfrac{1}{x}\,\kappa_w(\beta,\lambda) . \end{aligned}$$

Now, we have $\alpha_1 = \bar{\gamma}_k(2\lambda_f)$, and

$$\begin{aligned} \alpha_2 &= \sum_{n=k}^{\infty}\sum_{\ell=0}^{n-1}\tfrac{1}{\ell!}(2\lambda_f)^\ell\,(\lambda(1+\beta))^{n-\ell} \\ &= \sum_{\ell=0}^{\infty}\tfrac{1}{\ell!}(2\lambda_f)^\ell \sum_{n=\max(\ell+1,k)}^{\infty}(\lambda(1+\beta))^{n-\ell} \\ &= \tfrac{1}{1-\lambda(1+\beta)}\underbrace{\sum_{\ell=0}^{\infty}\tfrac{1}{\ell!}(2\lambda_f)^\ell\,(\lambda(1+\beta))^{\max(1,k-\ell)}}_{\alpha_3} \end{aligned}$$

where

$$\begin{aligned} \alpha_3 &= (\lambda(1+\beta))^k\sum_{\ell=0}^{k-1}\tfrac{1}{\ell!}\big(\tfrac{2\lambda_f}{\lambda(1+\beta)}\big)^\ell + \lambda(1+\beta)\sum_{\ell=k}^{\infty}\tfrac{1}{\ell!}(2\lambda_f)^\ell \\ &= (\lambda(1+\beta))^k\,\gamma_k\Big(\tfrac{2\lambda_f}{(1+\beta)\lambda}\Big) + \lambda(1+\beta)\bar{\gamma}_k(2\lambda_f) , \end{aligned}$$

which proves the final expression.

A.9 Proof of Theorem 3.3.1

For the covariance matrix $\bar{P}$ it suffices to notice from (3.10) that

$$\begin{aligned}\bar{P} &= A^{-1}(\bar{P}_r - I)A^{-\top} \\ &= (x - I)(A^\top A)^{-1} + x\beta \sum_{k\geq 1} \lambda^k A^{-1} U^k A^{-\top} ,\end{aligned}$$

where $(A^\top A)^{-1}$ is a positive semi-definite matrix, hence it is proportional to $\frac{1}{2}(I + U)$, and where $A^{-1}U^k A^{-\top}$ accounts for distance $k + 1$, because of the block-structure of H and A.

For $\bar{K}H$, instead, we look for an expression that avoids inversion, since inversion may provide unreliable results when $\lambda(1 + \beta)$ is close to 1. We therefore start from (3.4), that at steady state implies $\bar{P} = Q + F\bar{K}HRF^\top$, and solve for $\bar{K}H$ to have

$$\begin{aligned}\bar{K}H &= F^{-1}(\bar{P} - Q)F^{-\top}R^{-1} \\ &= F^{-1}A^{-1}(\bar{P}_r - I - AQA^\top)A^{-\top}F^{-\top}A^\top A \\ &= A^{-1}F_r^{-1}(\bar{P}_r - Z_r)F_r^{-\top}A ,\end{aligned}$$

where in the second equivalence we replaced $\bar{P}$ with $\bar{P}_r$ (as we did above using (3.10)), and in the third equivalence we simply used matrices from (3.11). We then observe that

$$F_r^{-1}(Z_r - xI)F_r^{-\top} = (1-x)F_r^{-1}F_r^{-\top} + F_r^{-1}G_r B_r G_r^\top F_r^{-\top} ,$$

where

$$\begin{aligned}F_r^{-1}G_r &= \Delta t e^{-\Delta t L_r}\left(e^{\Delta t L_r} - I\right) L_r^{-1} \\ &= \Delta t(e^{-\Delta t L_r} - I)\left(-L_r\right)^{-1} ,\end{aligned}$$

which is equivalent to G_r but for the fact that uses $-L_r$ in the Taylor series. However, this does not change the SED properties, so we have

$$F_r^{-1}(Z_r - xI)F_r^{-\top} = (1 + q - x)I + (1 + \beta_q - x)\,\mathcal{S}_0(2\lambda_f) ,$$

since for the same reason it is $F_r^{-1}F_r^{-\top} = \mathcal{S}(2\lambda_f)$. For the remaining part, instead, we have

$$\begin{aligned}F_r^{-1}(\bar{P}_r - xI)F_r^{-\top} &= x\beta\, F_r^{-1}\mathcal{W}_0(\lambda)F_r^{-\top} \\ &= x\beta\, e^{2\lambda_f/\lambda}\,\mathcal{W}_0(\lambda) ,\end{aligned}$$

where we used Theorem 3.2.4 and the bound $\gamma_k(\epsilon) < e^\epsilon$. The result follows by observing that multiplication by A and A^{-1} increases the distances by a factor of 1.

A.10 Proof of Theorem 3.4.1

To track the estimation error, we begin by stating the true system dynamics and measurement equations

$$\begin{aligned}\boldsymbol{x}_{k+1} &= F\boldsymbol{x}_k + \boldsymbol{w}_k \\ \boldsymbol{y}_{k+1} &= H\boldsymbol{x}_{k+1} + \boldsymbol{v}_{k+1} ,\end{aligned}$$

where F is the true state transition matrix, and $\boldsymbol{w}_k$, $\boldsymbol{v}_{k+1}$ are zero-mean process and measurement noise, respectively. The *pruned* Kalman filter state update is given by

$$\hat{\boldsymbol{x}}_{k+1}^\kappa = \hat{\boldsymbol{x}}_{k+1|k}^\kappa + K_{k+1}^\kappa(\boldsymbol{y}_{k+1} - H\hat{\boldsymbol{x}}_{k+1|k}^\kappa) ,$$

where $\hat{\boldsymbol{x}}_{k+1|k}^\kappa = F^\kappa \hat{\boldsymbol{x}}_k^\kappa$ is the pruned prediction. Substituting the prediction into the update equation

$$\hat{\boldsymbol{x}}_{k+1}^\kappa = F^\kappa \hat{\boldsymbol{x}}_k^\kappa + K_{k+1}^\kappa(\boldsymbol{y}_{k+1} - HF^\kappa \hat{\boldsymbol{x}}_k^\kappa) .$$

Now, substitute the true measurement equation $\boldsymbol{y}_{k+1} = H\boldsymbol{x}_{k+1} + \boldsymbol{v}_{k+1}$, to have

$$\begin{aligned}\hat{\boldsymbol{x}}_{k+1}^{\kappa} &= F^{\kappa}\hat{\boldsymbol{x}}_{k}^{\kappa} + K_{k+1}^{\kappa}(H\boldsymbol{x}_{k+1} + \boldsymbol{v}_{k+1} - HF^{\kappa}\hat{\boldsymbol{x}}_{k}^{\kappa}) \\ &= (I - K_{k+1}^{\kappa}H)F^{\kappa}\hat{\boldsymbol{x}}_{k}^{\kappa} + K_{k+1}^{\kappa}H\boldsymbol{x}_{k+1} + K_{k+1}^{\kappa}\boldsymbol{v}_{k+1} \,.\end{aligned}$$

Then, define the estimation error $\boldsymbol{e}_k = \hat{\boldsymbol{x}}_k^{\kappa} - \boldsymbol{x}_k$, which implies $\hat{\boldsymbol{x}}_k^{\kappa} = \boldsymbol{e}_k + \boldsymbol{x}_k$. Substituting this into the error equation $\boldsymbol{e}_{k+1} = \hat{\boldsymbol{x}}_{k+1}^{\kappa} - \boldsymbol{x}_{k+1}$ provides

$$\begin{aligned}\boldsymbol{e}_{k+1} = (I - K_{k+1}^{\kappa}H)F^{\kappa}(\boldsymbol{e}_k + \boldsymbol{x}_k) \\ + K_{k+1}^{\kappa}H\boldsymbol{x}_{k+1} + K_{k+1}^{\kappa}\boldsymbol{v}_{k+1} - \boldsymbol{x}_{k+1} \,.\end{aligned}$$

Expanding the equation and rearranging terms to isolate $\boldsymbol{e}_k$ we get

$$\begin{aligned}\boldsymbol{e}_{k+1} = (I - K_{k+1}^{\kappa}H)F^{\kappa}\boldsymbol{e}_k + \big[(I - K_{k+1}^{\kappa}H)F^{\kappa}\boldsymbol{x}_k \\ + K_{k+1}^{\kappa}H\boldsymbol{x}_{k+1} + K_{k+1}^{\kappa}\boldsymbol{v}_{k+1} - \boldsymbol{x}_{k+1}\big] \,.\end{aligned}$$

Let the term in the square brackets be $\boldsymbol{q}_k$. Substitute $\boldsymbol{x}_{k+1} = F\boldsymbol{x}_k + \boldsymbol{w}_k$ into $\boldsymbol{q}_k$ to have

$$\begin{aligned}\boldsymbol{q}_k &= (I - K_{k+1}^{\kappa}H)F^{\kappa}\boldsymbol{x}_k \\ &\quad + K_{k+1}^{\kappa}H(F\boldsymbol{x}_k + \boldsymbol{w}_k) + K_{k+1}^{\kappa}\boldsymbol{v}_{k+1} - (F\boldsymbol{x}_k + \boldsymbol{w}_k) \\ &= \big[(I - K_{k+1}^{\kappa}H)F^{\kappa} + K_{k+1}^{\kappa}HF - F\big]\,\boldsymbol{x}_k \\ &\quad + K_{k+1}^{\kappa}H\boldsymbol{w}_k + K_{k+1}^{\kappa}\boldsymbol{v}_{k+1} - \boldsymbol{w}_k \,.\end{aligned}$$

The term multiplying $\boldsymbol{x}_k$ can be simplified as

$$\begin{aligned}(I - K_{k+1}^{\kappa}H)F^{\kappa} + K_{k+1}^{\kappa}HF - F \\ &= F^{\kappa} - K_{k+1}^{\kappa}HF^{\kappa} + K_{k+1}^{\kappa}HF - F \\ &= (F^{\kappa} - F) - K_{k+1}^{\kappa}H(F^{\kappa} - F) \\ &= (I - K_{k+1}^{\kappa}H)(F^{\kappa} - F) \,.\end{aligned}$$

Thus, the estimation error propagates according to the relation

$$\begin{aligned}\boldsymbol{e}_{k+1} &= (I - K_k^{\kappa})F^{\kappa}\boldsymbol{e}_k + \boldsymbol{q}_k \\ \boldsymbol{q}_k &= (I - K_k^{\kappa})(F^{\kappa} - F)\boldsymbol{x}_k + K_k^{\kappa}\boldsymbol{v}_{k+1} - (I - K_k^{\kappa})\boldsymbol{w}_k \,.\end{aligned}$$

If we wish to analyze the expected bias terms $b_k = \mathbb{E}[\boldsymbol{e}_k]$, assuming zero-mean noises ($\mathbb{E}[\boldsymbol{v}_{k+1}] = \boldsymbol{0}$ and $\mathbb{E}[\boldsymbol{w}_k] = \boldsymbol{0}$), we then have the linear system

$$\begin{aligned}b_{k+1} &= (I - K_k^{\kappa})F^{\kappa}b_k + q_k \\ q_k &= (I - K_k^{\kappa})(F^{\kappa} - F)\mathbb{E}[\boldsymbol{x}_k] \,.\end{aligned}$$

By exploiting the theorem assumptions, the following bound can be defined for the norm values

$$\|b_{k+1}\| \leq C_1\|b_k\| + C_2\,\epsilon \,,$$

and the iterative application of the bound further provides

$$\|b_k\| \leq C_1^k\|b_0\| + \tfrac{1-C_1^k}{1-C_1}\,C_2\,\epsilon$$

so that the result is obtained by considering $\lim_{k\to\infty} C_1^k = 0$.

Appendix B

Proof of Theorems in Chap. 4

B.1 Proof of Lemma 4.2.1

Let $\overline{\sigma}(\cdot)$ and $\underline{\sigma}(\cdot)$ denote, respectively, the largest and smallest singular value of a matrix. From the definition of the Kalman gain for subsystem i,

$$K^{(i)}_{k+T_c^{(i)}} = P^{(i)}_{k+T_c^{(i)}|k} H^{(i,i)\top} \Big(H^{(i,i)} P^{(i)}_{k+T_c^{(i)}|k} H^{(i,i)\top} + R^{(i)} \Big)^{-1}. \tag{B.1}$$

Taking matrix norms and using the submultiplicative property gives

$$\|K^{(i)}_{k+T_c^{(i)}}\| \;\le\; \frac{\overline{\sigma}\big(P^{(i)}_{k+T_c^{(i)}|k} H^{(i,i)\top}\big)}{\underline{\sigma}\big(H^{(i,i)} P^{(i)}_{k+T_c^{(i)}|k} H^{(i,i)\top} + R^{(i)}\big)}. \tag{B.2}$$

Since all the involved matrices are symmetric positive definite, their singular values coincide with their eigenvalues. Applying the Rayleigh–Ritz characterization [0],

$$\overline{\sigma}\big(K^{(i)}_{k+T_c^{(i)}}\big) \le \frac{\overline{\sigma}\big(P^{(i)}_{k+T_c^{(i)}|k} H^{(i,i)\top}\big)}{\underline{\sigma}\big(H^{(i,i)} P^{(i)}_{k+T_c^{(i)}|k} H^{(i,i)\top}\big) + \underline{\sigma}\big(R^{(i)}\big)}. \tag{B.3}$$

Now, using Assumption 4.2.1, we have the uniform bounds

$$\underline{q}^{(i)} I \le P^{(i)}_{k+T_c^{(i)}|k} \le \overline{q}^{(i)} I, \qquad \underline{h}^{(i)} I \le \|H^{(i,i)}\| \le \overline{h}^{(i)} I, \qquad r_1 I \le R^{(i)}.$$

Hence,

$$\begin{aligned} \overline{\sigma}\big(P^{(i)}_{k+T_c^{(i)}|k} H^{(i,i)\top}\big) &\le \overline{q}^{(i)}\, \overline{h}^{(i)}, \\ \underline{\sigma}\big(H^{(i,i)} P^{(i)}_{k+T_c^{(i)}|k} H^{(i,i)\top}\big) &\ge \underline{q}^{(i)}\, \underline{h}^{(i)2}, \\ \underline{\sigma}\big(R^{(i)}\big) &\ge r_1. \end{aligned}$$

Substituting these into (B.3) yields

$$\|K^{(i)}_{k+T_c^{(i)}}\| \le \frac{\overline{q}^{(i)}\, \overline{h}^{(i)}}{\underline{q}^{(i)}\, \underline{h}^{(i)2} + r_1} \;=\; \tilde{\kappa}^{(i)}. \tag{B.4}$$

Since the right-hand side is independent of k, it defines a uniform constant upper bound for all update steps of subsystem i.

B.2 Proof of Lemma 4.2.2

Start from the (Joseph) covariance update for subsystem i at its correction instant:

$$P^{(i)}_{k+T_c^{(i)}} = \big(I - K^{(i)}_{k+T_c^{(i)}} H^{(i,i)}\big)\, P^{(i)}_{k+T_c^{(i)}|k}\, \big(I - K^{(i)}_{k+T_c^{(i)}} H^{(i,i)}\big)^\top + K^{(i)}_{k+T_c^{(i)}}\, \tilde{R}^{(i)}_{k+T_c^{(i)}}\, K^{(i)\top}_{k+T_c^{(i)}}, \tag{B.5}$$

with

$$P^{(i)}_{k+T_c^{(i)}|k} = F^{(i,i)}_k P^{(i)}_k F^{(i,i)\top}_k + \tilde{Q}^{(i)}_k, \quad \tilde{Q}^{(i)}_k := Q^{(i)} + \sum_{j\in M,\, j\neq i} F^{(i,j)}_k P^{(j)}_k F^{(i,j)\top}_k,$$

Substituting $P^{(i)}_{k+T_c^{(i)}|k}$ into (B.5) and grouping terms gives

$$P^{(i)}_{k+T_c^{(i)}} = \underbrace{\tilde{F}^{(i,i)} P_k^{(i)} \tilde{F}^{(i,i)\top}}_{\text{propagated term}} + \underbrace{\left(I - KH^{(i,i)}\right) \tilde{Q}_k^{(i)} \left(I - KH^{(i,i)}\right)^\top + K\,\tilde{R}^{(i)}\,K^\top}_{=:\ S_k^{(i)} \succeq 0}, \quad \text{(B.6)}$$

where, for brevity in what follows, $K \equiv K^{(i)}_{k+T_c^{(i)}}$. By Assumption 4.2.1,

$$\tilde{Q}_k^{(i)} \succeq q_1 I + \sum_{j \in M, j \neq i} \underline{p}^{(j)}\, \underline{f}^{(j)2} I \ =: \ q_*^{(i)} I, \qquad q_*^{(i)} > 0,$$

and $\tilde{R}^{(i)} \succeq r_1 I$ with $r_1 > 0$. Hence $S_k^{(i)} \succeq 0$ and, in particular,

$$P^{(i)}_{k+T_c^{(i)}} \ \succeq \ \tilde{F}^{(i,i)} P_k^{(i)} \tilde{F}^{(i,i)\top}. \quad \text{(B.7)}$$

By Löwner order inversion for positive definite matrices (monotonicity of the matrix inverse),

$$\left(P^{(i)}_{k+T_c^{(i)}}\right)^{-1} \ \preceq \ \left(\tilde{F}^{(i,i)} P_k^{(i)} \tilde{F}^{(i,i)\top}\right)^{-1} = \tilde{F}^{(i,i)\,-\top} P_k^{(i)\,-1} \tilde{F}^{(i,i)\,-1},$$

which implies the non-expansive inequality

$$\tilde{F}^{(i,i)\top} P^{(i)\,-1}_{k+T_c^{(i)}} \tilde{F}^{(i,i)} \ \preceq \ P_k^{(i)\,-1}. \quad \text{(B.8)}$$

To obtain a strict contraction factor, note from (B.6) that

$$P^{(i)}_{k+T_c^{(i)}} \ = \ \tilde{F}^{(i,i)} P_k^{(i)} \tilde{F}^{(i,i)\top} \ + \ S_k^{(i)}, \quad S_k^{(i)} \succeq 0.$$

Using the matrix inequality

$$(X+Y)^{-1} \ \preceq \ \left(I + X^{-1/2} Y X^{-1/2}\right)^{-1} X^{-1} \ \preceq \ \frac{1}{1+\mu}\, X^{-1} \quad \text{whenever } Y \succeq \mu X \text{ for some } \mu > 0,$$

with $X = \tilde{F}^{(i,i)} P_k^{(i)} \tilde{F}^{(i,i)\top}$ and $Y = S_k^{(i)}$, we only need a uniform $\mu > 0$ such that $S_k^{(i)} \ \succeq \ \mu\, \tilde{F}^{(i,i)} P_k^{(i)} \tilde{F}^{(i,i)\top}$. By Assumption 4.2.1, $\|F_k^{(i,i)}\| \ \leq \ \overline{f}^{(i)}$ and $P_k^{(i)} \ \preceq \ \overline{p}^{(i)} I$, hence $\tilde{F}^{(i,i)} P_k^{(i)} \tilde{F}^{(i,i)\top} \ \preceq \ \overline{p}^{(i)} \overline{f}^{(i)2} I$. On the other hand, since $\tilde{Q}_k^{(i)} \ \succeq \ q_*^{(i)} I$, we have $S_k^{(i)} \succeq 0 + 0 \succeq 0$ and (by inserting the positive terms without attenuation)

$$S_k^{(i)} \succeq q_*^{(i)} I \quad \Rightarrow \quad S_k^{(i)} \succeq \frac{q_*^{(i)}}{\overline{p}^{(i)} \overline{f}^{(i)2}}\, \tilde{F}^{(i,i)} P_k^{(i)} \tilde{F}^{(i,i)\top}.$$

Therefore we can choose

$$\mu \ = \ \frac{q_*^{(i)}}{\overline{p}^{(i)} \overline{f}^{(i)2}} \ = \ \frac{q_1 + \sum_{j \in M,\, j \neq i} \underline{p}^{(j)} \underline{f}^{(j)2}}{\overline{p}^{(i)} \overline{f}^{(i)2}} \ > \ 0.$$

Applying the inequality above yields

$$\left(P^{(i)}_{k+T_c^{(i)}}\right)^{-1} \ \preceq \ \frac{1}{1+\mu}\, \tilde{F}^{(i,i)\,-\top} P_k^{(i)\,-1} \tilde{F}^{(i,i)\,-1},$$

and multiplying on the left/right by $\tilde{F}^{(i,i)\top}$ and $\tilde{F}^{(i,i)}$ gives

$$\tilde{F}^{(i,i)\top} P^{(i)\,-1}_{k+T_c^{(i)}} \tilde{F}^{(i,i)} \ \preceq \ \frac{1}{1+\mu}\, P_k^{(i)\,-1}.$$

Defining

$$1 - \gamma^{(i)} \ := \ \frac{1}{1+\mu} \ = \ \left(1 + \frac{q_1 + \sum_{j \in M,\, j \neq i} \underline{p}^{(j)} \underline{f}^{(j)2}}{\overline{p}^{(i)} \overline{f}^{(i)2}}\right)^{-1} \in (0,1),$$

we obtain exactly (4.19).

B.3 Proof of Theorem 4.2.1

By definition, $V^{(i)}(\boldsymbol{e}_k^{(i)}) = \boldsymbol{e}_k^{(i)\,T} P_k^{(i)\,-1} \boldsymbol{e}_k^{(i)}$, and Assumption 4.2.1 yields

$$\frac{1}{\overline{p}^{(i)}} \|\boldsymbol{e}_k^{(i)}\|^2 \le V^{(i)}(\boldsymbol{e}_k^{(i)}) \le \frac{1}{\underline{p}^{(i)}} \|\boldsymbol{e}_k^{(i)}\|^2. \tag{B.9}$$

Using the linear error recursion at correction instants,

$$\boldsymbol{e}_{k+T_c^{(i)}}^{(i)} = \tilde{F}^{(i,i)}\, \boldsymbol{e}_k^{(i)} + \boldsymbol{n}_{k+T_c^{(i)}}^{(i)},$$

where $)\boldsymbol{n}_{k+T_c^{(i)}}^{(i)} = \left(I - K_{k+T_c^{(i)}}^{(i)} H^{(i,i)}\right) w_k^{(i)} + K_{k+T_c^{(i)}}^{(i)} v_{k+T_c^{(i)}}^{(i)}$, we compute the Lyapunov increment

$$\begin{aligned}
\Delta V^{(i)} &:= V^{(i)}(\boldsymbol{e}_{k+T_c^{(i)}}^{(i)}) - V^{(i)}(\boldsymbol{e}_k^{(i)}) \\
&= \boldsymbol{e}_k^{(i)\,T} \left(\tilde{F}^{(i,i)\,T} P_{k+T_c^{(i)}}^{(i)\,-1} \tilde{F}^{(i,i)} - P_k^{(i)\,-1} \right) \boldsymbol{e}_k^{(i)} + \boldsymbol{n}_{k+T_c^{(i)}}^{(i)\,T} P_{k+T_c^{(i)}}^{(i)\,-1} \boldsymbol{n}_{k+T_c^{(i)}}^{(i)} \\
&\quad + 2\, \boldsymbol{n}_{k+T_c^{(i)}}^{(i)\,T} P_{k+T_c^{(i)}}^{(i)\,-1} \tilde{F}^{(i,i)} \boldsymbol{e}_k^{(i)}.
\end{aligned} \tag{B.10}$$

By Lemma 4.2.2, there exists $\gamma^{(i)} \in (0,1)$ such that

$$\tilde{F}^{(i,i)\,T} P_{k+T_c^{(i)}}^{(i)\,-1} \tilde{F}^{(i,i)} - P_k^{(i)\,-1} \;\le\; -\gamma^{(i)}\, P_k^{(i)\,-1},$$

hence the first term in (B.10) is upper bounded by $-\gamma^{(i)} V^{(i)}(\boldsymbol{e}_k^{(i)})$. For the noise term, Assumption 4.2.1 and Lemma 4.2.1 imply

$$\|I - K_{k+T_c^{(i)}}^{(i)} H^{(i,i)}\| \le \frac{\overline{p}^{(i)}}{\underline{q}^{(i)}}, \qquad \|K_{k+T_c^{(i)}}^{(i)}\| \le \tilde{\kappa},$$

thus

$$\|\boldsymbol{n}_{k+T_c^{(i)}}^{(i)}\| \le \frac{\overline{p}^{(i)}}{\underline{q}^{(i)}}\, \overline{\boldsymbol{w}}^{(i)} + \tilde{\kappa}\, \overline{\boldsymbol{v}}^{(i)}. \tag{B.11}$$

Using $P_{k+T_c^{(i)}}^{(i)\,-1} \preceq \underline{p}^{(i)\,-1} I$ gives

$$\boldsymbol{n}_{k+T_c^{(i)}}^{(i)\,T} P_{k+T_c^{(i)}}^{(i)\,-1} \boldsymbol{n}_{k+T_c^{(i)}}^{(i)} \le \frac{1}{\underline{p}^{(i)}} \left(\frac{\overline{p}^{(i)}}{\underline{q}^{(i)}}\, \overline{\boldsymbol{w}}^{(i)} + \tilde{\kappa}\, \overline{\boldsymbol{v}}^{(i)} \right)^2 = \rho^{(i)}(\tilde{\kappa}).$$

For the cross term, by Cauchy–Schwarz and submultiplicativity,

$$2\, \boldsymbol{n}^\top P^{-1} \tilde{F}\, \boldsymbol{e} \;\le\; \boldsymbol{n}^\top P^{-1} \boldsymbol{n} + \boldsymbol{e}^\top \tilde{F}^\top P^{-1} \tilde{F}\, \boldsymbol{e} \;\le\; \rho^{(i)}(\tilde{\kappa}) + (1 - \gamma^{(i)})\, V^{(i)}(\boldsymbol{e}_k^{(i)}),$$

where the last inequality uses Lemma 4.2.2. Combining the three bounds in (B.10) yields

$$\Delta V^{(i)} \le -\gamma^{(i)} V^{(i)}(\boldsymbol{e}_k^{(i)}) + \rho^{(i)}(\tilde{\kappa}) + (1-\gamma^{(i)}) V^{(i)}(\boldsymbol{e}_k^{(i)}) = \underbrace{-\frac{\gamma^{(i)}}{\psi^{(i)}}}_{-\xi^{(i)} - (1-\gamma^{(i)})} V^{(i)}(\boldsymbol{e}_k^{(i)}) + \rho^{(i)}(\tilde{\kappa}),$$

which is the standard form

$$\Delta V^{(i)} \le \xi^{(i)}\, V^{(i)}(\boldsymbol{e}_k^{(i)}) + \rho^{(i)}(\tilde{\kappa}), \qquad \xi^{(i)} := \frac{\gamma^{(i)}(1 - \psi^{(i)})}{\psi^{(i)}} \in (-1, 0),\ \psi^{(i)} > 1.$$

Iterating gives

$$V^{(i)}(\boldsymbol{e}_{k+T_c^{(i)}}) \le (1 + \xi^{(i)})^{k+1} V^{(i)}(\boldsymbol{e}_0^{(i)}) + \frac{1 - (1 + \xi^{(i)})^{k+1}}{|\xi^{(i)}|}\, \rho^{(i)}(\tilde{\kappa}).$$

Finally, using (B.9) yields the bound (4.20). In the noise–free case, $\rho^{(i)}(\tilde{\kappa}) = 0$ and the exponential decay follows directly. This completes the proof.

B.4 Proof of Theorem 4.2.2

Assume that for $i \in \mathcal{V}$ and $k > 1$,

$$\mathbb{E}\left\{\left(\hat{\boldsymbol{x}}_k^{(i)} - \boldsymbol{x}_k^{(i)}\right)\left(\hat{\boldsymbol{x}}_k^{(i)} - \boldsymbol{x}_k^{(i)}\right)^T\right\} \leq P_k^{(i)}$$

holds. We define the estimated error as

$$\boldsymbol{e}_{k+T_c^{(i)}|k}^{(i)} = F_k^{(i,i)}(\hat{\boldsymbol{x}}_k^{(i)} - \boldsymbol{x}_k^{(i)}) + \sum_{j \in \mathcal{N}_i^{in}} F_k^{(i,j)}\left(\hat{\boldsymbol{x}}_{m_{ij}}^{(j)} - \boldsymbol{x}_k^{(j)}\right) - \boldsymbol{w}_k^{(i)}$$

Let us define a vector $\boldsymbol{\Delta}_k^{(i,j)}$ such that $\boldsymbol{\Delta}_k^{(i,j)} = \boldsymbol{x}_{m_{ij}}^{(j)} - \boldsymbol{x}_k^{(j)}$.

We define positive semidefinite and symmetrical matrix $D_{m_{ij}}^{(i,j)}$ following (4.26) for all $k \in [m_{ij}, m_{ij} + T_c^{(i)}]$. Then, since $\mathbb{E}\left\{\boldsymbol{e}_k^{(i)}\boldsymbol{\omega}_k^{(i)}\right\} = \mathbb{E}\left\{\boldsymbol{e}_{m_{ij}}^{(j)}\boldsymbol{\omega}_k^{(i)}\right\} = 0$, it is possible to write:

$$\begin{aligned}
&\mathbb{E}\left\{\left(\boldsymbol{e}_{k+T_c^{(i)}|k}^{(i)}\right)\left(\boldsymbol{e}_{k+T_c^{(i)}|k}^{(i)}\right)^T\right\} = \\
&= \mathbb{E}\left\{\left(F_k^{(i,i)}\boldsymbol{e}_k^{(i)} + \sum_{j \in \mathcal{N}_i^{in}} F_k^{(i,j)}\left(\boldsymbol{e}_{m_{ij}}^{(j)} + \boldsymbol{\Delta}_k^{(i,j)}\right)\right)\right. \\
&\qquad \left.\left(F_k^{(i,i)}\boldsymbol{e}_k^{(i)}) + \sum_{j \in \mathcal{N}_i^{in}} F_k^{(i,j)}\left(\boldsymbol{e}_{m_{ij}}^{(j)} + \boldsymbol{\Delta}_k^{(i,j)}\right)\right)^T\right\} + Q^{(i)}
\end{aligned}$$

Exploiting the Young's inequality [0], where $A\,B^T + B\,A^T \leq \theta A\,A^T + 1/\theta B\,B^T$, $\forall \theta > 0$, $\forall A, B \in \mathbb{R}^{n \times n}$ and setting $\theta = 1$, we obtain $(A+B)(A+B)^T \leq 2AA^T + 2BB^T$. In case of $(\sum_{j \in \mathcal{N}} A_j)(\sum_{j \in \mathcal{N}} A_j)^T \leq |\mathcal{N}| \sum_{j \in \mathcal{N}} A_j A_j^T$.

Therefore, we get:

$$\begin{aligned}
&\mathbb{E}\left\{\left(\boldsymbol{e}_{k+T_c^{(i)}|k}^{(i)}\right)\left(\boldsymbol{e}_{k+T_c^{(i)}|k}^{(i)}\right)^T\right\} \leq 2F_k^{(i,i)}P_k^{(i,i)}F_k^{(i,i)T} + \\
&\quad 2\mathbb{E}\left\{\sum_{j \in \mathcal{N}_i^{in}} F_k^{(i,j)}\left(\boldsymbol{e}_{m_{ij}}^{(j)} + \boldsymbol{\Delta}_k^{(i,j)}\right)\right. \\
&\qquad \left.\left(\sum_{j \in \mathcal{N}_i^{in}} F_k^{(i,j)}\left(\boldsymbol{e}_{m_{ij}}^{(j)} + \boldsymbol{\Delta}_k^{(i,j)}\right)\right)^T\right\} + Q^{(i)} \\
&\leq 2\,F_k^{(i,i)}P_k^{(i,i)}F_k^{(i,i)T} + Q^{(i)} + 2|\mathcal{N}_i| \cdot \\
&\quad \sum_{j \in \mathcal{N}_i^{in}} F_k^{(i,j)}\mathbb{E}\left(\boldsymbol{e}_{m_{ij}}^{(j)} + \boldsymbol{\Delta}_k^{(i,j)}\left(\boldsymbol{e}_{m_{ij}}^{(j)} + \boldsymbol{\Delta}_k^{(i,j)}\right)^T\right)F_k^{(i,j)T} \\
&\leq 2\,F_k^{(i,i)}P_k^{(i,i)}F_k^{(i,i)T} + \\
&\quad 4|\mathcal{N}_i| \sum_{j \in \mathcal{N}_i^{in}} F_k^{(i,j)}\left(P_{m_{ij}}^{(i,j)} + D_{m_{ij}}^{(i,j)}\right)F_k^{(i,j)T} + Q^{(i)}
\end{aligned}$$

At this point, we can set $\Pi_k^{(i,j)} = P_{m_{ij}}^{(i,j)} + D_{m_{ij}}^{(i,j)}$, and

$$\tilde{Q}_k^{(i)} = F_k^{(i,i)}P_k^{(i,i)}F_k^{(i,i)T} + 4|\mathcal{N}_i| \sum_{j \in \mathcal{N}_i^{in}} F_k^{(i,j)}\Pi_k^{(i,j)}F_k^{(i,j)T} + Q^{(i)},$$

then we have that:

$$\mathbb{E}\left\{\left(\boldsymbol{e}_{k+T_c^{(i)}|k}^{(i)}\right)\left(\boldsymbol{e}_{k+T_c^{(i)}|k}^{(i)}\right)^T\right\} \leq F_k^{(i,i)}P_k^{(i,i)}F_k^{(i,i)T} + \tilde{Q}_k^{(i)} =$$

$$= P^{(i,i)}_{k+1|k}$$

Now, notice that

$$\boldsymbol{e}^{(i)}_{k+T_c^{(i)}} = \left(I - K_k^{(i)} H_k^{(i,i)}\right)\left(\hat{\boldsymbol{x}}^{(i)}_{k+T_c^{(i)}|k} - \boldsymbol{x}^{(i)}_{k+T_c^{(i)}}\right) + K_k^{(i)} \boldsymbol{v}^{(i)}_{k+T_c^{(i)}}$$

we have that:

$$\mathbb{E}\left\{\left(\boldsymbol{e}^{(i)}_{k+T_c^{(i)}}\right)\left(\boldsymbol{e}^{(i)}_{k+T_c^{(i)}}\right)^T\right\} \leq \left(I - K_k^{(i)} H_k^{(i,i)}\right) P^{(i,i)}_{k+1|k} \cdot \left(I - K_k^{(i)} H_k^{(i,i)}\right) + K_k^{(i)} R^{(i)} K_k^{(i)T}$$

Knowing (4.8), we obtain

$$\mathbb{E}\left\{\left(\boldsymbol{e}^{(i)}_{k+T_c^{(i)}}\right)\left(\boldsymbol{e}^{(i)}_{k+T_c^{(i)}}\right)^T\right\} \leq (I - K_k^{(i)} H_k^{(i,i)}) P^{(i,i)}_{k+1|k} = P^{(i,i)}_k$$

and this completes the proof.

Appendix C

Proof of theorem in Chap. 4

C.1 Proof of the existence and uniqueness of K^*

By manipulating (7.8), we obtain the following inequality:

$$2K < \left(\frac{1}{\sqrt{2K+2C}} + d\right)^{-2}.$$

Adding $2C$ to both sides of the inequality, we have:

$$2K + 2C < \left(\frac{1}{\sqrt{2K+2C}} + d\right)^{-2} + 2C.$$

Substituting $\alpha = \sqrt{2K+2C}$, we can rewrite the inequality as

$$\alpha^2 < \left(\frac{1}{\alpha} + d\right)^{-2} + 2C.$$

For simplicity, define $f(\alpha) = \alpha^2$ and $g(\alpha) = \left(\frac{1}{\alpha} + d\right)^{-2} + 2C$.
We will show that there exists a unique value α^* such that $f(\alpha^*) = g(\alpha^*)$, and

$$\begin{cases} f(\alpha) < g(\alpha) & \text{if } \alpha < \alpha^* \\ f(\alpha) > g(\alpha) & \text{if } \alpha > \alpha^* \end{cases}$$

when $C, d \in (0, \infty)$. Note that α can only take positive values, so our interval of interest is $\alpha \in (0, \infty)$.

To prove this, we first show that $f(\alpha)$ and $g(\alpha)$ are continuous and differentiable in the interval $\alpha \in (0, \infty)$. $f(\alpha)$ is defined for every α and is therefore a continuous function, whereas $g(\alpha)$ is defined for $\alpha \neq -\frac{1}{d}$. Since d is a positive scalar and α only takes positive values, $g(\alpha)$ is also continuous in the interval of interest.

Next, consider their first order partial derivatives:

$$\begin{cases} \frac{\partial f(\alpha)}{\partial \alpha} = 2\alpha \\ \frac{\partial g(\alpha)}{\partial \alpha} = -\frac{2d}{(1+\alpha d)^3} + \frac{2}{(1+\alpha d)^2} \end{cases}.$$

Again, $\frac{\partial f(\alpha)}{\partial \alpha}$ is defined for every α and is therefore continuous. $\frac{\partial g(\alpha)}{\partial \alpha}$ is defined for $\alpha \neq -\frac{1}{d}$. Since d is positive and α only takes positive values, $\frac{\partial g(\alpha)}{\partial \alpha}$ is continuous in the interval of interest. Hence, both functions are continuous and differentiable in the interval.

Furthermore, by mathematical manipulation, it can be shown that both derivatives are strictly greater than 0 in the interval of interest, implying that they are monotonically increasing.

Through further mathematical manipulations, we can establish that $\frac{\partial f(\alpha)}{\partial \alpha} < \frac{\partial g(\alpha)}{\partial \alpha}$ for $\alpha \in (0, \infty)$. This implies that in the interval of interest, the function $f(\alpha)$ grows faster than the function $g(\alpha)$, and therefore $f(\alpha)$ and $g(\alpha)$ can intersect at most in one point.

Let us consider the limits of $f(\alpha)$ and $g(\alpha)$ as α tends to 0 from the right:

$$\lim_{\alpha \to 0^+} f(\alpha) = 0\,,$$

$$\lim_{\alpha \to 0^+} g(\alpha) = 2C$$

thus, it is clear that $f(0^+) < g(0^+)$. Next, consider the limits of $f(\alpha)$ and $g(\alpha)$ as α tends to $+\infty$, i.e.,

$$\lim_{\alpha \to +\infty} f(\alpha) = +\infty \,,$$

$$\lim_{\alpha \to +\infty} g(\alpha) = \frac{1}{d^2} + 2C \,.$$

Since C and d are both in $(0, \infty)$, it is clear that $f(\infty) > g(\infty)$. By the intermediate value theorem, we conclude that there exists at least one α^* such that $f(\alpha^*) = g(\alpha^*)$ in the interval $\alpha \in (0, \infty)$. Combining this with the fact that $\alpha^2 = 2K + 2C$, and $K, C \in (0, \infty)$, we conclude that for a given transmission cost C, there exists a unique K^* such that

$$K^* = \frac{\alpha^{*2} - 2C}{2}.$$

This completes the proof.

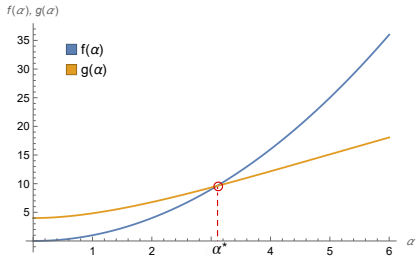


Figure C.1: $f(\alpha)$ and $g(\alpha)$ plot setting $d = 1$ and $C = 2$

.