Community detection as a tool for DMA identification

Available online at www.sciencedirect.com
SciVerse ScienceDirect
ELSEVIER
Procedia Engineering 00 (2013) 000–000
Procedia Engineering
www.elsevier.com/locate/procedia

12th International Conference on Computing and Control for the Water Industry, CCWI2013

Community detection as a tool for district metered areas identification

M. Scibetta[a,*], F. Boano[a], R. Revelli[a], L. Ridolfi[a]

[a]*DIATI - Department of Environment, Land and Infrastructure Engineering, Politecnico di Torino, Corso Duca degli Abruzzi 24, Turin 10121, Italy*

Abstract

Water losses, the portion of water introduced in a pipe network but not consumed by users, represent a significant problem in water distribution system (WDS) management. Modern guidelines suggest to divide the pipe network in clusters, in order to compute a water balance and measure water consumption by each group. These clusters are called district metered areas (DMAs). The division of a pipe network in DMAs is usually realized with a visual exam supported by technical experience. This approach, which is convenient for small WDSs, becomes difficult to apply to large WDSs characterized by thousands of user nodes and pipes. Therefore, it is necessary to have an automatic tool to recognize the affinity degree of neighbouring nodes and to decide how to assign a node to a particular DMA. We propose an automated approach to subdivide pipes, that only requires flow rates through the network. The method has been tested to a large WDS often used as benchmark. The approach successfully divides the pipe network in an acceptable number of DMAs. Each resulting DMA is characterized by a low number of external links and by a proper number of users.

Keywords: district metered area, water distribution system, water loss.

1. Introduction

Modern management guidelines suggest to divide the water distribution network in districts (IWA 2007a,b). Each district can be indipendently analyzed using flow meters and pressure trasducers to improve system knowledge, to realize the water balance of the area, and to locate water losses. Such equipped district is also called district metered area (DMA). Thus, a DMA is a set of nodes enclosed in specified and permanent boundaries defined in order to install the smallest number of metering devices. DMAs identification is commonly realized using an empiric approach based on technical experience (IWA 2007a) even if new methods, based on periodic acoustic surveys (Hunaidi 2012), graph theory tools (Tzatchkov and Alcocer-Yamanaka V. H. Ortiz 2006), and multi-agent systems (Herrera et al. 2012) have recently been proposed. The cited methods have a common limit: they become computationally expensive when applied to large networks.

* Corresponding author. Tel.: +39-011-090-5684 ; fax: +39-011-090-5698
E-mail address: marco.scibetta@polito.it

1877-7058

Nomenclature

Q	modularity
m	number of network edges
A_{ij}	element of the adjacency matrix
P_{ij}	expected number of edges between nodes i and j in the network null model
$\delta(a,b)$	equal to 1 if $a=b$, 0 otherwise
k_i	number of edges incident to node i
W_{ij}	element of the weights matrix
W	sum of the weights
s_i^{out}	outcome strength of node i
s_j^{in}	income strength of node j
q_{ij}	flow from node i to node j

In contrast, here we describe a procedure for the identification of DMA boundaries which is effective and fast for WDSs of any size. Since WDSs are networks in which user connections, tanks and reservoirs are nodes, and pipes, valves and pumps are links, our method draws inspiration from the community detection problem (Fortunato 2010), a key topic of the complex network theory. Complex network theory has grown exponentially in recent years (Watts and Strogatz 1998, Barabasi and Albert 1999, Albert et al. 2000, Strogatz 2001, Boccaletti et al. 2006). Its impressive growth is due principally to two aspects: the availability of data regarding to large networks, like Internet (Faloutsos et al. 1999), the World Wide Web (Albert et al. 1999), metabolic networks (Jeong et al. 2000) and citation networks (Redner 1998), and the development of tools to measure specific network characteristics (Albert and Barabasi 2002). Community detection aims to identify clusters of nodes in a generic network. A cluster is a set of nodes, also called community, in which nodes are better connected with intracluster nodes (nodes of the same community) than with extracluster nodes (nodes of another community).

In the following, after a description of the method proposed, an application to a large network (Ostfeld et al. 2008) will be discussed.

2. Method

Community detection methods aim to attribute each node to a distinct cluster to achieve a configuration in which intracluster nodes are well connected while nodes between two different clusters are not. In the last years many methods have emerged (see Fortunato 2010) like graph partitioning (Kernighan and Lin 1970), partitional clustering (Hlaoui and Wang 2004), hierarchical clustering (Hastie et al. 2009), and spectral methods (Donath and Hoffman 1973). Modularity Q, a measure of clustering goodness introduced by Newman and Girvan 2004, evaluates the fraction of edges in the network that connects vertices of the same type (i.e., within-community edges) compared to the expected value of the same quantity in a network with the same community divisions but with random connections between the vertices (null model). Modularity is defined as

$$Q = \frac{1}{2m}\sum_{ij}\left(A_{ij} - P_{ij}\right)\delta\left(C_i, C_j\right), \tag{1}$$

where A_{ij} are the elements of the adjacency matrix, m the number of edges, P_{ij} represents the expected number of edges between nodes i and j in the null model, and $\delta\left(C_i, C_j\right)$ is 1 if nodes are in the same community $\left(C_i = C_j\right)$ and 0 otherwise. P_{ij} is related to the joint probability that the same edge is connected to both i and j, i.e. $P_{ij} = k_i k_j/2m$, where k_i and k_j are the numbers of edges incident with nodes i and j, respectively. Substituting P_{ij} in Eq. (1), the latter becomes

$$Q = \frac{1}{2m}\sum_{ij}\left(A_{ij} - \frac{k_i k_j}{2m}\right)\delta\left(C_i, C_j\right), \tag{2}$$

with $Q \in [0, 1]$. A high value of modularity implies a good partition, while a low value entails partition of poor quality, with $Q = 0$ if all nodes are in the same group.

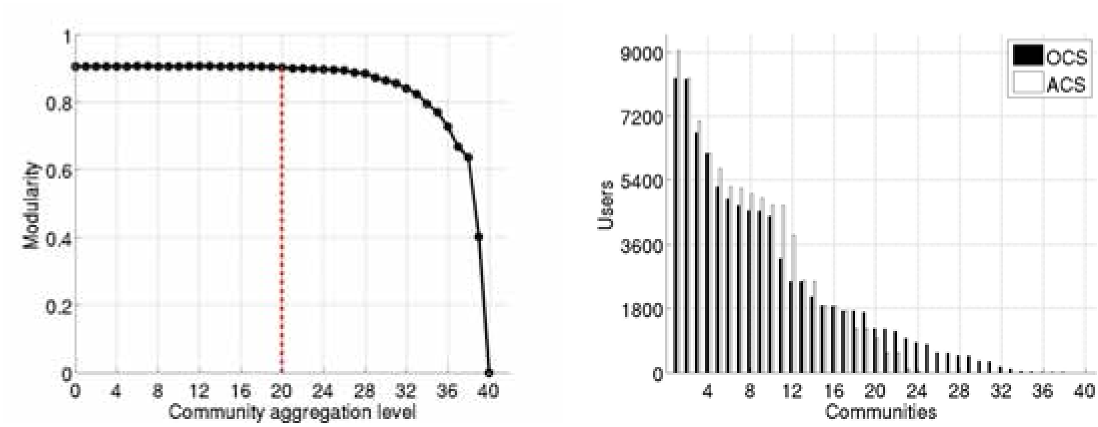
Generally, a network is a set of nodes connected by links. It is possibile to add two kind of information to the links: direction and weights (magnitudes). For WDS networks, direction and weight of links can consist of direction and amount of flow rate in pipes. Thus Eq. (2) becomes (Newman 2004, Yook et al. 2001, Leicht and Newman 2008)

$$Q = \frac{1}{W} \sum_{ij} \left(W_{ij} - \frac{s_i^{out} s_j^{in}}{W} \right) \delta \left(C_i, C_j \right), \tag{3}$$

where W_{ij} is the non-symmetric weights matrix, W indicates the sum of weights, s_i^{out} and s_j^{in} represent the outcome strength of node i and the income strength of node j, respectively; for the generic i-th node, they are evaluated as $s_i^{out} = \sum_j W_{ij}$ and $s_j^{in} = \sum_i W_{ij}$.

In order to obtain a set of communities characterized by a low number of points of incoming/outgoing flow, pipe flows at a single instant, q_{ij}, are chosen as weights of the W_{ij} matrix, where q_{ij} is the unidirectional flow from node i to node j. An aggregated indicator, as time-averaged flows, can be used in place of the flows at a single instant. In this case an analysis of how community boundaries vary with changes can be relevant. The unidirectionality of the flow ensures that $q_{ji} = 0$ if $q_{ij} \neq 0$. Therefore, the weights matrix reads

$$W_{ij} = q_{ij}. \tag{4}$$



(a) (b)

Fig. 1: *(a)* Modularity variation due to community aggregation. The red line shows the aggregation level at which modularity begins to decrease - *(b)* User number of each community.

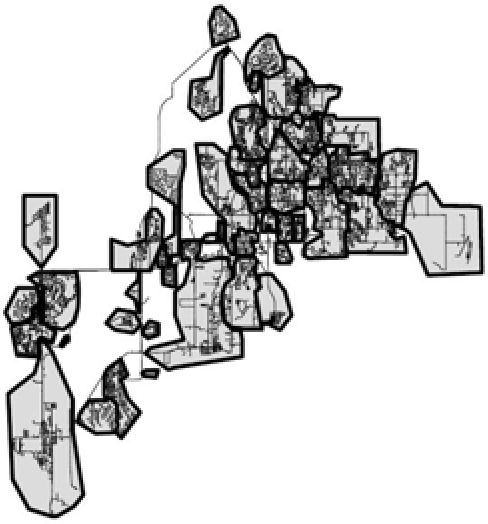
In Blondel et al. 2008, a fast and simple iterative algorithm for modularity evaluation is suggested. The method, based on fast-greeding technique, has demonstrated very good performance and computational complexity linear in the network size (Lancichinetti and Fortunato 2009). At the first step, each node is allocated to a different community. Then, for the i-th node, the algorithm evaluates the raise in modularity that would take place by moving that node from its community to another to which it is connected. The procedure goes on for all nodes till modularity no longer increases. In the second step a new network is built. Such network, in which communities created in the first step are nodes, is passed to the first step and so on. At the end of Blondel algorithm, although modularity is maximized, it is possible that some of the resulting communities have a impractically small size. To overcome this limit, we propose to go ahead in the aggregation procedure even if this could lower the modularity. Iteratively, the community with the smallest modularity is merged with one of its adjacent communities. The aggregation which provides the highest modularity is chosen in case of more than one adjacent community.

Diao et al. 2013 proposed a procedure based on Blondel approach. In particular they manually chose link weights to ensure that the method creates a new DMA at the point where a smaller pipe splits off the larger transmission main.

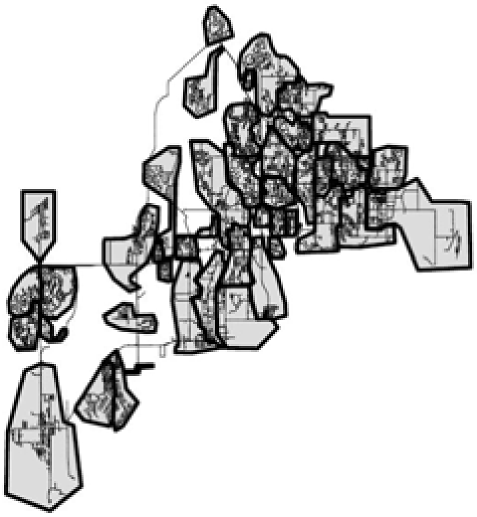
Thus DMA boundaries identification was influenced by the user's choice of weights. In contrast, the method here described uses flows as weights without any arbitrariness in the weight choice.

3. Application

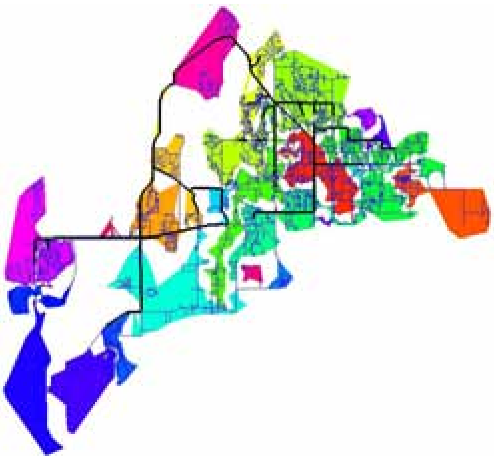
The proposed approach was tested on the Battle of Water-Sensors Network 2 (BWSN2) (Ostfeld et al. 2008). The network, consisting of 12'523 nodes and 14'822 pipes, has been chosen because it has already been used for DMA creation methods by Savic et al. 2009 and Diao et al. 2013. In both works, the authors subdivide the networks in DMAs.



(a) *43 DMAs - From Savic et al. 2009.*



(b) *41 DMAs - From Diao et al. 2013.*



(c) *41 DMAs - OCS case.*

(d) *21 DMAS - ACS case.*

Fig. 2: Case study network subdivided in DMAs.

Blondel algorithm provides good results, with 41 DMAs and a modularity $Q = 0.90$. The analysis has been executed in 0.2 s using the online optimization tool provided by ISI Foundation (`www.netcom-analyzer.org`). The analysis has been gone on, as described in previous section, performing community aggregation. Results displayed in Fig. 1(a) show a slight decrease in modularity after the 20th aggregation level, that becomes much higher after the

28th one. Modularity goes to zero after 40 aggregations, when only one district remains. Visually, the case at the 20th aggregation level has been identified as a good solution because of its high modularity, very close to the original case, and the size of communities. In the following we refer to the original selection as Original Community Structure (OCS, 41 communities), and to the aggregated case as Aggregated Community Structure (ACS, 21 communities).

Table 1: DMAs characteristics.

DMAs identification method	number of DMA	DMA with less than 500 user connections	DMA with more than 5000 user connections
Savic et al. 2009	41	1	2
Diao et al. 2013	43	3	1
OCS	41	10	5
ACS	21	4	8

The advantage of choosing ACS instead of OCS case lies in user connections distribution, that are points that represent connection between domestic water plants and WDS. To satisfy user demands, water must be moreover available with a suitable pressure. While in OCS case there are 10 DMAs with less than 500 connections and 5 larger than 5'000, in ACS, although there are 8 DMAs larger than 5'000 connections, only 4 DMAs have less than 500 connections (Fig. 1(b)), that is coherent with the others two works considered here (Tab. 1). Savic et al. 2009 manually redesigned the network dividing it into 43 DMAs,And as a result of these efforts three DMAs have less than 500 connections and only one more than 5'000. Diao et al. 2013 instead used an automatic procedure that gives 41 DMAs, only one of which has less than 500 connections and two more than 5'000 (Tab. 1). In Fig. 2 all the cited districtualizations applied to BWSN2 network are showed. Boundaries are similar for all cases. The only differences emerge in ACS case (21 DMAs), where smaller DMAs were aggregated in larger ones.

4. Conclusions

In this paper, the problem of DMA boundary identification has been faced using a complex network approach. The method, a fast-greeding algorithm, has been used because of its simplicity and efficiency. In fact, it presents a conceptual clarity that makes it easy to use; moreover it is one of the fastest algorithms in the field of community detection. The application on a benchmark network (BWSN2) allowed for a comparison with other similar works. Results are good, especially in terms of speed and quality. In fact the number of communities is acceptable and community sizes are coherent with the number of users per DMA suggested by technical guidelines. The case here described represents a good example of how complex network theory can provide novel useful methods to solve classical engineering problems.

Acknowledgements

We gratefully acknowledge SMAT Group for the financial support to this research and ISI Foundation for providing the online software.

References

Albert, R., Barabasi, A.L., 2002. Statistical mechanics of complex networks. Rev. Mod. Phys. 74, 47–97.
Albert, R., Jeong, H., Barabasi, A.L., 2000. Error and attack tolerance of complex networks. Nature 406, 378–382.
Albert, R., Jeong, H., L., B.A., 1999. Internet - Diameter of the World-Wide Web. Nature 401, 130–131.
Barabasi, A.L., Albert, R., 1999. Emergence of scaling in random networks. Science 286, 509–512.
Blondel, V.D., Guillaume, J.L., Lambiotte, R., Lefebvre, E., 2008. Fast unfolding of communities in large networks. J. Stat. Mech. .
Boccaletti, S., Latora, V., Moreno, Y., Chavez, M., Hwang, D.U., 2006. Complex networks: Structure and dynamics. Phys. Rep. 424, 175–308.
Diao, K., Zhou, Y., Rauch, W., 2013. Automated creation of district metered area boundaries in water distribution systems. Journal of Water Resources Planning and Management 139, 184–190.
Donath, W.E., Hoffman, A.J., 1973. Lower bounds for the partitioning of graphs. IBM Journal of Research and Development 17, 420–425.

Faloutsos, M., Faloutsos, P., Faloutsos, C., 1999. On power-law relationships of the internet topology. SIGCOMM Comput. Commun. Rev. 29, 251–262.

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

Hastie, T., Tibshirani, R., Friedman, J., 2009. The Elements of Statistical Learning. Springer, New York.

Herrera, M., Izquierdo, J., Perez-Garcia, R., Montalvo, I., 2012. Multi-agent adaptive boosting on semi-supervised water supply clusters. Adv. Eng. Soft. 50, 131–136.

Hlaoui, A., Wang, S., 2004. A direct approach to graph clustering., in: Neural Networks and Computational Intelligence, IASTED/ACTA Press. pp. 158–163.

Hunaidi, O., 2012. Economic comparison of periodic acoustic surveys adn DMA-based., in: Leakage 2005 Conference Proceedings, pp. 322–336.

IWA, 2007a. District metered areas - guidance notes.

IWA, 2007b. Leak location and repair - guidance notes.

Jeong, H., Tombor, B., Albert, R., Oltvai, Z., Barabasi, A., 2000. The large-scale organization of metabolic networks. Nature 407, 651–654.

Kernighan, B.W., Lin, S., 1970. An efficient heuristic procedure for partitioning graphs. The Bell system technical journal 49, 291–307.

Lancichinetti, A., Fortunato, S., 2009. Community detection algorithms: A comparative analysis. PHYSICAL REVIEW E 80. doi:{10.1103/PhysRevE.80.056117}.

Leicht, E.A., Newman, M.E.J., 2008. Community structure in directed networks. Phys. Rev. Lett. 100, 118703.

Newman, M.E.J., 2004. Analysis of weighted networks. Phys. Rev. E 70, 056131.

Newman, M.E.J., Girvan, M., 2004. Finding and evaluating community structure in networks. Phys. Rev. E 69, 026113.

Ostfeld, A., Uber, J., Salomons, E., Berry, J., Hart, W., Phillips, C., Watson, J., Dorini, G., Jonkergouw, P., Kapelan, Z., di Pierro, F., Khu, S., Savic, D., Eliades, D., Polycarpou, M., Ghimire, S., Barkdoll, B., Gueli, R., Huang, J., McBean, E., James, W., Krause, A., Leskovec, J., Isovitsch, S., Xu, J., Guestrin, C., VanBriesen, J., Small, M., Fischbeck, P., Preis, A., Propato, M., Piller, O., Trachtman, G., Wu, Z., Walski, T., 2008. The battle of the water sensor networks (bwsn): A design challenge for engineers and algorithms. Journal of Water Resources Planning and Management 134, 556–568.

Redner, S., 1998. How popular is your paper? An empirical study of the citation distribution. Eur. Phys. J. B. 4, 131–134.

Savic, D.A., Grayman, W.M., Murray, R., 2009. Effects of redesign of water systems for security and water quality factors, in: Integrating Water Systems, 10th International on Computing and Control for the Water Industry.

Strogatz, S.H., 2001. Exploring complex networks. Nature 410, 268–276.

Tzatchkov, V.G., Alcocer-Yamanaka V. H. Ortiz, V.B., 2006. Graph Theory Based Algorithms for Water Distribution Network Sectorization Projects. chapter 81. pp. 1–15.

Watts, D.J., Strogatz, S.H., 1998. Collective dynamics of 'small-world' networks. Nature 393, 440–442.

Yook, S., Jeong, H., Barabasi, A., Tu, Y., 2001. Weighted evolving networks. Phys. Rev. Lett. 86, 5835–5838.