



Jun 17th, 10:40 AM - 12:00 PM

Graph Clustering Based on Social Network Community Detection Algorithms

Enrique Campbell

Universitat Politècnica de València, encamgo1@upv.es

David Ayala-Cabrera

Universitat Politècnica de València, daaycab@upv.es

Joaquín Izquierdo

Universitat Politècnica de València, jizquier@upv.es

Rafael Pérez-García

Universitat Politècnica de València, rperez@upv.es

Follow this and additional works at: <https://scholarsarchive.byu.edu/iemssconference>



Part of the [Civil Engineering Commons](#), [Data Storage Systems Commons](#), [Environmental Engineering Commons](#), [Hydraulic Engineering Commons](#), and the [Other Civil and Environmental Engineering Commons](#)

Campbell, Enrique; Ayala-Cabrera, David; Izquierdo, Joaquín; and Pérez-García, Rafael, "Graph Clustering Based on Social Network Community Detection Algorithms" (2014). *International Congress on Environmental Modelling and Software*. 17.

<https://scholarsarchive.byu.edu/iemssconference/2014/Stream-G/17>

This Event is brought to you for free and open access by the Civil and Environmental Engineering at BYU ScholarsArchive. It has been accepted for inclusion in International Congress on Environmental Modelling and Software by an authorized administrator of BYU ScholarsArchive. For more information, please contact scholarsarchive@byu.edu, ellen_amatangelo@byu.edu.

Graph Clustering Based on Social Network Community Detection Algorithms

Enrique Campbell ^a, David Ayala-Cabrera^a, Joaquín Izquierdo^a and Rafael Pérez-García^a

^aFluIng-IMM, Universitat Politècnica de València
València, Spain

encamgo1@upv.es, daaycab@upv.es, jizquier@upv.es, rperez@upv.es

Abstract: Graph network clustering has been an important field of investigation in recent years with a wide spectrum of applications, including environmental, genetic and engineering studies. There are two main types of applications; their suitability depends of the modeling context. Cases where the number of subdivisions is a priori known are tackled with different techniques than those where such number is not clearly evident. The second falls within the field of community detection, which has been broadly studied by social network scientists. Traditionally, community detection in graphs has been tackled by means of hierarchical clustering, which groups nodes based on their similarities. However, given the uncertainty that entails the determination of the number of clusters, other more accurate and efficient methods have been put forward. Specifically, the so-called walktrap algorithm aims to detect communities in graphs based on the idea that random walks tend to get trapped within communities (areas with higher density of links and separated by few connections). In this paper, an application of some social network concepts in the field of engineering is described. As an example, they are used to sectorize the water supply network of the Battle of the Water Calibration Networks.

Keywords: Water; Sectorization; Network; Cluster; Graph; Communities

1 COMMUNITY DETECTION IN GRAPHS

Social network theory (SNT) is a field of investigation oriented to study the relations among individual elements bound by bridges entities. A social network (SN) can be abstracted from any system with inter-related individuals. For example, friendship among people is represented as SN in the web space. One of the most important aspects in SNT is community detection, which allows identifying groups of nodes sharing common patterns and differing from other groups, without knowing a priori the number of communities (or clusters) into which the graph should be subdivided.

1.1 Network community detection through hierarchical clustering

The detection of communities by means of hierarchical clustering (HC) is performed over a square matrix of dissimilarities between the nodes. In the first stage each node is set as a single cluster and then depending on a dissimilarity measure (e.g. euclidean distance) and on an agglomeration method (e.g. average) they are joined until all the clusters are part of a unique whole cluster. In this community detection process the hierarchical merges are depicted as a dendrogram, which is a graphical structure where all the nodes are initially represented as individual communities and at the end they are joined in a single community. Despite its simplicity, the technique has some disadvantages: it tends to create clusters of single nodes in the outer regions of the graphs and there are not clear criteria regarding the most representative clusters that can be found in a dendrogram.

1.2 Network community detection through random walks

The drawbacks of HC have inspired the design of new algorithms to detect communities in SNs in a more effective manner (Fortunato, 2010). One of them is the walktrap algorithm which works based on the idea that random walks throughout the graph tend to detect subgraphs (areas of the graph with high edge density) as there are only few links that lead outside a given community. The algorithm subdivides the network in communities in such a way that a score known as modularity is maximized. The former score reflects the fact that communities in a graph are formed by dense subgraphs connected by few links (Newman and Girvan, 2004; Clauset et al., 2004). Similarly to HC, when walktrap algorithm is executed, the node merges can be represented as a dendrogram. The exploration of this structure allows the different possible arrangements of communities in a given network to be determined. Performance comparisons of the method with other similar methods, using widely known networks as the "Zachary's club network" (Zachary, 1977) and the "protein interaction network" (Jeong et al., 2001) have demonstrated its advantages in term of the quality of results, and have ranked it among the best methods in terms of running time (Pons and Lapaty, 2006).

2 SECTORIZATION OF WATER SUPPLY NETWORKS

Sectorization is an operational approach currently used by many water utilities, mainly in Europe and Latin America, in order to have better control over the systems, especially over non revenue water (NRW) related aspects. The technique entails the subdivision of the water supply network (WSN) into sectors, which are areas of the WSN isolated from the rest of the network by closing some pipelines and installing flow meters in a single line of each sector. The main added value of implementing this technique is related to the potential to reduce the over costs produced by water losses. Water losses represent one of the main problems for water utilities. In some countries (especially developing countries), leakage is estimated to be of the order of 50% of the water into supply (Kingdom et al., 2006; Thorton et al., 2008).

This technique was first implemented in two cities in the UK in the early 1980s. Since this time, several reports and guidance have been published on this technique. The guidance document proposed by Morrison et al. (2007), is one of the most important and widely known. Despite the wide level of acceptance of the technique in many countries, most of the implementation cases follow a trial-error approach and also lack follow-up after implementation. The technical recommendations for selecting the size of the sectors are very general. The best recognized guidance recommends the use of a wide range of number of connections (500-3000 connections) (Morrison et al., 2007) or a wide range of pipe length (4-30 km), depending on the desired accuracy of water loss control (GIZ et al., 2011). In general, small sectors (whether it be a small extension pipe length or a small number of household connections) deliver more accuracy to detect faults. However, they also entail higher investment in the purchase of more valves and flow meters. Large sectors implementation is expected to be less expensive, but less sensitive to detecting significant changes in the inflow behavior caused by a fault (e.g a burst).

There are some remarkable drawbacks in implementing sectorization in a WSN: increase of head losses due to pipe wall friction, which may be translated into water supply shortages and downfall of the capacity of the network to overcome problematic situations. Also, as some of the pipes are closed, velocity may fall to critical values at dead end points. In these zones biological quality may be compromised.

Considering all the aspects discussed above it can be concluded that in a effective layout of sectorization, the level of pressure and leakage control and operational aspects should be optimally balanced. The layout also has to minimize the number of required boundary valves and flow meters, in order to make the project economically suitable.

During the last decade several computer based methods have been put forward in order to create sectorization schemes for WSNs. Essentially, these studies have used graph techniques along with

mathematical optimization tools to treat the sectorization problem as a graph partitioning problem. Most of them have focused on the detection of the influence zone of each water source in the WSN (Tzatchkov et al., 2008; Di Nardo and Di Natale, 2011; Herrera et al., 2012; Tzatchkov and Alcocer-Yamanaka, 2012; Hajebi et al., 2013; Alvisi and Franchini, 2014; Di Nardo et al., 2014).

It is very important to consider that if the WSN relies on many water sources, each sector might be defined around one or a few sources. In other networks (especially WSNs of extensive urban centers), the establishment of this type of layout might not be feasible if sources are located far from the consumers. In that case, water supply depends on a supply trunk. Therefore, the sectors cannot have their own water source (or sources) and the water entrances should be connected to the supply trunk. Although there are many cities with this type of layout, only a few studies have tackled this problem (Diao et al., 2013, Campbell et al., 2014).

2.1 Sectorization of WSN using SNT concepts

As stated above, SN can be abstracted from any system with inter-related individual elements. WSNs are not the exception. In this case pipes represent the edges that connect nodes (connection points of pipes). Once a SN of a WSN is transformed into a graph, it can be analyzed with methods offered by the SNT. Element's Centrality is a property of major interest. It describes the importance of elements (either node or edge) in the network. The centrality of each element is evaluated according to its capacity to connect regions of elements within the network. One of the centrality measures that better describes the global centrality in SNs is betweenness. In the case of edges, betweenness is equal to the number of shortest paths from all nodes to all others that pass through each edge.

When WSNs depends on a supply trunk the best sectorization approach is community detection. A WSN can be considered a union of subgraphs given that their layouts respond to urban needs. Any urban block (e.g. neighborhood or conglomeration of industrial facilities) relies on a subnetwork of minor diameter pipes connected to other subnetworks or to a trunk by means of a few larger diameter pipes. If trunk pipes are temporally segregated from the distribution network, unconnected sections of the graphs appear. In such disconnected sections the community detection algorithms can be applied to detect the community layouts. Then, by exploration of the dendrogram, the communities that best accomplish the operational criteria can be defined.

3 EXAMPLE OF IMPLEMENTATION

3.1 Outline of the method

In this section a methodology to sectorize WSNs is presented (see Figure1). It is based on the concepts discussed above. The steps of the method are the following: 1) analysis of the trunk centrality; 2) network community detection algorithm; 3) additional merges of communities based on operational criteria; 4) determination of feed pipe(s) for each sector; 5) hydraulic validation. The network used as example is the network of the Battle of the Water Calibration Networks (BWCN) (388 nodes, 1 reservoir, 7 tanks)(Ostfeld et al., 2012).

The trunk corresponds to the pipes with major centrality, except in the case of the pipes connected to the sources. However, these pipes are recognized by their larger diameters. It is worth noting that the definition of the trunk and segregation requires the establishment of a factor that defines the threshold from which pipes are considered part of the trunk. Such a factor has to be defined according to the experience of the technical staff of the water utilities.

Once the trunk is segregated from the distribution network, some completely isolated subgraphs with only one potential entrance appear. Those areas can be directly set as valid sectors if they meet the size limit constraints. There may be other areas that are too extensive to be considered single sectors.

To deal with this, all the potential subgraphs are estimated using community detection algorithms, then the communities are explored by means of the dendrogram. The community layout that best fits to the size limit requirement is selected as the appropriate one. It is worth noting that the appropriateness of the size limit has to be defined by the water utility technical staff based upon the required accuracy for detecting any change in the inflow of the sectors.

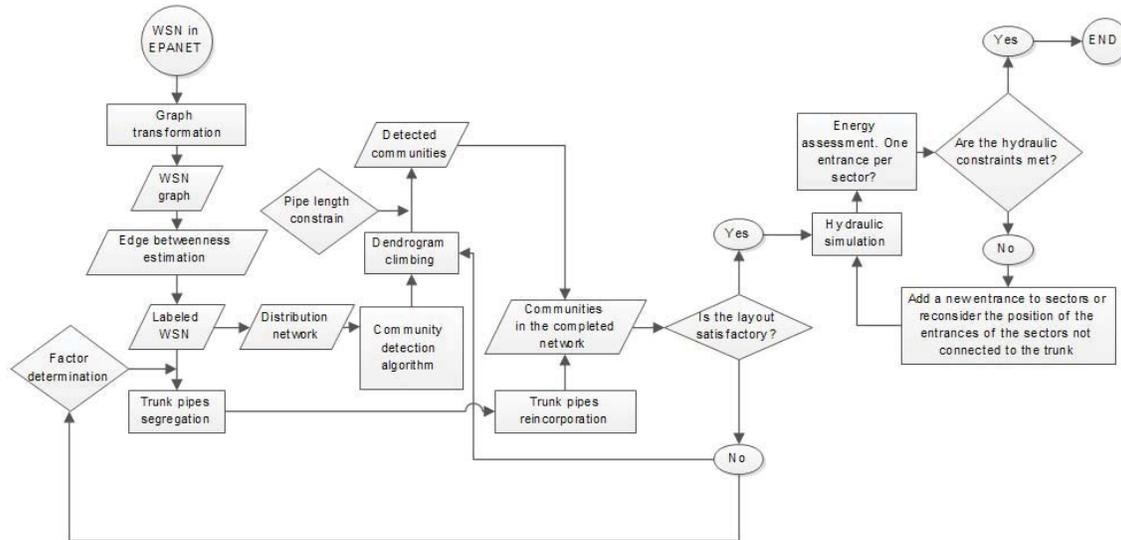
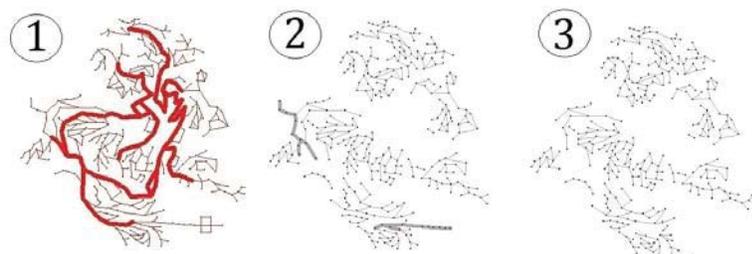


Figure 1. Flow chart of the proposed methodology

The methods used to detect communities in the unconnected graphs are: HC and the walktrap algorithm. By using the maximal modularity criterion, the walktrap algorithm tends to find small communities that might be out of the required limits of pipe length (normally 4-30 km). Through further merges, this constraint may be overcome.

3.2 Obtained results

In Figure 2 it can be seen that the trunk corresponds to the lines with the largest diameters. Also, it is evident how important the trunk is for transporting water from one point to other in the network.



- 1: Trunk network is characterized by high values of betweenness
- 2: Lines connected to sources have largest diameters
- 3: Most central lines and lines connected to sources are segregated, leaving only the distribution network

Figure 2. Trunk network segregation

By cutting the dendrogram produced by the application of HC at different heights, different schemes of communities are detected. For this examples the best configuration was found at height 0.1 (see Figure 3). Nevertheless, nodes belonging to the same cluster may be found in disjoint subgraphs (see circled areas), which is unacceptable.

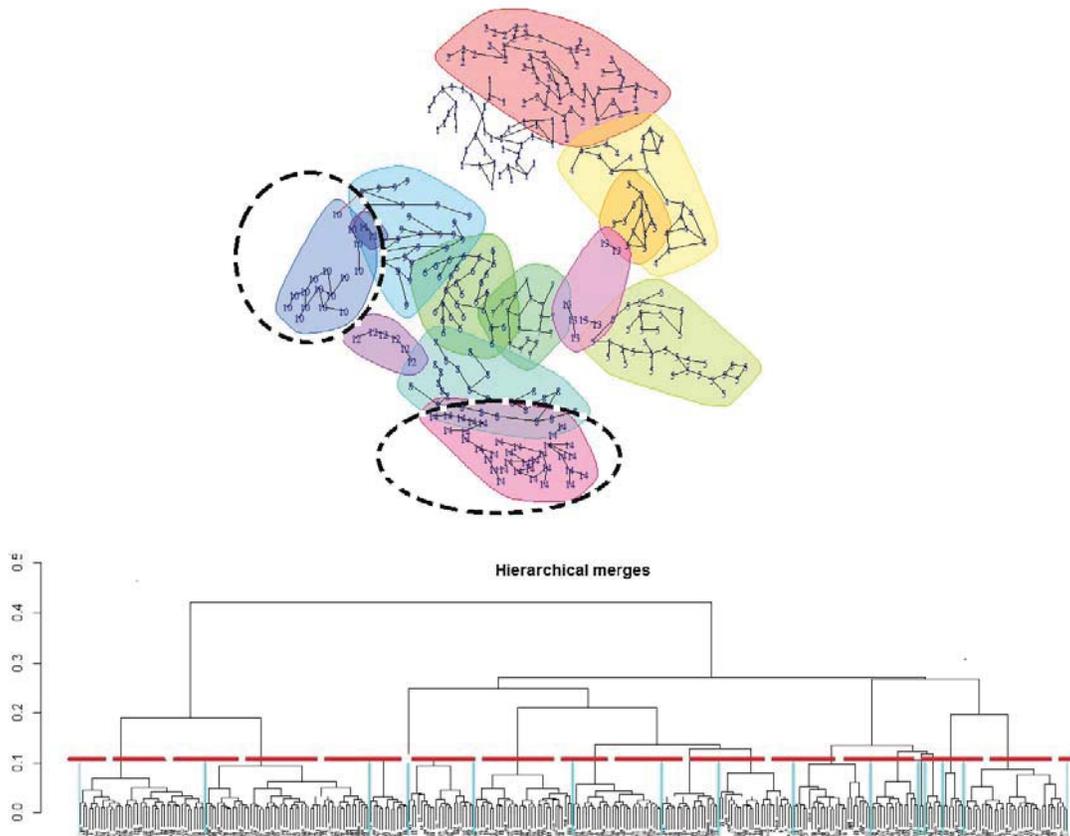


Figure 3. HC results. Top: communities detected. Bottom: dendrogram obtained

The outcome of the walktrap algorithm implementation is also a dendrogram, but it is rather different from the one above (Figure 3). The differences among the dendrograms are related to the differences in the merging criterion used by each one (HC uses dissimilarity measure combined with an agglomeration method whereas walktrap uses modularity score). Figure 4 shows the different communities detected at two levels of the dendrogram produced by the walktrap algorithm. Image 1 corresponds to the communities with the highest score of modularity. For this example small sectors are not desirable as they would represent larger investment in the purchase of flow meters. Therefore, further merges were conducted until the layout was satisfactory, which is represented by image 2. It is important to note that currently the additional merges have to be conducted by the modeler. Nonetheless, it is expected that in the future this can be automatically executed.

From the result above, it is evident that application of the walktrap algorithm provides better results than the use of HC. In this case 12 sectors were set (see Figure 5). It is worth stressing that the layout with the maximal modularity score detects small subgraphs, which can be acceptable in WSNs with very high rates of NRW.

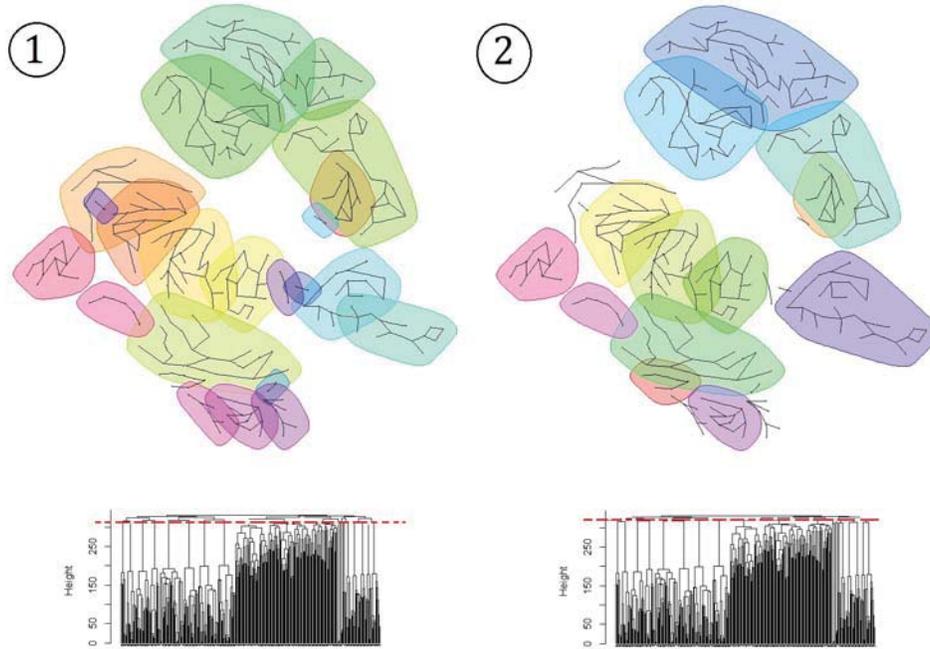


Figure 4. Outputs from the Walktrap algorithm. Left: communities with maximal modularity. Right: communities in the selected layout

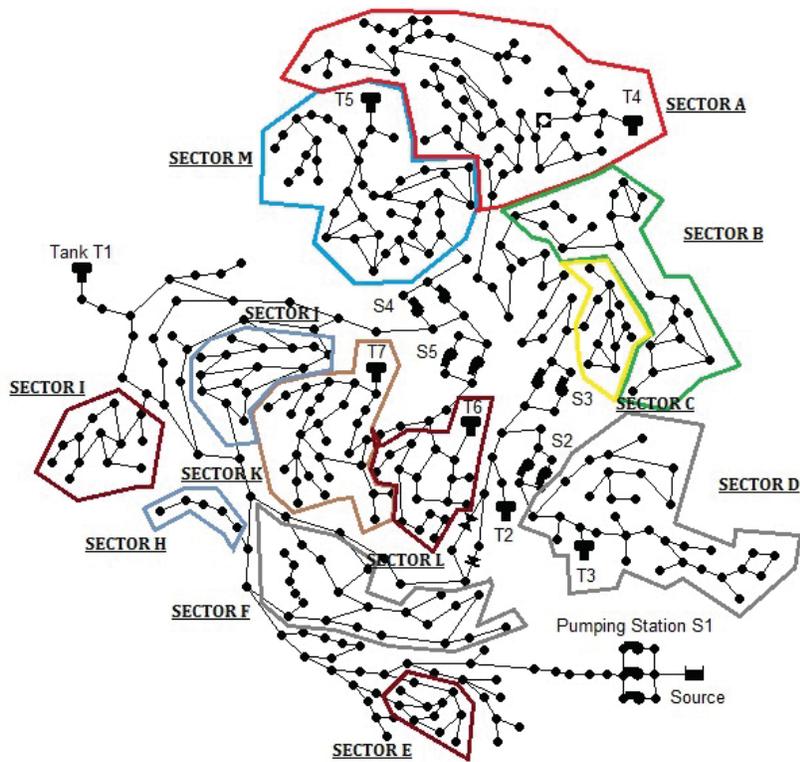


Figure 5. WSN sectorization final scheme

4 CONCLUSIONS AND RECOMMENDATIONS

This work describes the application of SNT concepts to detect communities in graphs. It has shown how the concept of clustering in SNT can be exploited to tackle real problems in WSN management. In this work the sectorization problem in networks where the number and size of sectors are initially unknown is approached by a combination of the centrality measure known as betweenness and community detection algorithms (HC and the walktrap algorithm). The example of implementation in the network of the BWCN demonstrated the appropriateness of the application of these techniques, although it is very important to challenge the method in a larger extension network. It is worth noting that this work embraces the sectorization of networks dependent of a trunk, which has been considered only by few previous and recent studies, despite the large significant number of networks that have this feature.

REFERENCES

- Alvisi, S., Franchini, M., 2014. A heuristic procedure for the automatic creation of district metered areas in water distribution systems. *Urban Water J* 11(2), 137-159.
- Campbell, E., Ayala-Cabrera, D., Izquierdo J., Pérez-García, R., Tavera, M., 2014. Label propagation algorithm based methodology for water supply networks sectorization. *Int.J.Comp.Syst.Sci.* 4(1), 35-39.
- Clauset, A., Newman, M.E.J., Moore, C., 2004. Finding community structure in very large networks. *Phys. Rev. E.* 70.
- Diao, K., Zhou, Y., Rauch, W., 2013. Automated creation of district metered area boundaries in water distribution systems. *J. Water Resour. Plann. Manage.* 139(2), 184 - 190.
- Di Nardo, A., Di Natale, M., 2011. A heuristic design support methodology based on graph theory for district metering of water supply networks. *Eng Optimiz* 43(2), 193-211.
- Di Nardo, A., Di Natale, M., Santonastaso, G. F., Tzatchkov, V., Alcocer-Yamanaka, V., 2014. Water network sectorization based on graph theory and energy performance indices. *J. Water Resour. Plann. Manage.* 140(5), 620-629.
- Fortunato, S., 2010. Community detection in graphs. *Phys. Rep.* 486(3-5), 75-174.
- GIZ (Deutsche Gesellschaft für Internationale Zusammenarbeit GmbH), VAG (Armaturen GmbH), FHNW (Fachhochschule Nordwestschweiz), KIT (Karlsruhe Institute of Technology), Institute for Ecopreneurship (IEC), 2011. Guidelines for water loss reduction: a focus on pressure management. Eschborn.
- Hajebi, S., Barrett, S., Clarke, A., Clarke, S., 2013. Multi-agent simulation to automate water distribution network partitioning, *Proc. 27th European Simulation and Modelling Conference - ESM2013.*
- Herrera, M., Izquierdo, J., Pérez-García, R., Montalvo, I., 2012. Multi-agent adaptive boosting on semi-supervised water supply clusters. *Adv Eng Softw* 50(August 2012), 131-136.
- Jeong, H., Mason, S.P., Barabási, A.-L., Oltvai, Z.N., 2001. Centrality and lethality of protein networks. *Nature* 411, 41-42
- Kingdom, B., Liemberger, R., Marin, P., 2006. The Challenge of Reducing Non-Revenue Water (NRW) in Developing Countries. How the private sector can help: a look at performance-based service contracting. The World Bank Group, Washington.
- Morrison, J., Stephen, T., Rogers, D., 2007. District Metered Areas: Guidance Notes. Technical Report 1. Water Loss Task Force - International Water Association (IWA).
- Newman, M.E.J., Girvan, M., 2004. Finding and evaluating community structure in networks. *Phys. Rev. E.* 69.
- Ostfeld, A., Salomons, E., Ormsbee, L., et al., 2012. Battle of the Water Calibration Networks. *J. Water Resour. Plann. Manage.* 138(5), 523-532.
- Pons, P., Lapaty, M., 2006. Computing communities in large networks using random walks. *J. Graph Algorithms Appl* 10(2), 191-218.
- Thornton, J., Sturm, R., Kunkel, G., 2008. *Water Loss Control.* McGraw-Hill.
- Tzatchkov, V., Alcocer Yamanaka, V., Bourguett Ortiz, V., 2008. Sectorización de redes de distribución de agua potable a Través de algoritmos basados en la teoría de grafos. *TI áloc AMH*, 40(Enero-Febrero 2008), 14-22.

- Tzatchkov, V., Alcocer-Yamanaka V., 2012. Graph partitioning algorithms for water distribution network sectorization projects, Proc. 10th International Conference on Hydroinformatics.
- Zachary, W., 1977. An information flow model for conflicts and fission in small groups. *J Anthropol Res*, 33, 452-473.