International Environmental Modelling and Software Society (iEMSs)
2012 International Congress on Environmental Modelling and Software
Managing Resources of a Limited Planet, Sixth Biennial Meeting, Leipzig, Germany
R. Seppelt, A.A. Voinov, S. Lange, D. Bankamp (Eds.)
http://www.iemss.org/society/index.php/iemss-2012-proceedings

A multi-level spectral clustering process to ascertain sensor location for mitigating effects of a potential contamination in a water supply network

M. Herrera [a], **J.A. Gutiérrez-Pérez**[a], **J. Izquierdo**[a], **and R. Pérez-García**[a]

[a] *Fluing-IMM, Universitat Politècnica de València, Spain (mahefe@upv.es, joagupre@upv.es, jizquier@upv.es, rperez@upv.es)*

Abstract: In this paper, we introduce a multi-level methodology based on the iteration of successive processes of spectral clustering to divide a water supply network (WSN) into metric subsystems, also called district metered areas (DMAs). Each one of the above-mentioned divisions is approached by spectral clustering. This is a graph clustering methodology in data analysis that improves the straightforward application of K-means, works well in non-convex spaces, and takes into account the underlying graph structure under study. Spectral clustering uses information obtained from computing the eigenvalues and eigenvectors of the Laplacian matrices obtained from partitioning the graphs and searches a minimum number of cut edges to achieve it.

Our aim is to propose suitable conditions to approach useful characterizations of DMAs. In addition, we try to take advantage from this reduction of the inspection area in technical management tasks, such as sensor location. In this work, an experimental study based on a real WSN is proposed. An iterative nested division into DMAs is used to locate sensors throughout the whole network and to perform inference tasks on the presence of contamination events based on those sensor signals. If the sensors are located in separated areas weakly interconnected with each other, it will be easier to mitigate the effects of potential contamination, having at one's disposal a minimal number of cut-off valves for security. In addition, the nested nature of the proposed cluster construction increases reliability in mitigation plans. The method is scalable and can be generalized to address other managerial issues related to water quality and leakages, among others.

Keywords: Water supply systems, sensor location, spectral clustering

1 INTRODUCTION

Over the past decade there have been important research contributions on optimal sensor location in a water supply network (WSN) to detect, intentional or accidental, contamination events. Thus, part of the fundamental references in this area are the works of Ostfeld and Salomons [2004]; Gueli [2006]; Propato [2006], among others. In all these works, the optimal sensor location is made by the study of the evolution of water flow jointly to the contaminant behaviour. These approaches involve a huge computational effort, which joins to the model calibration and the difficulty to simulate events in different time, place, and durations. A problem formulation based on optimisation models that only use hydraulic criteria (Berry et al. [2005]) find difficulties in large WSNs (Drezner and Hamacher [2002]) since this location problem is *NP-hard* and there are no algorithms that can resolve it in polynomial time. Sevkli and Guner [2006] and Shen et al.

[2011] provide heuristics that can give approximate solutions to the location problem in a general network. However, they do not take into account the hydraulic conditions of the network.

In this work we propose a different approach to the sensor location in a WSN problem[1]. Our idea is to place them such that facilitates the mitigation tasks of any possible contamination event, once detected. The objective is to provide a security system which, in an efficient manner, prevents the spread of contaminants throughout the WSN; doing so the consequences are small as possible. We will use Graph Theory and an iterative process for establishing clusters (hydraulic zones), which arranged in a nested architecture provide adequate reliability that supports possible mitigation tasks. The spectral clustering algorithm (Ng et al. [2001]) can divide the network taking into account the hydraulic, geographic, and topological characteristics of a WSN (Herrera et al. [2010]; Herrera [2011]). Therefore, this is the key part of the proposal of this work, together with a measure of the relative importance of each of the areas involving in this division (Gutiérrez-Pérez et al. [2012]).

A suitable combination of these WSN division processes into zones (and sub-zones) and the subsequent assessment of their relative importance, implicitly includes the design goals outlined in the Battle of the Sensor Water Networks (BSWN, Ostfeld and *et al.* [2008]). These objectives are:

- expected time of detection
- expected population affected prior to detection
- expected consumption of contaminated water prior to detection
- detection likelihood

The rest of this document is structured as follows. In the second section it is introduced the process of spectral clustering, suggesting understand a WSN as a graph of special characteristics. Later it is defined the nested clustering which ends with the sensor location in the network. The development of an experimental study is performed in the following Section 3. Section 4 discuss the results and closes the article proposing future challenges.

2 PROCESS DESCRIPTION AND STATED OBJECTIVES

The process proposed is related to the identification of those areas that are particularly sensitive to a contamination event, taking there a greater impact than in other areas the same event. The decision to locate a larger number of sensors will be in these areas. The process of spectral clustering represents a double advantage to search these interesting areas: it separates areas with a minimum number of arcs to each other areas and within each proposed cluster are found the high interconnected nodes. If these characteristics are transferred to the problem of sensor location in a WSN, we can see that our proposal has the following good properties:

- Areas of greater connectivity and relative importance of nodes (Herrera et al. [2011]) also represent an increased consumption and flow transfer. Thus, they are more sensitive in the occurrence of contamination events. Spectral clustering detects these areas.

[1]The approach involves simplifying assumptions respect the network contaminant transport and sensor response, among others.

- A poor connection from one area to other also causes delays in the spread of the contaminant. This means that despite the case we do not work with a sectioned WSN, early detection could prevent or minimise the consequences of contaminant dispersion. Again, spectral clustering algorithms minimises the number of arcs (pipes) connecting one area to another.
- A nested spectral clustering increases the accuracy working with most sensitive to contamination areas. This makes easy sensor location focusing the problem in greater interest and smaller size sub-systems.
- As a tool for mitigating potential effects of contaminant dispersion, spectral clustering proposes a sectorisation of the WSN (Herrera et al. [2010]) with a minimum number of cut-off valves.

In addition to this network division, it will be useful to work with a measure of the nodes that compose the previously obtained hydraulic areas (Grubesic et al. [2008]; Yazdani and Jeffrey [2010]). This could be both a measure of connectivity of each subgraph (Yazdani and Jeffrey [2010]) and a measure for ranking the sub-network nodes. This second option may have a suitable adaptation to work with the specific case of a WSN graph. This paper proposes an adaptation of Google's PageRank (Brin and Page [1998]) as a ranking of these nodes (Gutiérrez-Pérez et al. [2012]; Herrera [2011]).

The advantage of consider this algorithm is that it is based on similar calculations to the spectral clustering algorithm, since both are based on eigenvectors matrices: spectral clustering is performed by a certain transformation of the Laplacian of the graph and PageRank by the Google's matrix of transitions between nodes.

2.1 Spectral clustering

Cluster analysis based on the spectrum of a matrix (Ng et al. [2001]) is a powerful technique in data analysis that has found increasing support and application in many areas. It improves the straightforward application of k-means, working well in non-convex spaces taking into account the possible graph structure under study. Spectral methods are based on the eigenvalues and eigenvectors of a block-diagonal matrix conveniently associated with the graph, and which understands a graph as a collection of k disjoint cliques. Their normalised Laplacian is a block-diagonal matrix that has an eigenvalue of zero with multiplicity k; and the corresponding eigenvectors serve as indicator functions of membership in the corresponding cliques. Any deviation caused by introducing edges between the cliques causes $k-1$ out of the k eigenvalues that were zero to become slightly larger than zero and also causes the corresponding eigenvectors to change. This phenomenon is the basis of spectral clustering, where an eigenvector or a combination of several eigenvectors is used as a vertex similarity measurement for computing the clusters. Thus, the k eigenvectors, which correspond to the k smallest eigenvalues of the affinity matrix's Laplacian, are used to form an $n \times k$ matrix, where each column is normalised to unit length. Treating each row of this matrix as a data point, the algorithm of k-means is finally used to cluster the points.

Spectral clustering is summarised in Table 1, and was introduced by Herrera et al. [2010] in the water sectorisation field.

2.2 Ranking nodes of a network: Google's PageRank application

Chung and Zhao [2008] generalised the Google's PageRank (Brin and Page [1998]) applications to establish sorting vertices by importance in any graph. Understanding the

Table 1: The spectral clustering process

algorithm: supply clusters by spectral clustering
1. abstraction from WSN to a graph
2. construction of Laplacian matrix
3. Matrix transformation into kernel matrix
4. calculus of the matrix spectrum
5. k-means on the 'top' eigenvectors (associated with smallest eigenvalues)
6. Results re-assigning into the original data

WSN as a graph of special characteristics (Herrera [2011]), we can abstract the notion of web-site as a consumption node in the network. Links between pages can be now understood as pipes connecting different nodes. Finally, the random walk idea behind PageRank can be seen as the path which makes a particle throughout a WSN, starting at any node.

The adaptation of Google's PageRank algorithm for WSN graphs raises the possibility of working with a proven efficiency on large databases and has a simple adaptation to a hydraulic network. Ranking nodes can be an useful tool to search efficient criteria in sectorisation. It also is useful for working with vulnerability indices and may be related to studies of rehabilitation plans, water quality analysis, or sensor location (such as is discussed in this paper). The next sub-section shows the role of the PageRank vector in this last application.

2.3 Multi-level clustering

The so called multi-level spectral clustering process is based on following subdivisions of each cluster initially obtained. The idea is to start a new partition in the cluster with a high relative importance nodes. Thus, it is calculated the index that we call *Google's hydraulic* index (gh index) on each cluster (cf. Equation 1).

$$gh_j = \sum gh_{ij} = \frac{\sum w_i d_i}{n_j} \tag{1}$$

where w_i is a certain index of relative importance (ranking) of the node i that is computed with respect the other the ither nodes and links on cluster j (PageRank, in our case); C_j represents each one of the c clusters in which the network is divided; d_i is the water demand at node i; and n_j is the number of nodes in cluster j. This process is repeated until obtain as many clusters as the number of sensors to locate (initially, it is considered a fixed number). The configuration of this multi-level process will be more "depth" in the areas where it is a greater degree of connectivity and relative importance of the nodes. Thus, these areas will be most sensitive to any supply problem, but will be most benefited by the arrangement of the sensors we are proposing. In this case, we try to place sensor (around) at a point i of the cluster j with higher gh (see Equation 1), such that their ranking is maximum. That is:

$$\arg\max_{w_i}\{gh_j \ : C_j \ \ j = 1, \ldots, c\} \tag{2}$$

The final process follows the flow of the algorithm proposed in Table 2.

Table 2: Proposed sensor location process

algorithm: sensor location by spectral clustering + PageRank
1. Spectral clustering of the graph
2. multi-level spectral-clustering
2.0. n-Clusters $<$ n-Sensors: Continue
Otherwise: Go to Step 4.
2.1. gh index calculus
2.2. Spectral clustering in the higher gh index area
Go to 2.0
3. Calculus of PageRank in each subgraph (cluster)
4. Locate the sensor in the higher PageRank node of each cluster

3 EXPERIMENTAL STUDY

In order to show the performance of the presented process it is considered a real case: the WSN of the central area of Celaya (Guanajuato, Mexico); where we want to place 5 sensors. This area is fed by one reservoir ($D1$) and five tanks ($E1, \ldots, E5$) with five pumping stations. The network is made of 479 lines and 333 consumption nodes; total pipe length is 42.5 km and the node elevation average is 156 meters; the total consumed flowrate amounts to 91 l/s (see Figure 1).

A priori, the network is divided into three sectors (in this paper whenever we talk about of divisions, it is understood that they are made by spectral clustering) in which is calculated the gh index to decide which sector is subdivided first. Next, the process is iterated calculating gh in each partition of the graph until obtain as many sectors as we want to place sensors.

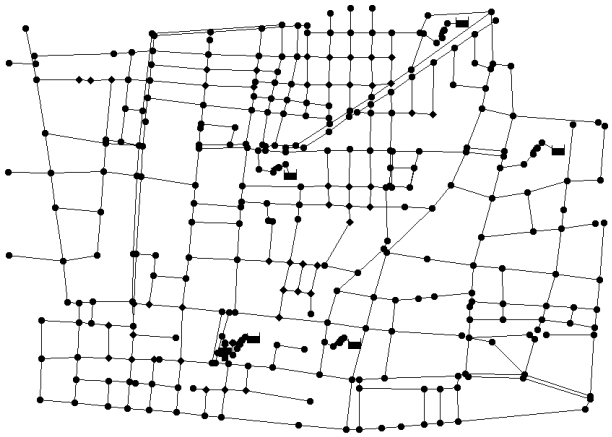

Figure 1: Layout of the WSN of the Central District of Celaya (Gto., Mexico).

Figure 2 (right side) shows a first division of the network using the spectral clustering algorithm. In this first iteration of the algorithm, we have added the restriction that at least have a supply point for each of the hydraulic areas, thus becoming, in this case, the process in a semi-supervised algorithm (Herrera [2011]).

Figure 2 (left side) shows the initial configuration of the relative importance of the nodes

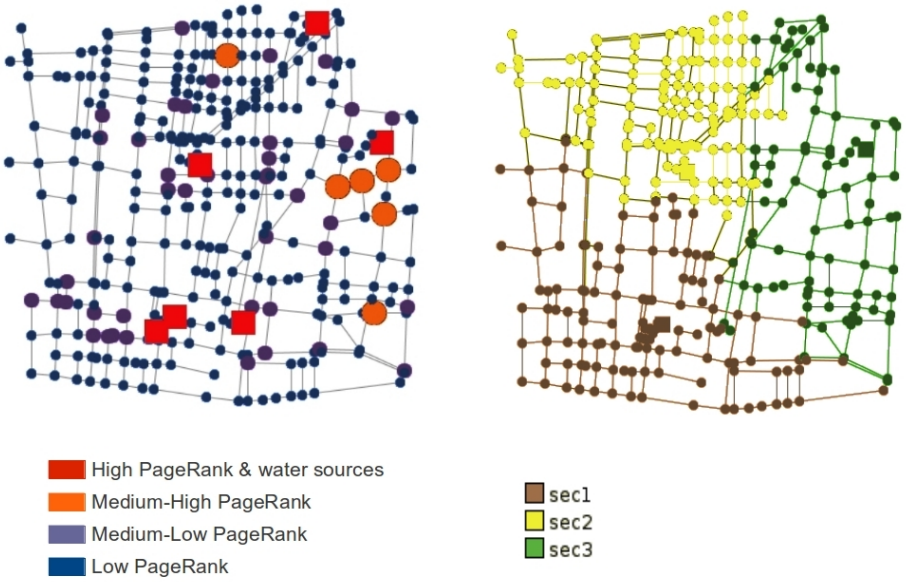

Figure 2: Spectral sectorisation and PageRank distribution in the case-study.

according to their PageRank. The red nodes are the most important, the colour orange is PageRank medium-high and the violet and blue values correspond to medium-low and low, respectively.

Figure 3 shows the final location of the 5 sensors proposed by this method based on hydraulic zones. Based on the index gh, the first subdivision in sub-areas (starting from Figure 2 -right side- configuration) would be on the Sector 1 which introduced the first step of the process. Subsequently, the process updates the values of this index and returns to the decision to subdivide. The second iteration of the process starts on the initial Sector 1. The recalculated gh index for each node supports the decision to place a sensor in the neighbourhood of these nodes of greater relative importance (highest value of gh). In this way, 5 hydraulic areas are originated, where we will place a sensor in each one. These nodes are: $N73$ (as in Figure 3 lies in the marked area approximately by nodes brown), $N101$ (area marked by green nodes - Figure 3), $N151$ (zone with nodes orange - Figure 3), $N280$ (an area with yellow nodes - Figure 3) and $N314$ (area with blue nodes - Figure 3).

Spectral clustering and PageRank algorithms are implemented in the packages *specc* (Karatzoglou [2006]) and *igraph* (Csárdi and Nepusz [2006]) of R Language (R Development Core Team [2011]); from which we import data from EPANET (Rossman [2000]). It is interesting point out that EPANET continue working, by its input/output with R relationship (Herrera [2011]). Thus, we can simulate contamination events in EPANET and continue analysing our data and/or validating our experiments from an hydraulic point of view.

4 Conclusions

This paper proposes an iterative WSN division into DMAs for sensor location and establishes inferences about contamination events based on signals from these sensors. In contrast to other existing methods that specifically approach the sensor location problem, our proposal address it by approaching divisions and subdivisions into hydraulic areas.

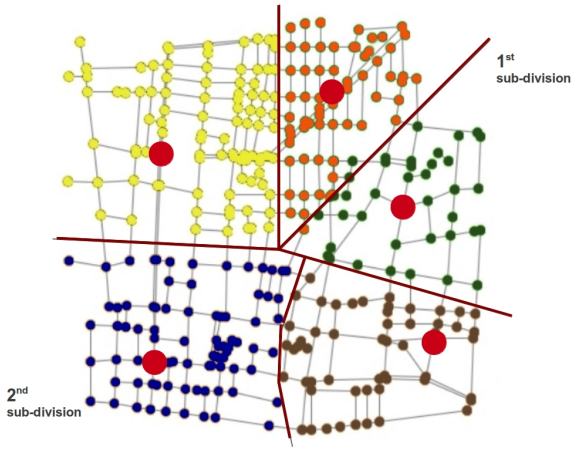

Figure 3: Final location of 5 sensors in the WSN of our case-study.

The proposed method is scalable and can be generalised. In addition, our aim also is mitigating the consequences of contamination by slowing or stopping the contaminated water in their spreading to not affected areas.

As future work we plan to formalise the concepts introduced here, working in order to demand a minimum number of sensors by losing a minimum detection sensitivity of contaminant agents. Another challenge to address is to apply our proposal to already studied WSN (such as the introduced in Ostfeld and *et al.* [2008]), in order to compare the results of our method respect to other alternatives already developed.

ACKNOWLEDGMENTS

This work has been performed under the support of the project IDAWAS, DPI2009-11591 of the Dirección General de Investigación del Ministerio de Ciencia e Innovación (Spain) and ACOMP/2011/188 of the Conselleria de Educació de la Generalitat Valenciana.

REFERENCES

Berry, J., L. Fleisher, W. Hart, C. A. Phillips, and J. Watson. Sensor placement in municipal water networks. *Journal of Water Resource Planning and Management*, 131(3): 237–243, 2005.

Brin, S. and L. Page. The anatomy of a large-scale hypertextual web search engine. *http://infolab.stanford.edu/ backrub/google.html*, 1998.

Chung, F. and W. Zhao. Pagerank and random walks on graphs. In *Proceedings Fete of Combinatorics and Computer Science Conference in honours of Laci Lovász*, 2008.

Csárdi, G. and T. Nepusz. igraph reference manual. *http://igraph.sourceforge.net/doc/html/index.html*, 2006.

Drezner, Z. and H. Hamacher. *Facility location: Applications and theory*. Springer, Berlin, 2002.

Grubesic, T., T. Matisziw, A. Murray, and D. Snediker. Comparative approaches for assessing network vulnerability. *International Regional Science Review*, 31(1):88–112, 2008.

Gueli, R. Predator-prey model for discrete sensor placement. In *Proceedings of 8th Annual Water Distribution System Annalysis Symp.*, pages 591–595, 2006.

Gutiérrez-Pérez, J. A., M. Herrera, R. Pérez-García, and E. Ramos-Martínez. Application of graph-spectral methods in the vulnerability assessment of water supply networks. *Mathematical Computing and Modelling*, in press, 2012.

Herrera, M. *Improving water network management by efficient division into supply clusters*. PhD thesis, 2011.

Herrera, M., S. Canu, A. Karatzoglou, R. Pérez-García, and J. Izquierdo. An approach to water supply clusters by semi-supervised learning. In *International Congress on Environmental Modelling and Software*, 2010.

Herrera, M., J. A. Gutiérrez-Pérez, J. Izquierdo, and R. Pérez-García. Ajustes en el modelo pagerank de google para el estudio de la importancia relativa de los nodos de la red de abastecimiento. In *X Seminario Iberoamericano de planificación, proyecto y operación de sistemas de abastecimiento de agua*, 2011.

Karatzoglou, A. *Kernel methods software, algorithms and applications*. PhD thesis, 2006.

Ng, A. Y., M. I. Jordan, and Y. Weiss. On spectral clustering: Analysis and an algorithm. In *Advances in Neural Information Processing Systems 14*, pages 849–856, 2001.

Ostfeld, A. and *et al.* The battle of the water sensor networks (bswn): A design challenge for engineers and algorithms. *Journal of Water Resource Planning and Management*, 134:556–568, 2008.

Ostfeld, A. and E. Salomons. Optimal layout of early warning detection stations for water distribution systems security. *Journal of Water Resource Planning and Management*, 130(5):377–385, 2004.

Propato, M. Contamination warning in water networks: General mixed-integer linear models for sensor location design. *Journal of Water Resource Planning and Management*, 132(4):225–233, 2006.

R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2011. ISBN 3-900051-07-0.

Rossman, L. *EPANET-User's Manual*. United States Environmental Protection Agency (EPA), 2000.

Sevkli, M. and A. Guner. *Lecture Notes in Computer Science 4150*, chapter Optimisation algorithm for uncapacited facility location problem, pages 316–323. Dorigo et al. Eds., 2006.

Shen, Z., R. Zhan, and J. Zhang. The reliable facility location problem: Formulations, heuristics, and approximation algorithms. *Journal in Computing*, 23(3):470–482, 2011.

Yazdani, A. and P. Jeffrey. A complex network approach to robustness and vulnerability of spatially organized water distribution networks. *e-print: http://arxiv.org/abs/1008.1770v2*, 2010.