AGU PUBLICATIONS

Water Resources Research

RESEARCH ARTICLE

10.1002/2014WR016067

A novel infrastructure modularity index for the segmentation of water distribution networks

O. Giustolisi[1] **and L. Ridolfi**[2]

[1]Department of Civil and Environmental Engineering, Politecnico di Bari, Bari, Italy, [2]Department of Environmental, Land, and Infrastructure Engineering, Politecnico di Torino, Torino, Italy

Key Points:
- A general quantitative approach for water distribution network segmentation
- No resolution limit
- Proposed approach is discussed and tested on real water networks

Correspondence to:
L. Ridolfi,
luca.ridolfi@polito.it

Citation:
Giustolisi, O., and L. Ridolfi (2014), A novel infrastructure modularity index for the segmentation of water distribution networks, *Water Resour. Res.*, *50*, 7648–7661, doi:10.1002/2014WR016067.

Received 1 JUL 2014
Accepted 4 SEP 2014
Accepted article online 8 SEP 2014
Published online 2 OCT 2014

Abstract The search for suitable segmentations is a challenging and urgent issue for the analysis, planning and management of complex water distribution networks (WDNs). In fact, complex and large size hydraulic systems require the division into modules in order to simplify the analysis and the management tasks. In the complex network theory, modularity index has been proposed as a measure of the strength of the network division into modules and its maximization is used in order to identify community of nodes (i.e., modules) which are characterized by strong interconnections. Nevertheless, modularity index needs to be revised considering the specificity of the hydraulic systems as infrastructure systems. To this aim, the classic modularity index has been recently modified and tailored for WDNs. Nevertheless, the WDN-oriented modularity is affected by the resolution limit stemming from classic modularity index. Such a limit hampers the identification/design of small modules and this is a major drawback for technical tasks requiring a detailed resolution of the network segmentation. In order to get over this problem, we propose a novel infrastructure modularity index that is not affected by the resolution limit of the classic one. The rationale and good features of the proposed index are theoretically demonstrated and discussed using two real hydraulic networks.

1. Introduction

In the last decade, studies about complex networks are significantly increased [e.g., *Albert and Barabasi*, 2002; *Boccaletti et al.*, 2006; *Barrat et al.*, 2008; *Newman*, 2010]. This fact is caused by the availability of a huge amount of data coming from different kinds of networks, going from nature (e.g., foodweb and neural connections) to economy (trade networks), from infrastructures (e.g., water, gas, transportation and electrical networks) to social networks (e.g., World Wide Web). Moreover, the pervasive presence of networks is continuously encouraging researchers to develop novel mathematical paradigms and tools in order to better understand network behavior and features [e.g., *Havlin et al.*, 2012].

In the astonishing variety of the natural and man-made networks, the infrastructure ones are a relevant subset. They are continuously used by the people and are fundamental in the modern society. The infrastructure networks are becoming more and more complex and interconnected with the social development and this fact raises a number of crucial issues to their analysis, planning and management. The distribution networks are a remarkable and widespread example of infrastructure networks which convey water, electricity, gas, data, etc. In the last years, several works were motivated by the necessity of facing the increased complexity of infrastructure networks. Some researchers investigated methods which were developed in the complex network theory (CNT) and results were promising. We here report the studies on the topology of transportation networks [e.g., *Sienkiewicz and Holyst*, 2005; *Kurant and Thiran*, 2006; *von Ferber et al.*, 2009; *Lin and Ban*, 2013], on the structure and resilience of water distribution networks [*Yazdani and Jeffrey*, 2011, 2012] and on the vulnerability of power grids [*Albert et al.*, 2004; *Arianos et al.*, 2009; *Bompard et al.*, 2012].

The infrastructure networks can significantly benefit from approaches and tools of the CNT. Nevertheless, CNT-based conceptual tools are not specifically intended for infrastructure networks and for this reason it is necessary to pay attention to the correct technical transfer to different kind of networks. The infrastructure networks are in fact characterized by specificities – strong spatial constrains (two-dimensionality, urban structure, demand locations, distance-based costs, etc.), material links between nodes (differently from the immaterial links which are typical of the social networks), occurrence of a wide set of possible devices installed on links (e.g., pumping stations, valves, metering devices, etc.) – which make them different from other categories of networks. *Barthélemy* [2011] recently discussed the need of avoiding the uncritical application of complex network studies and tools for spatial networks. In fact, the specificity of infrastructure networks needs to be considered in developing *ad hoc* methods stemming from the emerging ideas of CNT studies. Novel studies are therefore necessary in order to adapt CNT methods, which are generally developed considering immaterial networks (e.g., social networks), to the specific nature of infrastructure networks.

We here focus on the segmentation of water distribution networks (WDNs), which is a relevant issue of the modern water system analysis, planning and management. Segmentation of WDNs deals with the design or analysis of system modules/segments [*Giustolisi and Ridolfi*, 2014]. The segmentation design can be performed considering several network characteristics: e.g., topological (the connection degree), structural (pipe features like diameter, length, material, age, etc.) or associated to the network fluxes (water flows, leakages, etc.). The segmentation (also named districtualization or sectoralization) strategy was recently used for several purposes: system analysis, model calibration, efficient metering planning, monitoring optimization for early warning, management cost minimization, network resilience and reliability enhancement, etc. [e.g., *Deuerlein*, 2008; *Perelman and Ostfeld*, 2011; *Zecchin et al.*, 2012; *Alvisi and Franchini*, 2014; *Ferrari et al.*, 2014]. Even though the segmentation of WDNs is beneficial from many technical purposes, wide-ranging methods are however not available.

In order to develop general methods for segmentation of WDNs, some preliminary works recently proposed to refer to community detection methods developed in the analysis of complex social networks [*Scibetta et al.*, 2013; *Diao et al.*, 2013]. The key idea was to interpret a module of WDNs as a community to be identified. Community detection methods aim at identifying separated or overlapped groups of nodes into the network whose information exchanges are greater than with other groups [*Fortunato*, 2010]. In particular, among the community detection methods, the approach based on the modularity concept [*Newman and Girvan*, 2004] was used. The related modularity index is based on the comparison of the real networks with a random network realization sharing some general topological characteristics, i.e., number of links and mean degree of nodes. The index is maximized in order to identify the communities which are the network modules. The modularity index is able to embed different network features and, moreover, efficient optimization (maximization) algorithms for very large size problems are available. This fact explains the successful application to a wide range of cases [*Fortunato*, 2010].

In a previous work [*Giustolisi and Ridolfi*, 2014], we proposed a modularity-based index tailored for WDNs and a multiobjective segmentation design for a general integrated planning. In order to correctly perform a technical transfer of the CNT method to the specific infrastructure network, we focused on: (i) the tailoring of the topology-based and the weight-based (embedding pipe properties) classic modularity index [*Newman and Girvan*, 2004] by means of the topological incidence matrix, commonly used to describe the WDN, (ii) the developing a cut position-sensitive modularity index in order to account for the actual position along pipes (i.e., close to ending nodes) of devices generating segments/modules, and (iii) the framing the modularity maximization into a multiobjective strategy, with the aim to develop a decision support tool for segmentation suitable for an integrated, dynamical planning whose resolution increases over time.

Nevertheless, the WDN-oriented modularity index is affected by a resolution limit that increases with network size. This limit stems from classic modularity index [*Fortunato and Barthélemy*, 2007] and bounds the detection of modules with respect to the size. Therefore, the design of modules which are smaller than a threshold depending on the number of links of the network is not possible. This is a major drawback for technical tasks requiring a detailed resolution of the network segmentation, as for example in the case of the isolation valve system design.

The purpose of the present work is to overcome the resolution limit of classic modularity index by proposing a novel infrastructure modularity index starting from WDN-oriented which was developed in *Giustolisi and Ridolfi* [2014]. To this aim, we demonstrate that the classic modularity maximization is a special tradeoff solution of the general two-objective optimization problem that involves the two components of the classic modularity index. In this framework, we explain the origin of the resolution limit. Then, we propose the novel infrastructure modularity index and theoretically demonstrate the overcoming of the resolution limit. Finally, the new metric is discussed in the framework of the general two-objective optimization, in order to show some special features with respect to the classic one.

The work is organized as follows. In the next section, classic modularity index and its tailoring for WDNs are briefly recalled. In the third section, the maximization of classic modularity index is interpreted in a general two-objective optimization framework in order to discuss the origin of the resolution limit. The novel infrastructure modularity index is presented in the fourth section, where the overcoming of the resolution limit is demonstrated and features with respect to the classical index are described using the two-objective optimization framework. The fifth section reports the segmentation of two real networks in order to show the key characteristics of the novel infrastructure index versus the classic index and, finally, the last section draws the conclusions.

2. Modularity Index Tailored for WDNs

The modularity, Q, is a measure of the strength of a network or graph division into communities/modules. Hereinafter, we will use the word "module" being more usual for infrastructure networks. Modularity is defined as [*Newman and Girvan*, 2004]

$$Q=\frac{1}{2n_l}\sum_{ij}\left(A_{ij}-P_{ij}\right)\delta\left(\mathrm{M}_i,\mathrm{M}_j\right)=\frac{1}{2n_l}\sum_{ij}\left(A_{ij}-\frac{k_ik_j}{2n_l}\right)\delta\left(\mathrm{M}_i,\mathrm{M}_j\right) \tag{1}$$

where n_l is the number of network links, A_{ij} are the elements of the adjacency matrix, P_{ij} is the expected fraction of links between nodes i and j in the null/random network (i.e., the expected number of links in the network if they were randomly distributed), M_i is the identifier of network modules, δ is the function to apply the summation to the elements of the same module (i.e., $\delta=1$ if $\mathrm{M}_j=\mathrm{M}_i$ and $\delta=0$ otherwise), and summation runs on all the possible node couples (i,j), with $i\neq j$. In equation (1), the expected fraction P_{ij} is computed using node degree: i.e., k_i (k_j) is the degree of the i-th (j-th) node, being degree the number of links incident in the node.

The formulation in equation (1) of the modularity index is not always appropriate for WDNs because the adjacency matrix does not store the information on parallel links (i.e., pipes of hydraulic systems) which are common in those networks [*Giustolisi and Ridolfi*, 2014]. For this reason, we adopt the equivalent formulation of the modularity index

$$\begin{aligned} Q&=\sum_{m=1}^{n_m}e_{mm}-\sum_{m=1}^{n_m}a_m^2 \\ a_m&=\sum_i\frac{k_i}{2n_p}\quad\forall i\in\mathrm{M}_m \end{aligned} \tag{2}$$

where e_{mm} is the fraction of the pipes with both the end nodes belonging to the m-th module, n_m is the number of network modules, n_p ($=n_l$) is the number of network pipes (hereinafter "pipe" is used instead of "link"), and a_m is the fraction of pipes having at least one end node in the module m, half counting the pipes dividing modules. Therefore, a_m is half the summation, divided by n_p, of the number of incident pipes (degree) in nodes falling in the module m. It is possible to write

$$\sum_{m=1}^{n_m}e_{mm}=1-\frac{n_c}{n_p} \tag{3}$$

where n_c is the number of pipes linking modules of the infrastructure, namely the number of "cuts" in the network.

2.1. Topological Matrix Description of the Infrastructure Network

WDNs are usually described by the general topological incidence matrix $\bar{\mathbf{A}}_{pn}$, [see e.g., *Giustolisi et al.*, 2008]. The matrix $\bar{\mathbf{A}}_{pn}$ is composed by n_p rows, each corresponding to one pipe of the hydraulic system, having two elements different from zero corresponding to the i-th and j-th ending nodes. The values of the two nonnull elements in each row of $\bar{\mathbf{A}}_{pn}$ are {1,–1} depending on the assumed positive flow direction in the pipe. The number of columns of the matrix $\bar{\mathbf{A}}_{pn}$ is n_n, that is the number of nodes including the water source(s), i.e., reservoirs.

Herein we will refer to the absolute value of the general topological incidence matrix, $|\bar{\mathbf{A}}_{pn}|$, and to the unitary n_p-size vector $\mathbf{u}_p$. By definition of topological incidence matrix, the summation by columns $|\bar{\mathbf{A}}_{pn}|$ – i.e., $|\bar{\mathbf{A}}_{pn}|^T\mathbf{u}_p$ (superscript ^{T} indicates the transpose) – is the vector of nodal degrees.

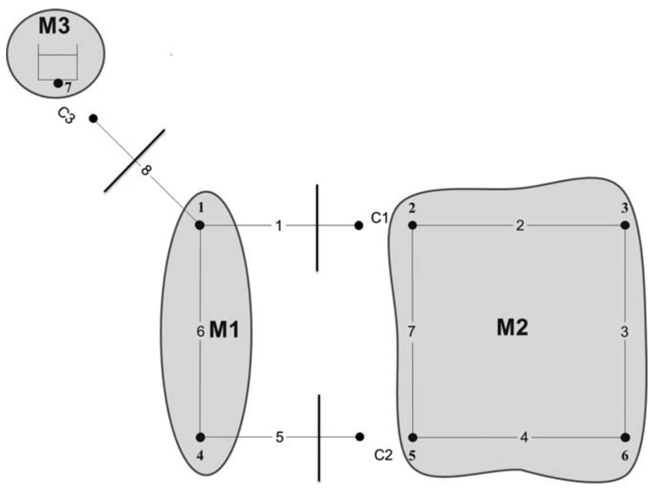

Figure 1. Network with three actual cuts along pipes 1, 5, and 8 close to one ending node. According to the classical modularity evaluation, the pipes are counted half for each adjacent module and $Q = 0.1328$.

The adjacency matrix, $\bar{\mathbf{A}}_{nn}$, of the graph related to the network is the Boolean version of the matrix product $\bar{\mathbf{A}}_{pn}^{T} \times \bar{\mathbf{A}}_{pn}$ with null diagonal, being node loops not allowed in the topological representation of WDNs.

The modularity index for WDNs can be then written using the general topological matrix as

$$Q=1-\frac{n_c}{n_p}-\sum_{m=1}^{n_m}\left[\sum_{i=1}^{n_n}\frac{\left(|\bar{\mathbf{A}}_{pn}^{T}|\mathbf{u}_p\right)_i\delta(\mathrm{M}_m,\mathrm{M}_i)}{2n_p}\right]^2 \quad (4)$$

where $\mathrm{M_i}$ indicates the module where *i*-th pipe falls and symbol $(\cdot)_i$ indicates the *i*-th component of the vector inside the round brackets. In equation (4), the term

$$\frac{|\bar{\mathbf{A}}_{pn}^{T}|\mathbf{u}_p}{2} \quad (5)$$

is half of the number of pipes incident in each node. Therefore, the squared term in equation (4) states that, for each module, we consider the number of pipes with the two ending nodes belonging to that module and half of the number of pipes separating it from the other (n_m – 1) modules.

2.2. Actual Representation of the Pipe Cut in Infrastructure Networks

In any modularity-based index, virtual pipe cuts are implicitly assumed in the middle of pipes. However, this assumption is misleading for WDNs being significant the position of devices which actually segment the infrastructure networks. In fact, the virtual cuts become real devices installed into the networks being divided into modules and those devices are usually installed close to the ending nodes of the pipes. The installation of devices close to one of the two ending nodes of pipes corresponds to two different technical solutions which should correspond to two different values of the modularity index, i.e., the metric should be cut position-sensitive. Instead, the metrics based on the classic modularity (e.g., see equation (4)) neglect the actual position of devices along pipes; therefore, they are not cut position-sensitive [*Giustolisi and Ridolfi*, 2014]. For example, Figure 1 shows a simple WDN where three cuts are located close to specific nodes of the pipes. Computing the classic modularity index, i.e., not considering the actual position of the cuts, $Q = 0.1328$ is obtained.

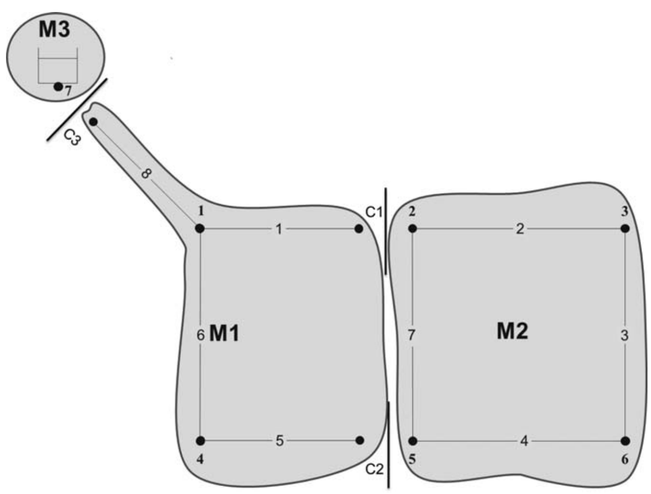

Figure 2. Modules of the network in Figure 1 accounting for cut positions. The pipes are assigned to module according to the position of the actual gap and $Q = 0.125$.

Giustolisi and Ridolfi [2014] introduced a different modularity evaluation aiming at embedding the information about the position of devices, i.e., in order to develop a cut position-sensitive modularity. Figure 2 shows the strategy of the cut position-sensitive index. The cut pipes are assigned to one of the adjacent module, instead of half for each module as in the case of the classic modularity index, based on the position of the device, i.e., pipes 1 and 5 are entirely assigned to module M_1 being cuts C1 and C2 (i.e., devices) close to nodes 2 and 5, respectively. Consequently, the modified modularity index accounting for actual position of cuts becomes $Q = 0.125$ instead of $Q = 0.1328$. Moving the cut C1 close to the node 1, the value of modularity becomes $Q = 0.094$. Therefore, the modularity index becomes cut position-sensitive.

The mathematical formulation of the cut position-sensitive modularity index is [*Giustolisi and Ridolfi*, 2014]

$$Q=1-\frac{n_c}{n_p}-\sum_{m=1}^{n_m}\left[\sum_{k=1}^{n_p}\frac{(\mathbf{u}_p)_k\delta(\mathrm{M}_m,\mathrm{M}_k)}{n_p}\right]^2 \tag{6}$$

where the summation inside the square brackets is now related to pipes and δ is the function to sum the pipes of the same module (i.e., $\delta = 1$ if $\mathrm{M}_m = \mathrm{M}_k$ and $\delta = 0$ otherwise). It is therefore sufficient to count pipes belonging to modules instead of using nodal degrees as in the classic modularity.

It is here useful to note that the external summation in equation (6) generally decreases with the number of components n_m. In fact, the following property holds

$$\sum_{m=1}^{n_m}\left[\sum_{k=1}^{n_p}\frac{(\mathbf{u}_p)_k\delta(\mathrm{M}_m,\mathrm{M}_k)}{n_p}\right]=1 \tag{7}$$

and, therefore, the summation of the square of n_m numbers (occurring in the equation (6)) – whose summation is unitary – tends to be lower with increasing of n_m. In particular, if a division generates identical modules, each term of the external sum in equation (7) would be equal to $1/n_m$ and, consequently, modularity index would be equal to

$$Q_{\max}=1-\frac{n_c}{n_p}-\frac{1}{n_m}\overbrace{=}^{n_c=n_m-1}1-\frac{n_m-1}{n_p}-\frac{1}{n_m} \tag{8}$$

The first of equations (8) states that the upper bound of the modularity index for a network composed of n_p pipes, given the number of "cuts" n_c and the number of modules n_m. The second of equations (8) specifies the upper bound of the modularity when n_m modules can be obtained with a minimum number of cuts ($n_c = n_m - 1$). The second of equations (8) will be useful in order to introduce the infrastructure modularity index in the next sections.

3. Modularity Maximization as a Particular Case of a More General Optimization

The modularity index – both the equivalent formulations in equations (1), (2), or (4) and the modified formulation in equation (6) which, differently, computes a_m in order to account for actual location of cuts on pipes – is the sum of two components, $Q = Q_1 + Q_2$, equal to

$$\begin{aligned} Q_1&=1-\frac{n_c}{n_p}\\ Q_2&=-\sum_{m=1}^{n_m}a_m^2=-\sum_{m=1}^{n_m}\left[\sum_{k=1}^{n_p}\frac{(\mathbf{u}_p)_k\delta(\mathrm{M}_m,\mathrm{M}_k)}{n_p}\right]^2 \end{aligned} \tag{9}$$

The first component, Q_1, strictly decreases with the number of cuts and penalizes the excess of cuts for a given number of modules. On the contrary, Q_2 generally is an increasing function of the number of modules (and generally of n_c) and it drives the search to the set of most similar modules to each other for a given number of cuts.

Q_1 and Q_2 are conflicting function of n_m, i.e., of n_c, and the maximization of their sum, the modularity index Q, can be interpreted as a single-objective strategy to balance those components.

However, the most general formulation of the mathematical problem is the following two-objective optimization,

10.1002/2014WR016067

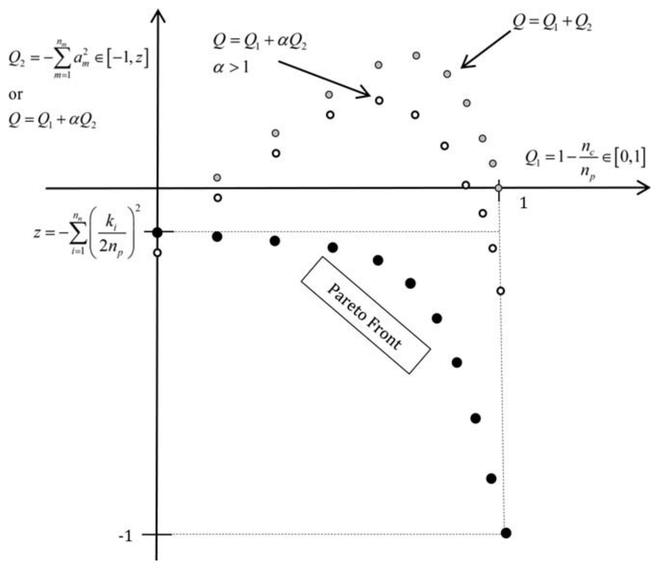

Figure 3. Two-objective optimization (Q_1 versus Q_2) and classic modularity index.

$$\begin{cases} [\mathrm{M}, n_c, n_m] = Connectivity\left(|\bar{\mathbf{A}}_{pn}|, I_c\right) \\ \underbrace{\max}_{I_c}(Q_1) = \underbrace{\max}_{I_c}\left(1-\frac{n_c}{n_n}\right) \\ \underbrace{\max}_{I_c}(Q_2) = \underbrace{\max}_{I_c}\left(-\sum_{m=1}^{n_m} a_m^2\right) \end{cases} \tag{10}$$

where $|\bar{\mathbf{A}}_{pn}|$ is the absolute value of the general topological matrix and I_c is the set of cuts in the network, i.e., the decision variable of the optimization problem. The topological representation of the network with gaps (see Figure 2) could be that reported as in *Giustolisi and Savic* [2010]

Consequently, the identification of the network modules by the maximization of the modularity index Q is a special case of a single-objective optimization,

$$\begin{cases} [\mathrm{M}, n_c, n_m] = Connectivity\left(|\bar{\mathbf{A}}_{pn}|, I_c\right) \\ \underbrace{\max}_{I_c}(Q_1+\alpha Q_2) = \underbrace{\max}_{I_c}\left(1-\frac{n_c}{n_n}-\alpha\sum_{m=1}^{n_m} a_m^2\right) \end{cases} \tag{11}$$

where Q_1 and Q_2 are summed by means of the weight α which is assumed unitary. Therefore, the maximization of the classic modularity index identifies a specific single solution (i.e., a set of modules) for the network segmentation corresponding to equal weights of Q_1 and Q_2 ($\alpha = 1$).

In order to better clarify the meaning of the classic modularity index with respect to the optimization standpoint, Figure 3 shows the general two-objective optimization in (10) providing the so-called Pareto front of solutions, which is the set of the best trade-offs of Q_1 versus Q_2. If $n_c = n_p$, i.e., each pipe is cut, $Q_1 = 0$ and, being each module composed of one node only, in the case of cuts in the middle of pipes $Q_2 = z$ (see Figure 3) is obtained. If the network is not segmented, $n_m = 1$, $Q_1 = 1$, $Q_2 = -1$, and $Q = 0$. The Pareto front is therefore bounded as follows

$$\begin{aligned} Q_1 &\in [0,1] \\ Q_2 &\in \left[-1, z=-\sum_{i=1}^{n_n}\left(\frac{k_i}{2n_p}\right)^2\right] \end{aligned} \tag{12}$$

where k_i stays for the degree of the i-th node. Figure 3 shows the pattern of the classic modularity index, where ordinate axis represents $Q = Q_1 + \alpha Q_2$, characterized by a maximum value and two bound values equal to $Q = z$ and $Q = 0$. Moreover, Figure 3 highlights that the maximization of the classic modularity index is a particular case of a single-objective optimization, which is a pragmatic solution to a two-objective optimization.

However, the pragmatic approach of the classic modularity is paid with the occurrence of a resolution limit in the module detection. In fact, the maximum of the curve ($Q_1 + Q_2$) is due to the reaching of Q_1 dominance with respect to Q_2, namely the magnitude of Q_1 is always larger than Q_2. Such mathematical dominance always occurs because the two components, Q_1 and Q_2, of the modularity index are conflicting, but it generates a resolution limit of the modularity index which depends on the number of pipes n_p [*Fortunato and Barthélemy*, 2007]. In other words, for a given network division in n_m modules, any identification of a new module is subject to the need that Q_2 is not absolutely dominated by Q_1 in order to have the possibility to increase their sum with a further optimal cut. Since in order to identify a new module is necessary at least one new cut in the network, the following inequality holds

$$\sum_{m=1}^{n_m} a_m^2 - \sum_{m=1}^{n_m+1} b_m^2 > \frac{1}{n_p} \quad (13)$$

where a_m and b_m are the divisions in n_m and n_m+1 modules, respectively, and the first summation is greater than the second because

$$\sum_{m=1}^{n_m} a_m = 1 \quad \text{and} \quad \sum_{m=1}^{n_m+1} b_m = 1 \quad (14)$$

Equation (13) clarifies that the resolution is proportional to the inverse of the square root of n_p.

In the single objective perspective, the resolution can be increased by introducing the weight $\alpha > 1$ in front of Q_2. In fact, as shown in Figure 3, the maximum value of $Q_1 + \alpha Q_2$ moves toward a lower Q_1 (i.e., a greater n_c) indicating the identification of a greater number of modules n_m. Accordingly, equation (13) becomes

$$\sum_{m=1}^{n_m} a_m^2 - \sum_{m=1}^{n_m+1} b_m^2 > \frac{1}{\alpha n_p}, \quad (15)$$

namely $\alpha > 1$ makes lower the resolution limit constraint decreasing the weight of Q_1 with respect to Q_2.

Even though the introduction of the coefficient α can reduce the resolution limit, in infrastructure networks this strategy is unpractical. In fact, the task is to support a decision on segmentation whose resolution depends on the technical purpose. Consequently, it would be necessary to select α depending on the specific aim of the segmentation and this sounds artificial and arbitrary. Moreover, even by changing the weight α, it would be impossible to identify all the possible module configurations if the trade-off curve (e.g., similar to that shown in Figure 3) has concave parts. It is instead more appealing to provide a unique decision support tool for segmentation as it will be described in the next section.

It is worth to note that one could solve the general problem in (10) and obtain the Pareto set of optimal solutions involving conflicting objectives. Nevertheless this approach exhibits some drawbacks: (i) increasing n_p, it provides a very large number of solutions – most of them useless – and causes an exponential growth of the computational burden even for not very large network sizes; (ii) for WDNs, the decision support tool for segmentation needs of a cost function which can be a nonlinear function of the number of cuts, see *Giustolisi and Ridolfi* [2014] for further details.

4. Infrastructure Modularity Index

4.1. Topology-Based Index

In the previous section, the constraint of equation (13) indicates that classic modularity index has a resolution limit which is a drawback for its application to infrastructure networks, when a specific technical task requires the identification of the small modules.

In order to remove such limit, we propose to modify the classic modularity index given by equation (8) by adding the term $(n_m - 1)/n_p$ representing the minimum fraction of cuts to obtain n_m modules as from equation (8). The proposed infrastructure-oriented modularity index is

$$IQ = Q + \frac{n_m - 1}{n_p} \Rightarrow IQ = 1 - \frac{n_c - (n_m - 1)}{n_p} - \sum_{m=1}^{n_m} \left[\sum_{k=1}^{n_p} \frac{(\mathbf{u}_p)_k \delta(\mathrm{M}_m, \mathrm{M}_k)}{n_p} \right]^2. \quad (16)$$

Without loss of generality with respect to equation (4), the modification is applied to equation (8) which is better suited for WDNs.

The maximum value of the infrastructure modularity index IQ results (from equation (8))

$$IQ_{\max} = 1 - \frac{1}{n_m} \quad (17)$$

and shows that it strictly depends on the number of modules. It is worth noting that the maximum value of IQ is asymptotically upper bounded to unit (for an infinite number of modules) as well as in the case of Q, while $IQ = 0$ for an unsegmented network.

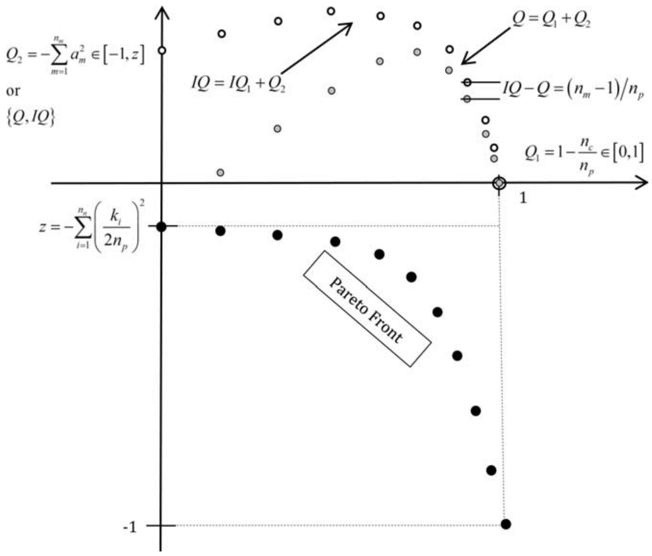

Figure 4. Two-objective optimization (Q_1 versus Q_2): classic and infrastructure modularity index.

Let us consider equations (9) and assume $IQ = IQ_1 + Q_2$, where

$$IQ_1 = 1 - \frac{n_c - (n_m - 1)}{n_p} \tag{18}$$

and the term Q_2 is the same of equation (8). Hence, the constraint in equation (13) becomes

$$\sum_{m=1}^{n_m} a_m^2 - \sum_{m=1}^{n_m+1} b_m^2 > 0 \tag{19}$$

Equation (19) demonstrates that the constraint with respect to the identification of a module by a single cut does not exist for IQ; in other words, IQ_1 never dominates Q_2. This fact significantly increases the resolution of the infrastructure modularity index with respect to any based on classic one, e.g., equation (8), without completely eliminating the dependence of the index on the number of cuts, in IQ_1, in order to obtain the modules.

Let us consider Figure 4, which is similar to Figure 3 but reports also the pattern of IQ. For sake of simplicity and without loss of generality, we assume that each solution (i.e., points) in Figure 4 corresponds to a different number of modules n_m, i.e., Q_1 can be interpreted as a monotone (decreasing) function of n_m.

Figure 4 shows that IQ coincides with $Q = 0$ for $n_m = 1$ (i.e., $n_c = 0$); in fact the additional term $(n_m - 1)/n_p$ (see equation (16)) becomes null, therefore the difference between IQ and Q linearly increases with n_m and decreases with Q_1. This fact shifts the maximum value of IQ toward lower values of Q_1 and higher values of the number of modules, i.e., the maximization of IQ finds more modules than that of Q. Moreover, the fact that the difference between IQ and Q depends only on the number of modules (n_p is in fact fixed once a specific network is chosen) guarantees that, fixed the number n_m lower than the maximum allowed by Q, the maximization of Q and IQ finds the same modules.

Finally, it is useful to note that the component IQ_1 of IQ results less conflicting with Q_2 because it increases with n_m similarly to Q_2 (remind that this latter is negative). Therefore, the infrastructure modularity index can be better used in the two-objective optimization max(IQ) versus min(*segmentation cost*) in order to support technical decisions for different purposes. This feature of IQ solves the drawbacks (points (i) and (ii)) reported at the end of the previous section, which are related to the optimization Q_1 versus Q_2, see problem in (10). In fact, the Pareto set is upper bounded by the maximum of IQ, without the need of exploring the whole domain of Q_1, point (i). Furthermore, IQ is conflicting with any cost objective function monotonically increasing with the number of cuts, because its dependence on n_c is smoothed in IQ_1 by the term $(n_m - 1)/n_p$. For this reason, the index IQ is much more appropriate than Q in order to build a decision support tool for segmentation, i.e., to solve a two-objective optimization problem, because IQ is more conflicting with a cost function based on the number of cuts with respect to Q which is linearly increasing with n_c. Nevertheless, IQ maintains the most relevant key features of Q, point (ii). In fact, as discussed above, maximizations of Q and IQ (once n_m is fixed) find the same modules.

Moreover, the cost-benefit, min(*cost*) versus max(IQ), optimization can start from any existing segmentation and can be a decision tool for any technical purpose of the network segmentation. In fact, the infrastructure modularity index is not influenced by an initial nonoptimal preexisting segmentation (biasing the problem) because IQ_1 has an offset depending on n_m. On the contrary, preexisting segmentations, especially nonoptimal, influence the actual resolution of the classic modularity index.

4.2. Pipe Weight-Based Index and Technical Constraints

In order to extend the formulation of the infrastructure modularity to the case of pipe weights, the equation (4) (see *Giustolisi and Ridolfi* [2014] and equation (16)) is rewritten as follows

$$IQ(\mathbf{w}_p)=1-\frac{n_c-(n_m-1)}{n_p}-\sum_{m=1}^{n_m}\left[\sum_{i=1}^{n_n}\frac{\left(|\bar{\mathbf{A}}_{pn}^T|\mathbf{w}_p\right)_i\delta(\mathrm{M}_m,\mathrm{M}_i)}{2W}\right]^2 \quad (20)$$

where W is the sum of pipe weights w_k ($k=1, \ldots, n_p$) stored in the vector $\mathbf{w}_p$. Accordingly, the cut position-sensitive formulation of the equation (20) is

$$IQ(\mathbf{w}_p)=1-\frac{n_c-(n_m-1)}{n_p}-\sum_{m=1}^{n_m}\left[\sum_{k=1}^{n_p}\frac{(\mathbf{w}_p)_k\delta(\mathrm{M}_m,\mathrm{M}_k)}{W}\right]^2 \quad (21)$$

where the summation inside the square brackets considers pipe weights. The case $\mathbf{w}_p=\mathbf{u}_p$ corresponds to the topology-based infrastructure modularity in equation (16).

Equation (19) holds for the infrastructure modularity index based on pipe weights because the term in the right hand of the equation (13) becomes null also in the generalized form of *IQ* of equation (21) being the added term $(n_m-1)/n_p$ independent on pipe weights. It follows that the (cut position-insensitive or sensitive) generalized forms of the infrastructure modularity index given in equations (20) and (21) are characterized by the same resolution features discussed in the previous section.

Finally, it is to remark that the term $(n_m-1)/n_p$ of *IQ* contains the number of modules and, therefore, the metric allows the separation of the most peripheral nodes of the branched parts of the networks without the reduction of the index. In order to exclude the identification of such a modules composed by a single node when *IQ* is used, it is possible to consider, inside the term IQ_1, for n_m the modules with at least one pipe. This remark allows to further extend the formulation of *IQ*

$$IQ=1-\frac{n_c-[n_m(w_m\geq w_{th})-1]}{n_p}-\sum_{m=1}^{n_m}\left[\sum_{k=1}^{n_p}\frac{(\mathbf{w}_p)_k\delta(\mathrm{M}_m,\mathrm{M}_k)}{W}\right]^2 \quad (22)$$

where $n_m(w_m\geq w_{th})$ is the number of modules having the weights w_m greater than the value w_{th} that is a threshold for the lowest technical size of a network segment.

5. Case Study

The infrastructure modularity index is here used to segment a real water distribution network as illustrative case study. The network is C-TOWN [*Ostfeld et al.*, 2012]; its layout is reported in Figure 5. The network is composed of 444 pipes and 396 nodes. The system is already divided in five district metering areas (DMAs) by proper flow devices placed downstream of the eleven pumps, labeled PM in Figure 5. Any segmentation of the C-TOWN network is then biased by such existing DMAs. This fact represents a common situation for infrastructure networks where segmentation has to match preexistent (nonoptimal) modules.

Furthermore, there are seven tanks and one reservoir, indicated in Figure 5 by the nodal rectangles and label H_0, respectively. Finally, there are also three pressure control valves, labeled PV in Figure 5. Assuming for the 11 pumps, eight sources of water and three valves one flow measurements, the 22 existing devices are observations acting as conceptual cuts already segmenting, in a nonoptimal way, the network in 9 modules.

We start using the network in Figure 5 without considering the 22 devices, in order to obtain the same type of diagram shown in the Figure 4 but for a real infrastructure network. Accordingly, Figure 6 reports the diagrams which refer to the solution of the optimization problem Q_1 versus Q_2 of the real network in Figure 5, as formulated in (10), solved using genetic algorithms [*Goldberg*, 1989; *Giustolisi and Ridolfi*, 2014]. Therefore, Q_1 spans the interval [0,1] corresponding to n_c ranging in [n_p=444,0], while Q_2 ranges from -1 to $z=-3.27{\cdot}10^{-3}$.

Similarly to Figure 4, on the top of the Figure 6a, the patterns of Q and IQ are reported as a function of Q_1. The diagram of Q exhibits a maximum value equal to 0.867 at Q_1=0.935 and corresponds to $n_c=n_m=29$.

 10.1002/2014WR016067

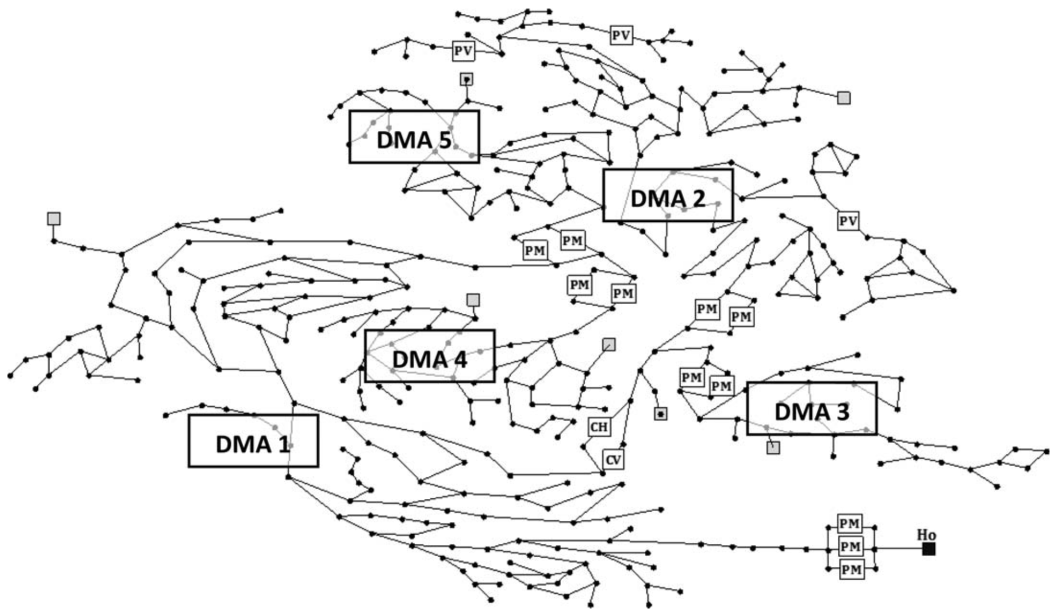

Figure 5. Layout of TOWN-C network. (PM = pump; CV = tank controlled valve; CH = check valve; PV = pressure reduction valve).

The diagram of *IQ* exhibits a maximum value equal to 0.959 at Q_1=0.836 and corresponds to $n_c = 73$ and $n_m = 68$. The similarity of the diagrams in Figures 4 and 6a confirms the theoretical discussion of the previous section and the fact that the infrastructure index has the maximum value corresponding to a greater number of modules, 29 versus 68, i.e., its resolution limit is much lower considering the medium/small size of the network.

However, as reported in Figure 6b zooming the diagrams close to maxima, it is evident the smooth behavior of *Q* while *IQ* is characterized by small fluctuations although its pattern indicates the existence of a clear maximum. The values of *Q* do not fluctuate because it is the sum of the two variables, Q_1 and Q_2, used for the optimization. The fluctuations of *IQ* are instead caused by the fact that the number modules does not correspond to an unique number of cuts (embedded in the abscissa Q_1) as assumed in the Figure 4. In fact, a given number of modules can be obtained by means of (little) different numbers of cuts and, being the difference ($IQ - Q$) equal to $(n_m - 1)/n_p$, this fact causes the small fluctuations in the pattern $IQ = IQ(Q_1)$.

Figure 7 shows the Pareto set of the best trade-off minimization of n_c versus maximization of *IQ*, now starting from the 22 cuts corresponding to the already existing flow devices in the network which generate nine preexisting modules. Since the optimization of *IQ* is performed, the fluctuations do not occur: for a given n_m the best solution (i.e., minimum n_c) is in fact obtained during the multiobjective optimization.

It is worth noting that the maximum of *Q* correspond to for $n_c = 38$ with $n_m = 24$, while the maximum of *IQ* corresponds to $n_c = 133$ with $n_m = 118$.

Two facts need to be emphasized: (i) the maximum of *IQ* corresponds to a number of cuts that is much smaller than $n_p = 444$, especially considering that the network is not particularly looped and large sized; and (ii) the *IQ* index is less dependent on n_c than *Q* and this is confirmed by the flat pattern of the diagram increasing the number of modules (see Figure 7). This fact is useful, as above reported, because *IQ* results much more conflicting with any cost function monotonically increasing with n_c and much more suitable for a multiobjective segmentation strategy, e.g., involving the cost of newly installed devices [*Giustolisi and Ridolfi*, 2014].

Finally, we use another real infrastructure network of a medium size hydraulic system, named Exnet [*Giustolisi et al*., 2008]. The Exnet network layout is reported in Figure 8.

The network is composed of 1894 nodes (reservoir comprised) and 2471 pipes. It is a powerful test in order to prove the resolution of proposed infrastructure index, because it is larger than Town-C and much more compact, i.e., more looped and generally requiring more than one cut to separate modules. Minimization of

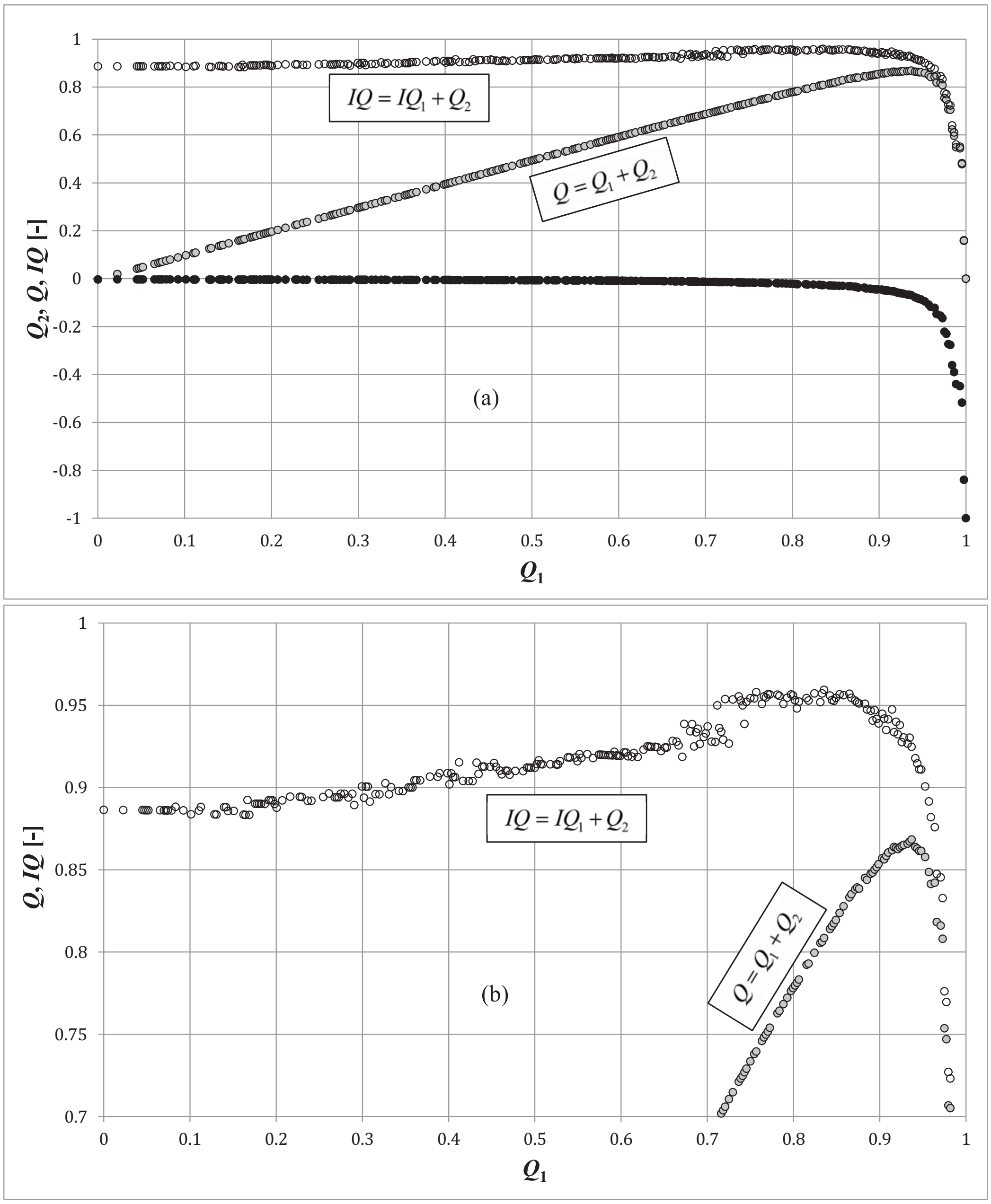

Figure 6. (a) Two-objective optimization, Q_1 versus Q_2 and patterns of the classic (Q) and infrastructure modularity (IQ) index. (b) Zooming of the Q and IQ patterns close to maxima.

n_c versus the maximization of IQ is obtained using genetic algorithms [*Goldberg*, 1989; *Giustolisi and Ridolfi*, 2014]. The infrastructure index is assumed topology-based while existing devices are not considered.

Figure 9 reports the Pareto set of the best trade-off n_c versus IQ and the same diagram assuming Q instead of IQ. It demonstrates that the infrastructure index IQ extends the resolution of the segmentation to $n_m = 465$ with $n_c = 517$ which is much higher than $n_m = 36$ with $n_c = 89$ achievable with the classic cut position-sensitive modularity index, Q.

6. Conclusions

Recently proposed mathematical tools of the complex network theory can play a relevant role in order to face analysis, planning and management tasks of infrastructure networks. This work focuses on the problem

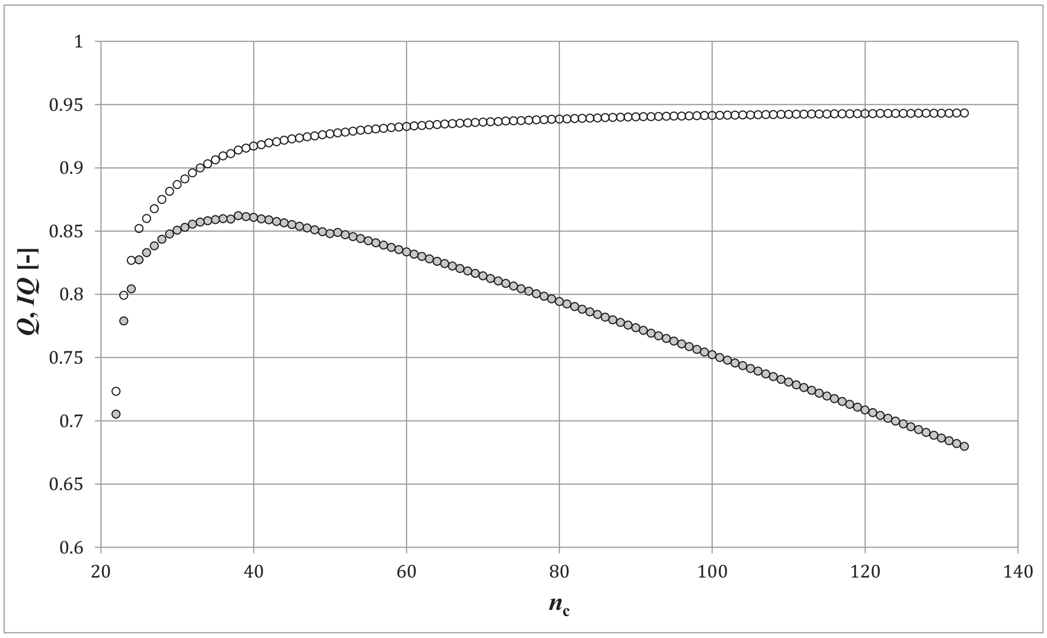

Figure 7. Two-objective optimization n_c (biases by existing devices) versus *IQ* (white points) and classic modularity index (gray points).

of the segmentation of water distribution networks. To this aim, the infrastructure modularity index has been proposed; it is a modification of the classic modularity index in order to overcome the resolution limit. This is a key point because several technical tasks require a decision support tool able to segment networks with a resolution more detailed than the resolution limit typical of the classical modularity index.

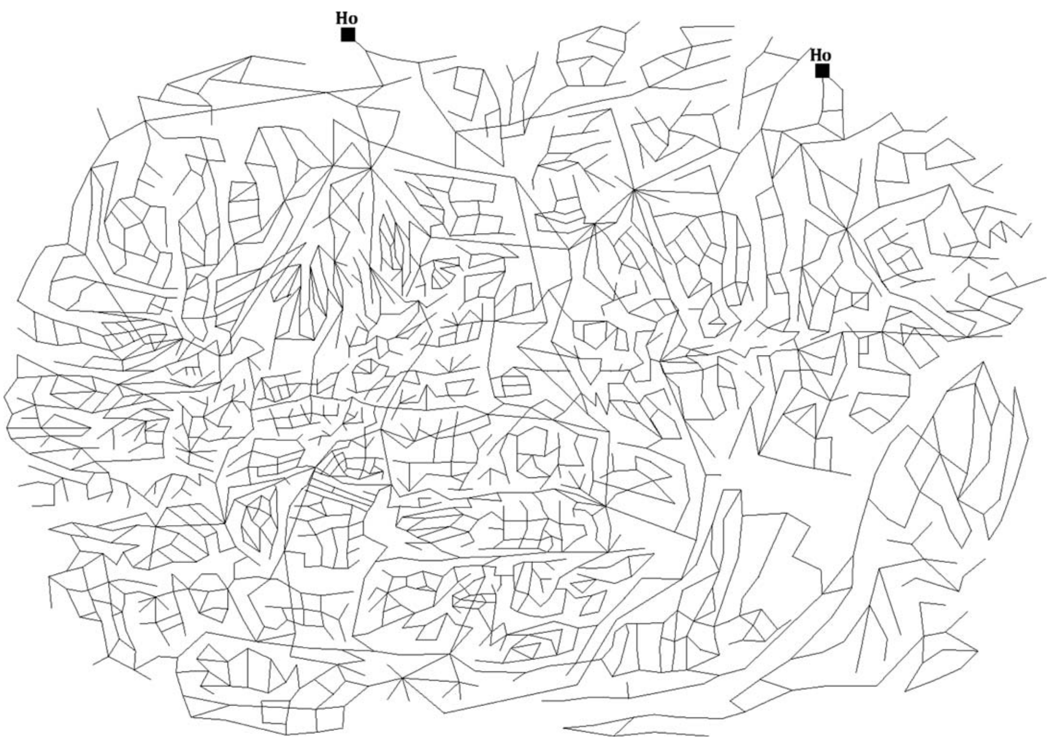

Figure 8. Layout of Exnet network.

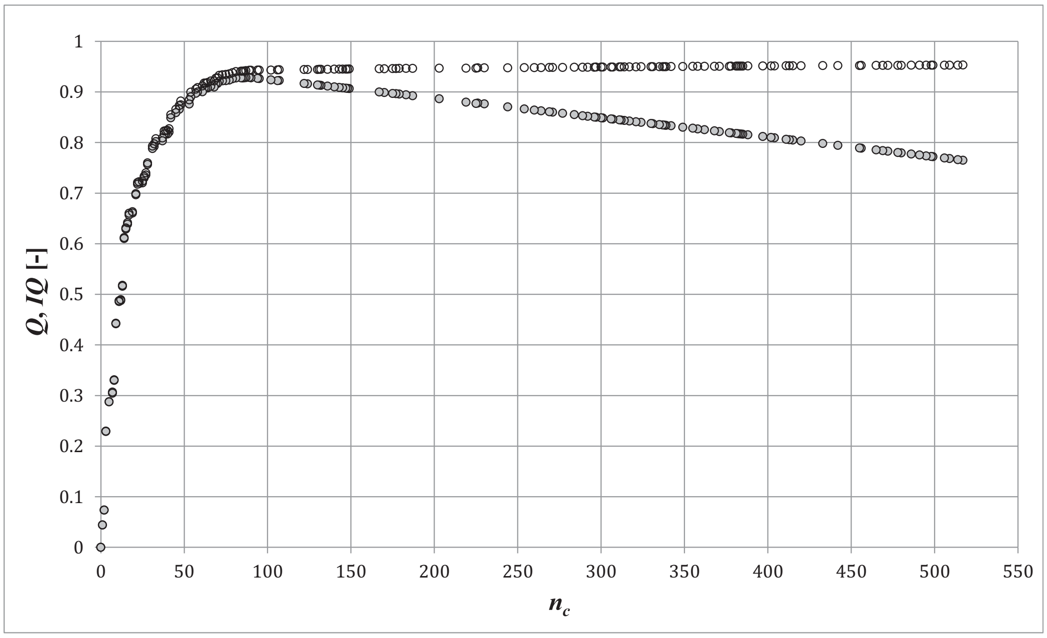

Figure 9. Two-objective optimization n_c versus *IQ* (white points) and classic modularity index (gray points).

The infrastructure modularity index maintains the nice features of the classic modularity index, but it is better suited for water distribution networks because the resolution limit: (i) does not exist with respect to the separation of a module using one cut; (ii) is not influenced by existing segmentations/devices biasing the optimization problem; and (iii) is less dependent on the number of cuts, which makes the metric more conflicting with a cost function (monotonically increasing with the number of cuts, i.e., the installed devices in the network) in the formulation of the optimization as multiobjective. Furthermore, constraints about the minimum segment size with respect to selected pipe weights can be easily introduced in the infrastructure modularity index. They can be implemented by means of a constrained counting of the number of modules/segments of the term $(n_m - 1)/n_p$.

Finally, segmentation of two real infrastructure networks, reported as case studies, shows and confirms the effectiveness of the proposed infrastructure modularity index.

Acknowledgments

All data used in this paper were obtained from papers cited in the reference list. The research reported in this paper was funded by the Italian Scientific Research Program of National Interest PRIN-2012 "Tools and Procedures for the Advanced and Sustainable Management of Water Distribution Systems."

References

Albert, R., and A. L. Barabasi (2002), Statistical mechanics of complex networks, *Rev. Mod. Phys.*, *74*, 47.

Albert, R., I. Albert, and G. L. Nakarado (2004), Structural vulnerability of the North American power grid, *Phys. Rev. E*, *69*, 025103.

Alvisi, S., and M. Franchini (2014), A heuristic procedure for the automatic creation of district metered areas in water distribution systems, *Urban Water J.*, *11*, 137–159.

Arianos, S., E. Bompard, A. Carbone, and F. Xue (2009), Power grid vulnerability: A complex network approach, *Chaos*, *19*, 013110.

Barrat, A., M. Barthélemy, and A. Vespignani (2008), *Dynamical Processes on Complex Networks*, Cambridge Univ. Press, Cambridge, U. K.

Barthélemy, M. (2011), Spatial networks, *Phys. Rep.*, *499*, 1–101.

Boccaletti, S., V. Latora, Y. Moreno, M. Chavez, and D. U. Hwang (2006), Complex networks: Structure and dynamics, *Phys. Rep.*, *424*, 175–308.

Bompard, E., E. Pons, and D. Wu (2012), Extended topological metrics for the analysis of power grid vulnerability, *IEEE Syst. J.*, *6*, 481–487.

Deuerlein, J. W. (2008), Decomposition model of a general water supply network graph, *J. Hydraul. Eng.*, *134*, 822–832.

Diao, K., Y. Zhou, and W. Rauch (2013), Automated creation of district metered area boundaries in water distribution systems, *J. Water Resour. Plann. Manage.*, *139*, 184–190.

Ferrari, G., D. A. Savic, and G. Becciu (2014), A graph theoretic approach and sound engineering principles for design of district metered areas, *J. Water Resour. Plann. Manage.*, doi:10.1061/(ASCE)WR.1943-5452.0000424, in press.

Fortunato, S. (2010), Community detection in graphs, *Phys. Rep.*, *486*, 75–174.

Fortunato, S., and M. Barthélemy (2007), Resolution limit in community detection, *Proc. Natl. Acad. Sci. U. S. A.*, *104*, 36–41.

Giustolisi, O., and L. Ridolfi (2014), A new modularity-based approach to segmentation of water distribution networks, *J. Hydraul. Eng.*, doi:10.1061/(ASCE)HY.1943-7900.0000916.

Giustolisi, O., and D. A. Savic (2010), Identification of segments and optimal isolation valve system design in water distribution networks, *Urban Water J.*, *7*, 1–15.

Giustolisi, O., D. A. Savic, and Z. Kapelan (2008), Pressure-driven demand and leakage simulation for water distribution networks, *J. Hydraul. Eng.*, *134*, 626–635.

Goldberg, D. E. (1989), *Genetic Algorithms in Search, Optimization and Machine Learning*, Addison-Wesley, Reading, Mass.
Havlin, S., et al. (2012), Challenges in network science: Applications to infrastructures, climate, social systems and economics, *Eur. Phys. J.*, *214*, 273–293.
Kurant, K., and P. Thiran (2006), Extraction and analysis of traffic and topologies of transportation networks, *Phys. Rev. E*, *74*, 036114.
Lin, J. Y., and Y. F. Ban (2013), Complex network topology of transportation systems, *Transp. Rev.*, *33*, 658–685.
Newman, M. E. J. (2010), *Networks: An Introduction*, Oxford Univ. Press, Oxford, U. K.
Newman, M. E. J., and M. Girvan (2004), Finding and evaluating community structure in networks, *Phys. Rev. E*, *69*, 026113.
Ostfeld, A., et al. (2012), Battle of the water calibration networks, *J. Water Resour. Plann. Manage.*, *138*, 523–532.
Perelman, L. and A. Ostfeld (2011), Topological clustering for water distribution systems analysis, *Environ. Modell. Software*, *26*, 969–972.
Scibetta, M., F. Boano, R. Revelli, and L. Ridolfi (2013), Community detection as a tool for complex pipe network clustering, *EPL*, *103*, 48001.
Sienkiewicz, J., and J. A. Holyst (2005), Statistical analysis of 22 public transport networks in Poland, *Phys Rev. E*, *72*, 046127.
von Ferber, C., T. Holovatch, Y. Holovatch, and V. Palchykov (2009), Public transport networks: Empirical analysis and modeling, *Eur. Phys. J. B*, *68*, 261–275.
Yazdani, A. and P. Jeffrey (2011), Complex network analysis of water distribution systems, *Chaos*, *21*, 016111.
Yazdani, A., and P. Jeffrey (2012), Water distribution system vulnerability analysis using weighted and directed network models, *Water Resour. Res.*, *48*, W06517, doi:10.1029/2012WR011897.
Zecchin, A. C., P. Thum, A. R. Simpson, and C. Tischendorf (2012), Steady-state behavior of large water distribution systems: Algebraic multigrid method for the fast solution of the linear step, *J. Water Resour. Plann. Manage.*, *138*(6), 639–650.