International Conference on Hydroinformatics

2014

Modularity Index For The Segmentation Of Water Distribution Networks

Orazio Giustolisi

Luca Ridolfi

11th International Conference on Hydroinformatics
HIC 2014, New York City, USA

MODULARITY INDEX FOR THE SEGMENTATION OF WATER DISTRIBUTION NETWORKS

ORAZIO GIUSTOLISI (1), LUCA RIDOLFI (2)
(1): Department of Civil Engineering and Architecture, Technical University of Bari, via E. Orabona 4, Bari 70125, Italy
(2): Department Environmental, Land, and Infrastructure Engineering, Politecnico di Torino, Corso Duca degli Abruzzi 24, Torino 10129, Italy.

The search for suitable segmentations is a challenging issue for analysis, planning and management of 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 complex network theory, modularity index was proposed as a measure of the strength of the network division into communities. Nevertheless, modularity index needs to be revised considering the specificity of the hydraulic systems. Accordingly, the classic modularity index is firstly presented and, then, tailored and modified for WDNs. Furthermore, a multi-objective strategy for optimal segmentation is presented. The optimization framework is based on the maximization of the WDN-oriented modularity-based index versus the minimization of the cost of newly installed devices in order to segment WDNs.

INTRODUCTION

The search for suitable segmentations is a challenging and important issue [2, 8, 9] and it is related to a number of important technical tasks as for example: to perform system analysis and model calibration, to plan efficient metering systems, to design pressure-control zones, etc. Giustolisi and Ridolfi [4] recently proposed the use of the modularity index, from complex network theory, as metric of WDNs segmentation in a multi-objective framework.

The network topology is the most important driver of the segmentation problem, although also the hydraulic and asset characteristics can play a relevant role. In fact, network segmentation needs to account for specific technical purposes asking for segmenting considering specific asset features. Furthermore, the segmentation process needs to account for the constrains on budget limit/uncertainty, pre-existing segmentations/devices and the objective of capital/operational cost minimization. The last point is relevant because the optimal segmentation for WDNs needs to be performed at least versus the cost minimization of the newly installed devices, therefore, it is biased by the pre-existing devices [4].

In complex network theory, modularity index is used to detect communities [3, 7], i.e. separated or overlapped group of nodes that privilege internal exchanges with respect to other groups. Modularity index can be interpreted from a functional standpoint: it is a specific tradeoff between the number of edges bridging clusters and a measure accounting for the number of modules and the similarity among each other. The clustering method is based on maximizing the modularity index and higher values of that metric mean a more evident community structure of the network. However, Barthlemy [1] pointed out that a carefulness when applying the modularity is advisable. In fact, the original formulation of that metric in WDNs is not worthwhile because they are infrastructure networks [1].

Therefore, the classic formulation of the modularity index needs to be tailored for WDNs. Furthermore, being the segmentation of WDNs performed for different technical tasks, its formulation needs to be extended in order to use asset features as further driver of the segmentation [4]. In addition, the modularity index needs to be modified in order to obtain a cut position-sensitive metric. In fact, the devices segmenting WDNs are usually installed close to the end nodes of pipes, while the classic modularity assumes that they are in the middle of pipes [4]. Finally, infrastructure networks require the development of a decision support tool, namely to solve an optimization problem exploring the best tradeoffs between segmentation cost of the newly devices versus a metric of the benefit of the division into modules [4].

To this purposes, the WDN-oriented modularity is used in a multi-objective (MO) optimization. The optimization is faced with MO genetic algorithm (MOGA) [4].

MODULARITY INDEX FOR WDNS

Topological representation of WDNs is a graph whose edges and vertices are the pipes and nodes of the hydraulic system, respectively. The number of pipes is n_p, the number of nodes is n_n (nodes of tanks/reservoirs are comprised), and the nodal degree, k, is the number of pipes/edges incident the node/vertex. A simple network with n_p=8 and n_n=7 is reported in Figure 1 and will be used as an example.

The network graph can be described by the general topological incidence matrix $\bar{\mathbf{A}}_{pn}$ which is commonly used to represent hydraulic systems. 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, remarking that closed pipes are not represented in the matrix $\bar{\mathbf{A}}_{pn}$. They are the i-th and j-th nodes linked by that pipe. The values of the two non-null elements in each row of $\bar{\mathbf{A}}_{pn}$ are {1,–1} and depend on the conventional positive direction assumed for pipe flow. The summation by columns of absolute value of the general topological incidence matrix, $|\bar{\mathbf{A}}_{pn}|$, allows to obtain the number of pipes incident the nodes (i.e., the nodal degrees) corresponding to each of the n_n columns of the matrix $\bar{\mathbf{A}}_{pn}$. The adjacency matrix, $\bar{\mathbf{A}}_{nn}$, of the network graph related to WDNs is the Boolean version of the matrix product $\bar{\mathbf{A}}_{pn}^{T} \times \bar{\mathbf{A}}_{pn}$ (superscript ^{T} indicates the transpose).

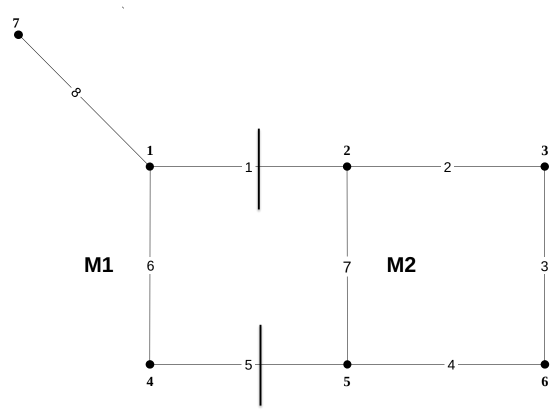

Figure 1. Example network with two modules (M1 and M2) generated by cuts on pipes 1 and 5.

The modularity, Q, is a measure of the strength of the network or graph division into communities/modules, we will use the word "module". Modularity is defined as [7]

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

where A_{ij} are the elements of the adjacency matrix $\bar{\mathbf{A}}_{nn}$, P_{ij} is the expected fraction of pipes between nodes i and j in the null/random network (i.e. the expected number of pipes in the network if they are 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 pairs (i,j), with $i \neq j$. In Eq. (1), the expected fraction P_{ij} is computed using nodal degree: i.e. k_i (k_j) is the degree of the i-th (j-th) node (i.e., the number of pipes incident that node). The modularity can be formulated also as

$$Q = \sum_{m=1}^{n_m} e_{mm} - \sum_{m=1}^{n_m} a_m^2 \quad \text{where} \quad \sum_{m=1}^{n_m} e_{mm} = 1 - \frac{n_c}{n_p}; \quad a_m = \sum_i \frac{k_i}{2n_p} \quad \forall i \in \mathrm{M}_m \tag{2}$$

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, a_m is the fraction of pipes having at least one end node in the module m and n_c is the number of pipes separating modules, namely the number of "cuts" in the network. The pipes dividing modules are counted ½ when computing a_m. In fact a_m is ½ of the summation of nodal degrees of the module m, divided by n_p. The n_c "cuts" are virtual and relate to the division into blocks (modules) of the network graph. Eq. (2) and the matrix $\bar{\mathbf{A}}_{pn}$ yield [4]

$$Q = 1 - \frac{n_c}{n_p} - \sum_{m=1}^{n_m}\left[\sum_{i=1}^{n_n} \frac{\left(\left|\bar{\mathbf{A}}_{pn}^T\right|\mathbf{u}_p\right)_i \delta\left(\mathrm{M}_m, \mathrm{M}_i\right)}{2n_p}\right]^2 \tag{3}$$

where M_i indicates the module where i-th node falls and $\mathbf{u}_p$ is the unitary n_p-size vector (symbol $(\cdot)_i$ indicates the i-th component of the vector inside the round brackets). The third term in Eq. (3) counts twice the number of pipes having both the nodes belonging to the m-th module and half of the number of pipes separating it from the other modules, consistently with the definition of a_m. If $n_c = 0$, the network is composed of one module ($n_m = 1$) and $a_m = 1$, being $2n_p$ the number of non-zero elements in $\bar{\mathbf{A}}_{pn}$; consequently, the modularity index Q is null. On the other side, Q is asymptotically upper bounded to unit as reported in the appendix section.

The formulation in Eq. (3) tailors the classic modularity index for WDNs using the topological incidence matrix and the number of "cuts". The use of the incidence matrix is important because it stores the information about parallel pipes which are common in infrastructure networks. On the contrary, the modularity index formulated in Eq. (1), i.e. by means of the adjacency matrix, does not store that information unless specific weights are used. In order to consider other network characteristics, the modularity can be also written [6]

$$Q = \frac{1}{2W}\sum_{ij}\left(W_{ij} - \frac{k_i^w k_j^w}{2W}\right)\delta\left(\mathrm{M}_i, \mathrm{M}_j\right) \tag{4}$$

where W is the sum of pipe weights w_k (k=1, …, n_p), W_{ij} are the elements of the weight matrix (i.e., $W_{ij} = W_{ji} = w_k$, where the k-th pipe connects nodes i and j), and k_i^w (k_j^w) is the degree based on pipe weights of the i-th (j-th) node. Then, Eq. (4) for WDNs becomes [4]

$$Q = 1 - \frac{\sum_{k=1}^{n_c} w_k}{W} - \sum_{m=1}^{n_m} \left[\sum_{i=1}^{n_n} \frac{\left(\left| \bar{\mathbf{A}}_{pn}^T \right| \mathbf{w}_p \right)_i \delta(\mathrm{M}_m, \mathrm{M}_i)}{2W} \right]^2 \tag{5}$$

The matrix product allows pipe weights in $\mathbf{w}_p$ to be concentrated in the nodes using the coefficient ½. The summation over k=1:n_c accounts for the pipe weights corresponding to cuts. Weight-based definition of Eq. (5) is much more versatile than the formulation of Eq. (3) [4].

However, the maximization of modularity index in Eq. (3) implies the minimization of the number of cuts in order to obtain the highest number of modules which are similar to each other, while the maximization of weight-based modularity in Eq. (5) implies the minimization of the sum of the pipe weights where cuts occur in order to obtain the greatest number of modules which are similar to each other with respect to the selected w_k. As we are interested in the number of cuts because they will be mapped in a total cost of newly installed devices, the following modification of Eq. (5), substituting the number of cuts to the pipe weights [4],

$$Q = 1 - \frac{n_c}{n_p} - \sum_{m=1}^{n_m} \left[\sum_{i=1}^{n_n} \frac{\left(\left| \bar{\mathbf{A}}_{pn}^T \right| \mathbf{w}_p \right)_i \delta(\mathrm{M}_m, \mathrm{M}_i)}{2W} \right]^2 \tag{6}$$

is better suited as modularity-based index when it is used for infrastructure networks.

TOPOLOGICAL REPRESENTATION OF MODULES FOR WDNs

In any modularity-based index, virtual pipe cuts are implicitly assumed in the middle of pipes, as reported in Figure 1. However, this assumption can be misleading for WDNs being significant the position of devices which actually segment the infrastructure networks. In fact, the virtual cuts become devices installed in networks being divided into modules.

The division can be (i) real, by means of the isolation valve system, in order to be able to detach the segments, (ii) conceptual, by means of flow observations, or (iii) mixed. In any case, those devices are installed close to the end nodes of pipes. The fact that a device is installed close to one or the other node of a pipe corresponds to two different technical solutions which should have a different modularity value. Therefore, the modularity index of Eq. (6) needs to be modified in order to become sensitive to the actual cut position. In other words, the value of the modularity should depend on the position devices close to the end nodes of pipes.

Figure 2 shows the same network reported in Figure 1, now with gaps corresponding to three devices, of any type, installed close to nodes. A new cut is assumed on pipe 8; it produces the single-node module M_3 corresponding to the isolation of a tank/reservoir node. The modularity by the classical cut position-insensitive formulation is Q = 0.1328. In order to embed the information about the position of valve/flow devices in the modularity the topological representation of networks divided into modules used in *Giustolisi and Savic* [5] is here adopted. Figure 3 reports the division into modules accounting for the cut position and assigning the cut pipes to one of the adjacent module. For example, pipes 1, 5 and 8 belong to module M_1 and, therefore, the corresponding pipe weights are assigned totally to module M_1. Consequently, the modularity index for the network becomes $Q = 0.125$. Assuming the cuts C1, C2, and C3 close to the nodes 1, 4, and 1, respectively, the position-sensitive modularity is equal to $Q = 0.03125$. Therefore, we obtain two different values which are both dissimilar from that of the classical cut position-insensitive modularity (i.e., $Q = 0.1328$).

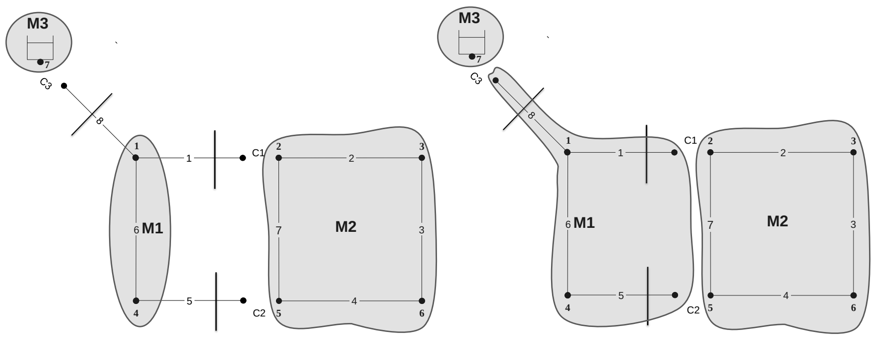

Figure 2. Modules of the network in Figure 1 segmented by three devices (left) classic modularity (right) modularity accounting for cut positions.

In general, in order to account for the position of valve/flow devices, it is sufficient to count pipes belonging to modules as separated in the network with gaps instead of using nodal degrees. Therefore, Eq. (7) is modified as follows [4]:

$$Q = 1 - \frac{n_c}{n_p} - \sum_{m=1}^{n_m} \left[\sum_{k=1}^{n_p} \frac{(\mathbf{w}_p)_k \, \delta(\mathrm{M}_m, \mathrm{M}_k)}{W} \right]^2 \quad (7)$$

where the summation inside the square brackets now considers pipes and δ is the function to sum the pipe weights of the same module (i.e. $\delta = 1$ if $\mathrm{M}_m = \mathrm{M}_k$ and $\delta = 0$ otherwise). The case $\mathbf{w}_p = \mathbf{u}_p$ corresponds to the topology-based metric.

SUPPORTING SEGMENTATION USING MULTI-OBJECTIVE OPTIMIZATION

In order to segment WDNs, a MO optimization is proposed [4],

$$\begin{cases} [\mathrm{M}, n_m, n_c] = Connectivity\left(\left| \bar{\mathbf{A}}_{pn-mod} \right|, I_c \right) \\ \max\limits_{I_c}(Q) = \max\limits_{I_c} \left(1 - \frac{n_c}{n_p} - \sum_{m=1}^{n_m} \left[\sum_{k=1}^{n_p} \frac{(\mathbf{w}_p)_k \, \delta(\mathrm{M}_m, \mathrm{M}_k)}{W} \right]^2 \right); \quad \min\limits_{I_c}(Cost) = \min\limits_{I_c} \left(\sum_{j=1}^{n_d} C_j \right) \end{cases} \quad (8)$$

where $\bar{\mathbf{A}}_{np}$ is the incidence matrix, I_c is the set of n_c cuts in the network (i.e., the decision variable of the optimization problem), *connectivity*($|\bar{\mathbf{A}}_{np}|$, I_c) stands for component analysis of the graph for the given cuts and C_j are the costs of the n_d new devices installed in the WDN in order to segment. In problem (8) the metric used is the cut position-sensitive modularity defined in Eq. (7). We remark the important fact that $n_d \leq n_c$, i.e. the pre-existing devices (selected in the optimization procedure among those candidate by a prior engineering judgment in order to be eventually used for segmentation) are not considered in the cost function. On the contrary, they are considered in the metric measuring the strength/efficiency of the segmentation.

The optimization problem in (8) provides a Pareto set of solutions which are optimal tradeoffs between the cost of newly installed devices versus the best segmentation whose metric is the modularity. These solutions are a support to the technical decisions considering budget uncertainty and a dynamical planning, from a coarser to a finer segmentation, possibly driven by the increase of the system knowledge [4].

CASE STUDY AND MO STRATEGY USING GENETIC ALGORITHMS

WDNs are not large size networks (at least if compared with other typical complex networks) and the division into modules is an activity which needs to be performed few times and, generally, off-line in the whole life of the hydraulic system. For this reason, MOGA is an efficient and flexible optimization strategy in order to achieve a good sub-optimal Pareto set of solutions. In fact, MOGA is a global exploration of the search space which is particularly effective for combinatorial problems and it is very flexible for mapping the cuts of the modularity metric through a cost function, possibly adding a further system analysis.

The Apulian network [5] is here used, because the aim is to show the MO strategy by means of a simple network and the length-based segmentation is shown without losing generality. Then, the problem in (8) using the modularity of Eq. (7) using w_k = k-th pipe length. The MOGA optimization starts from the pre-existing devices, in the case study the assumption of a flow and level measurements downstream to the network reservoir as shown in Figure 3.

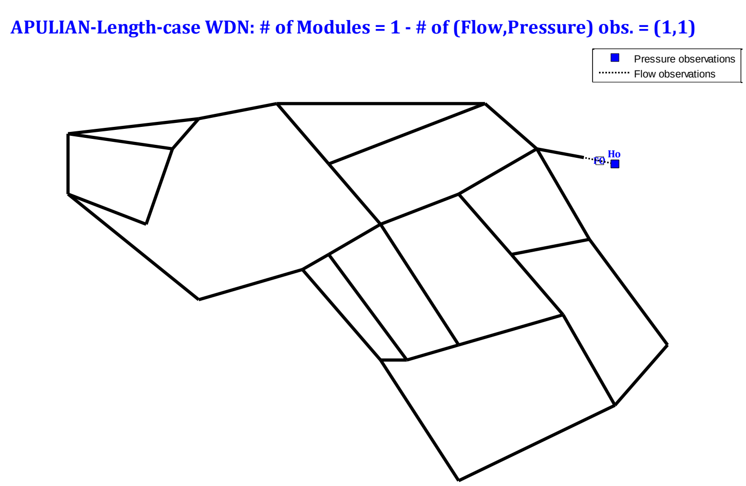

Figure 3. Apulian network as segmented by the pre-existing devices close to the reservoir.

Also the pressure observations in the nodes close to the devices concur to the cost function; this in order to achieve the whole set of pressure/flow observations allowing the hydraulic determination of the boundary conditions for each module [4]. To the purpose of the case study, the cost function is surrogated by the number of newly installed devices (comprised the pressure observations). Clearly, those measurement devices segment the network separating the reservoir.

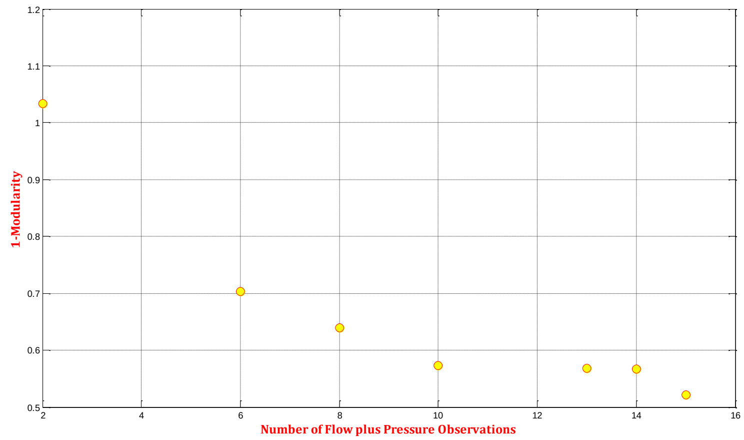

Figure 4. Pareto set of optimal solution (y-axis is 1 – *Q*).

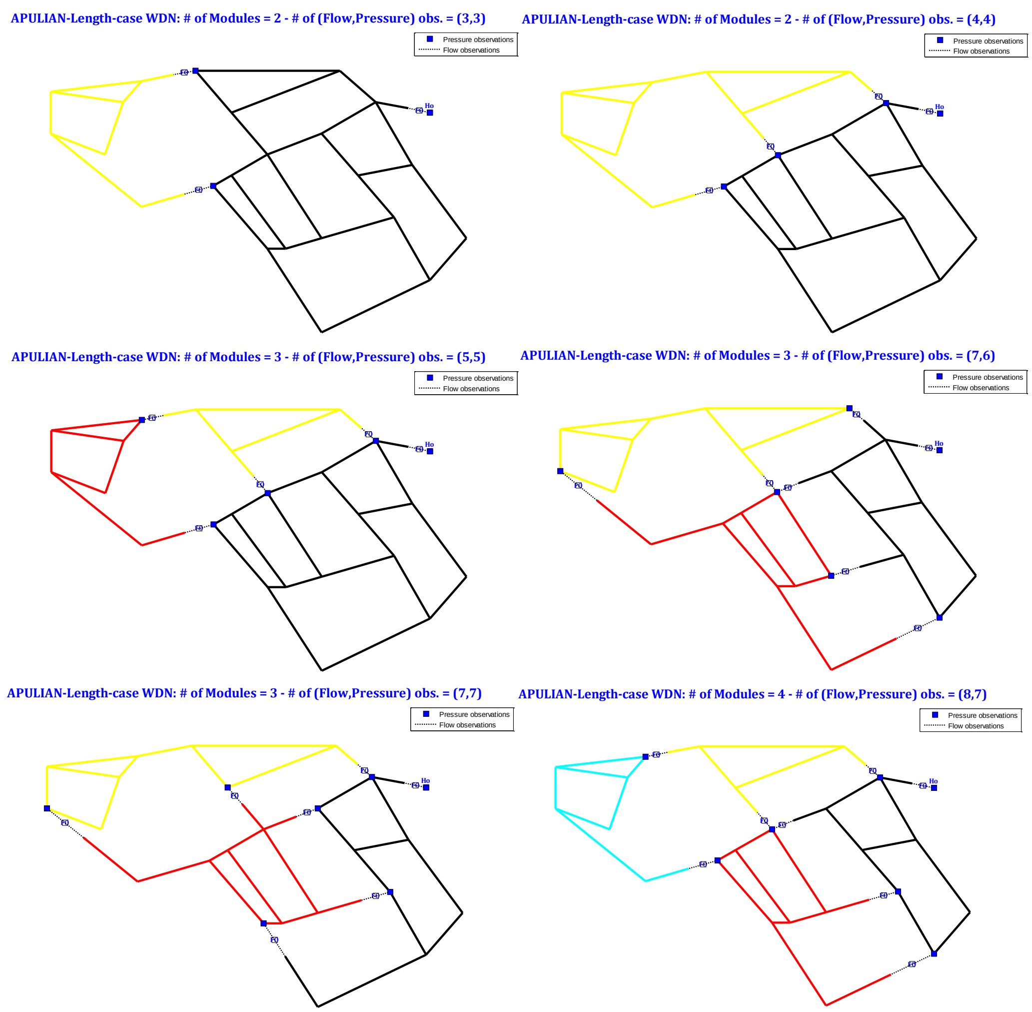

Figure 5. Optimal segmentations considering pre-existing devices close to the reservoir.

Figure 4 reports the Pareto set of optimal solution (trade-offs number between of new devices to be installed and modularity maximization) and Figure 5 shows the diagrams corresponding to the six solutions not considering the first, Figure 4, biasing the optimization.

The Pareto set in Figure 4 and the segmentations from coarser to finer reported in Figure 5 allows to support the decision on the segmentation. In fact, the six solution can used to evaluate the level of segmentation of the network versus the budget (i.e. the cost of new devices to be installed which is here surrogated as number of devices) also accounting for budget uncertainty planning a strategy from a coarser to a finer segmentation.

It is noteworthy that the technical purpose of the segmentation is relevant for the planning strategy and the weight used should be consistent with it. However, it is possible to argue that [4] to argue that the metrics with different weights are not conflicting for coarser segmentations because the main driver of the division in modules is the network topology. While stronger differences could emerge for finer segmentations using different weights corresponding to different technical needs of the segmentation [4]. Therefore, the problem in (8) extended to several metrics could be used (i) to explore the differences among metrics for a specific WDN and (ii) to achieve a basic coarse segmentation which is an optimal solution common to all the metrics being a mainly topology-based. Such coarse segmentation can be refined over time with a finer one which is a common optimal or sub-optimal solution considering all the metrics in an integrated planning perspective.

CONCLUDING REMARKS

The modularity index is here tailored in order to be WDN-oriented. Then it is proposed multi-objective strategy for the segmentation of WDNs. It allows to obtain a Pareto set of solutions which is a decision support tool for planning activities considering the specific technical task of the segmentation. Furthermore, the strategy allows considering an integrated and dynamical framework during planning activities when multiple WDN-oriented metrics are used.

REFERENCES

[1] Barthlemy, M.. “Spatial networks”. *Phys. Rep.*, Vol. 499, No. 13, (2011), pp 1-101.

[2] Diao, K., Zhou, Y., Rauch, W. “Automated Creation of District Metered Area Boundaries in Water Distribution Systems.” *J. Water Res. Plan. and Manage.*, Vol. 139, No. 2, (2013), pp 184–190.

[3] Fortunato, S. “Community detection in graphs.” *Phys. Rep.*, 486(3-5), (2010) 75{174}.

[4] Giustolisi, O, Ridolfi, L. “A new modularity-based approach to segmentation of water distribution networks.” *J. Hydr. Eng.*, (2014) accepted for publication.

[5] Giustolisi, O, Savic, D.A. “Identification of segments and optimal isolation valve system design in water distribution networks.” *Urban Water J.*, Vol. 7, No. 1, (2010), pp 1–15.

[6] Newman, M.E.J. “Analysis of weighted networks.” *Phys. Rev.* E, 70, 05613, (2004).

[7] Newman M.E.J., Girvan M. “Finding and evaluating community structure in networks.” *Phys. Rev.* E, 69, 026113, (2004).

[8] Perelman, L., Ostfeld, A. “Topological clustering for water distribution systems analysis.” *Environ. Modell. Software*, Vol. 26, No. 7, (2011), pp 969–972.

[9] Scibetta M., Boano F., Revelli R., Ridolfi L. “Community detection as a tool for complex pipe network clustering.” EPL, 103, 48001, (2013).