Abstract: For large water-distribution systems fully detailed models result in a substantial amount of data, making it difficult to manage, monitor, and understand how the main structure of the system works. A possible way to cope with this difficulty is to gain insight to the system behavior by simplifying its operation through topological/connectivity analysis. The objective of this study is to develop and demonstrate a generic topological-based scheme to aid in the analysis of water-distribution systems. The methodology relies on clustering, which divides the distribution system into strongly and weakly connected sub-graphs using the depth first search (DFS) and breadth first search (BFS) graph algorithms. The partitioning results in a connectivity matrix that represents the interconnections between clusters, which can support, for example, a response modeling plan in case of a contamination intrusion incident. A detailed illustrative example and a real complex water-distribution system are explored for demonstrating the developed model capabilities. Possible applications of the proposed algorithm are suggested. DOI: 10.1061/(ASCE)WR.1943-5452.0000173. © 2012 American Society of Civil Engineers.

CE Database subject headings: Water distribution systems; Algorithms.

Author keywords: Model; Simplification; Connectivity; Clustering; Analysis; Water-distribution systems; Graph theory.

Introduction

A water-distribution system is a collection of hydraulic control elements connected together to convey quantities of water from sources to consumers. A distribution system is typically subjected to dynamic loading conditions and controls, making systems properties variable with time. For large systems containing hundreds to thousands of nodes and links, it is difficult to control the structure of the system and the interactions of its components. Existing hydraulic simulation models are capable of providing a very detailed model with high accuracy but not an improved understanding of the main structure of the underlying system. Such systems inherently raise challenges in managing, monitoring, and understanding their behavior.

Management models for water-distribution systems can be classified into three categories (Ostfeld 2005): (1) layout (analysis of system connectivity/topology); (2) design (system sizing given a layout); and (3) operation (system operation given a design), with reliability/risk considerations related to all three. This work is associated with category (1). Its goal is to develop a generic tool of network simplification for visualization and data processing through subnetworks. The developed algorithm divides the system into clusters for a given time duration according to the flow directions in pipes. The resulted clusters create a connectivity relationship that further aggregates and simplifies the system interconnections. An illustrative example and a real complex water-distribution system are explored for demonstrating the developed methodology capabilities.

This work is aimed at introducing a generic connectivity-based analysis tool that can be further explored for different applications, such as those highlighted in the Possible Applications section. This is somewhat different than typical studies because of it provides an open model for discovering the underlying structure of a system, not predicting anything specific, rather than a closed model for a specific target. This study extends Perelman and Ostfeld (2011).

Literature Review

Clustering is the process of partitioning a set of objects into subsets of similar properties. The objects are alike in a given cluster and different with respect to objects in other clusters. Clustering in general refers to unsupervised learning and statistical data analysis, including machine learning, data mining, pattern recognition, image analysis, and bioinformatics. Examples of clustering include groups of independent computer servers interconnected through a dedicated network to work as one centralized data processing resource, in marketing—finding collection of customers with similar behavior, in Biology—classification of plants and animals with similar features, on the Internet—gathering web log data to assemble groups of similar access patterns, etc. A survey of graph clustering methods and applications can be found in Schaeffer (2007).

For water-distribution systems there has been limited work to date on topological/connectivity analysis, with most of the literature concentrating on reliability or aggregation/skeletonization related models. Following is a brief review.

Wagner et al. (1988) applied analytical methods using the algorithms of Satyanarayana and Wood (1982) and Rosenthal (1977) for computing: (1) Connectivity the probability that a given demand node is connected to a source; and (2) Reachability the probability that all demand nodes are connected to a source. Hamberg and Shamir (1988) presented a simplification methodology for the design of water-distribution systems. The simplification was developed.
through using a step-wise combination of system elements and through a nonlinear continuum representation of the system as a bundle. Anderson and Al-Jamal (1995) introduced a methodology for the simplification of complicated hydraulic networks using a parameter-fitting approach. The layout of the final simplified network is specified a priori; then, pipe conductances and demand distribution are determined by maximizing the fitness between the simplified and the full system performances. Ulanicki et al. (1996) presented a mathematical model for the hydraulic aggregation of water-distribution systems. The approach computes a network model equivalent to the original system with fewer components by analytical elimination of system components. The reduced nonlinear model preserves the non-linearity of the original model and approximates the original system within a wide range of operating conditions. Yang et al. (1996) utilized a minimum cut set method to evaluate the impact of link failures on source-demand connectivity. Grayman and Rhee (2000) utilized a skeletonization process for representing a water-distribution system by selected pipes. Davidson et al. (2005) used supervisory control and data acquisition (SCADA) data to create connectivity matrices for encapsulating the worst-case scenarios of possible contamination spreads. Ostfeld (2005) described a methodology for establishing the most flexible pair of operational and backup digraphs (Ostfeld and Shamir 1996) of a water-distribution system that maintains Kirchoff’s Laws 1 and 2, and yields (if possible) a one-level system redundancy. Perelman and Ostfeld et al. (2008) extended the aggregation methodology of Ulanicki et al. (1996) to water quality analysis. Deuerlein (2008) introduced a decomposition method of a network graph according to its connectivity properties. The model decomposes the network into basic small components, allowing the solution of the flow and head equations of each component separately and combining all of them to receive the full hydraulic system solution. Grayman et al. (2009) presented rehabilitation of supply systems based on district metered areas (DMAs). The network topology was manually redesigned by adding, deleting, and modifying existing pipes and by opening and closing valves to attain a system having, in most cases, a single inflow and no outflow to the transmission system or to other zones. Xu et al. (2010) proposed a methodology that divides the system into subareas using a facility location model that is further linked to a Bayesian network.

There are probably many ways to partition a water-distribution system (WDS) to clusters, primarily depending on the objective and the selected clustering algorithm. These works represent the system using undirected or directed graphs with applications to small examples. Recently, the importance for clustering in a distribution system for its design, operation, and security has been noticed and started to be investigated (Grayman et al. 2009). However, no formal definition of clusters or groups in distribution systems is yet to exist. This work attempts to partition system nodes to clusters considering the dynamic underlying conditions of the system (i.e., system flow patterns). The developed network simplification methodology through clustering allows qualitative formulation of the system graph. The method outcome facilitates in the formulation of simplified network graph of the system for the ease of visualization, analysis, and application of appropriate solution techniques.

### Problem Formulation

A cluster herein is thought to be a set of nodes that have more and/or better connections between its inner nodes than to the remaining nodes of the system. The metric for grouping nodes is based on different levels of connectivity between pairs of nodes—strongly or weakly connected (Fig. 1). The objective is to partition the network into strongly and weakly connected clusters. The depth first search (DFS) (Tarjan 1972) and breadth first search (BFS) (Pohl 1969) graph algorithms are used to compute graph connectivity and clustering.

Two public domain tools are employed: a hydraulic analysis model for exploring flow directions in pipes [EPANET (Environmental Protection Agency 2012)] and a geographical information system (GIS) for clusters structure presentation [MapWindow-GIS 2012].

The clustering algorithm is demonstrated in detail on a small illustrative example and on the water supply system of Dover Township area, New Jersey (Maslia et al. 2000, 2001).

### Methodology

The proposed methodology relies on basic notions of graph theory, which are briefly outlined bellow.

### General Graph Notions

#### Connected Graph

A water supply system can be represented as a graph \( G (V, E) \) using graph theory. Consumers and source nodes are mapped as graph vertices/nodes \( V \) and the connecting links (i.e., pipes, pumping units, and valves) as the edges/links \( E \). The undirected graph is assumed to be connected if it contains a path between every pair of nodes.

#### Problem Formulation

A cluster herein is thought to be a set of nodes that have more and/or better connections between its inner nodes than to the remaining nodes of the system. The metric for grouping nodes is based on different levels of connectivity between pairs of nodes—strongly or weakly connected (Fig. 1). The objective is to partition the network into strongly and weakly connected clusters. The depth first search (DFS) (Tarjan 1972) and breadth first search (BFS) (Pohl 1969) graph algorithms are used to compute graph connectivity and clustering.

Two public domain tools are employed: a hydraulic analysis model for exploring flow directions in pipes [EPANET (Environmental Protection Agency 2012)] and a geographical information system (GIS) for clusters structure presentation [MapWindow-GIS 2012].

The clustering algorithm is demonstrated in detail on a small illustrative example and on the water supply system of Dover Township area, New Jersey (Maslia et al. 2000, 2001).

### Methodology

The proposed methodology relies on basic notions of graph theory, which are briefly outlined bellow.

#### General Graph Notions

#### Connected Graph

A water supply system can be represented as a graph \( G (V, E) \) using graph theory. Consumers and source nodes are mapped as graph vertices/nodes \( V \) and the connecting links (i.e., pipes, pumping units, and valves) as the edges/links \( E \). The undirected graph is assumed to be connected if it contains a path between every pair of nodes.
distinct nodes. For the directed graph, all possible edge directions are established by performing extended period simulation of network hydraulics and setting edge direction as the direction of flow in each link. The configuration of the undirected graph is fixed, whereas the configuration of the directed graph is time dependent, because flow direction in links may alter. Intuitively, the undirected graph represents network topology and the directed graph interactions between network nodes.

**Strongly Connected Component**
A strongly connected component is a set of nodes in which at least one directed path exists between every pair of nodes $u$, $v$ of that component, i.e., $u \rightarrow v$ and $v \rightarrow u$. A strongly connected component is a directed cyclic component (i.e., flow direction in links belonging to that component can reverse). Fig. 1(a) demonstrates four nodes that are strongly connected. It can be seen that all pairs of nodes are connected by at least one directed path.

**Weakly Connected Component**
A directed subgraph is weakly connected if its underlying undirected graph is connected (i.e., it contains a path between every two nodes $u$ and $v$). A weakly connected component can comprise acyclic looped subgraphs and/or tree components. In both cases, flow direction in links is constant and cannot reverse. Fig. 1(b) shows four nodes that are weakly connected. Because flow in all links is constant and cannot reverse, not all nodes are connected. For example, node C is not connected to nodes A, B, and D.

**Depth First Search**
The depth first search (DFS) (Tarjan 1972) algorithm is used to explore the connectivity of a graph. The algorithm starts at some node, traversing in the direction of the outgoing edges as far as possible before retrieving. At each step the procedure continues by choosing an unexplored edge of the recently reached node. When no more unexplored edges are left, the algorithm tracks back to a previously reached node that still has unexplored edges. The search completes when all nodes of the graph have been visited. The DFS time complexity is linearly proportional to the size of the graph $O(|V| + |E|)$. Some of the DFS applications are for testing connectivity of a graph, finding a path between two nodes of a graph, computing a cycle in a graph, etc.

**Breadth First Search**
Breadth first search (BFS) (Pohl 1969) is a different type of search algorithm that is used herein to find all nodes within one weakly connected component. It begins at a root node, and explores all of the adjoining nodes, and then for each of those nearest nodes it explores all their unexplored adjoining nodes, continuing until there are no more adjacent unvisited nodes. Time complexity is again linearly proportional to the size of the graph $O(|V| + |E|)$. The BFS does not visit all nodes of the graph, only those reachable from the starting node.

**Clustering Basic Algorithm**
The clustering algorithm attempts to group network nodes based on their connectivity as to attain a simplified representation of the distribution system. The procedure is based on weakly and strongly definitions of nodal connectivity, utilizing two graph search algorithms, all previously described, capable of identifying those types of connections.

Because a real water-distribution system is typically subjected to varying loading and operational conditions, its hydraulic behavior is dynamically changing over time. Primarily, flow directions alternate over time, which affects cluster formation. Consequently, first the starting time and duration of the clustering procedure needs to be defined before the algorithm can be invoked. The proposed methodology comprises the following stages.

**Water-Distribution Systems Mapping**
The distribution system is mapped to a graph in which the nodes represent the consumers, sources, and tanks and the edges of the connecting pipes, pumps, and valves. The interaction between network elements is described solely by the direction of flow in the network links during the specified time period.

**Strongly Connected Clusters**
In this stage, the breadth first search (BFS) algorithm is utilized. The BFS begins at a root node and finds all accessible nodes that comprise a single weakly connected component. Consequently, for each root node a new weakly connected component is identified. A weakly connected cluster (WCC) is defined as the maximal weakly connected subgraph (i.e., all accessible nodes from the root node). The BFS is executed several times, each time beginning at a different starting point, reservoirs, tanks, source nodes, and strongly connected cluster boundary nodes, until all remaining nodes of the network are explored.

The identification of a weakly connected cluster is not unique opposed to a strongly connected cluster. This is attributable to two main reasons: (1) the order in which BFS is executed influences the defined set of nodes reachable from the root node, and (2) a weakly connected set of nodes can be formulated based on different criteria, for example, setting upper and lower bounds on the size of each set, minimizing the maximum number of levels of connectivity of the sets of nodes, etc. Point (1) is addressed below where the algorithm is further extended; point (2) is dealt with by defining a weakly connected cluster as the maximal weakly connected subgraph for a given root node (i.e., all weakly connected nodes that can be reached from a given root node).

**Cluster Structure Formalization**
On the completion of the previous two steps, all nodes of the network are identified as strongly or weakly connected and grouped into independent strongly connected or weakly connected clusters. The clusters are connected by edges in which flow direction is, by definition, known and constant (i.e., because if flow direction could reverse than the link would have been part of a strongly connected cluster). As a result, a new network topology is formulated, describing the clusters and their connecting links. Furthermore, a connectivity matrix can be formed representing a new system topology and the connections between its clusters.
**Illustrative Example**

The illustrative example incorporates a detailed description of the proposed methodology using a slightly modified version of EPA-NET Example 1. The system is subject to a variable demand pattern of 24 h and consists of 9 nodes, 11 pipes, 1 source, 1 storage tank, and 1 pumping unit. Fig. 2 shows the system layout and each of the steps of the proposed method. Those are:

1. **Water-distribution systems mapping**—the system is mapped to a directed graph based on 24 h simulation results. Fig. 2(b) shows the resulting graph and all the edge directions.
2. **Strongly connected clusters**—the DFS is executed once identifying all strongly connected nodes and grouping them into unique strongly connected clusters. All the strongly connected nodes connected by links in which flow can reverse are circled in Fig. 2(b). Fig. 2(c) shows all the strongly connected nodes grouped into two strongly connected clusters: SC1 (i.e., strongly connected cluster 1) and SC2 (i.e., strongly connected cluster 2).
3. **Weakly connected clusters**—the remaining nodes are weakly connected and can be partitioned into weakly connected clusters. As outlined above, the BFS starts at several root nodes until all weakly connected nodes are explored. For example Fig. 2(d) shows the construction of weakly connected cluster 1 (WC1, nodes 1 and 10) starting the BFS at the source (node 10); Fig. 2(e) shows WC2 (nodes 4 and 7) starting the BFS at

---

**Fig. 2.** Illustrative example layout and clustering partition

---

**Fig. 3.** Extended clustering: illustrative example layout and clustering partition
node 3 (a boundary node of SC1); and finally Fig. 2(f) shows WC3 (nodes 8 and 9) starting the BFS at node 5 (a boundary node of SC2).

4. **Cluster structure formalization**—once all nodes of the system are explored and classified into strongly connected and weakly connected clusters, the connections between the clusters are established based on the original mapping [Fig. 2(b)]. Fig. 2(f) shows the final cluster structure of the system.

**Extended Clustering Algorithm**

The proposed clustering algorithm was further extended to improve and generalize its outcome. By applying the clustering model described in the previous section all system nodes are classified into strongly or weakly connected clusters. However, the procedure may result in a substantial number of varying size clusters in the case of large water-distribution systems. In such cases, some clusters can be very small with even only two nodes. For example, small number of pendant nodes (branched dead end) without sources can form an independent cluster or two strongly connected nodes form a corresponding strongly connected cluster. It is thus beneficial for small clusters to be merged with other clusters to form a coarser overall system partition. This can be viewed as a multilayered clustering process, in which each top level connected clusters subclusters.

The clustering method is thus further developed to attain a coarser clustering by merging and coupling subclusters. Assuming

```plaintext
Water distribution systems (WDSs) clustering algorithm

[ NodeList ] = WDS_CL ( WDS, T, ClMinSize )

INPUT: 1) WDS 2) T - time period 3) ClMinSize - cluster minimum size

OUTPUT: A list of nodes classified to clusters NodeList (node) = {c₁, ..., cₙ}

/* STEP 1 - network graph */
map WDS to a digraph DG (V,E) for a given T

for each ( node ∈ DG ) do
  NodeList (node) = unclassified
  /* mark node as unclassified */

/* STEP 2 - identify SC components */
[ SC ] = DFS ( NodeList )

for each ( node ∈ DG ) do
  if ( node ∈ SCᵢ )
    NodeList (node) = SCᵢ
    /* classify node as SCᵢ */

/* STEP 3 - identify WC components */

for each ( root node ) do
  /* root node = source + SCᵢ boundary node */

  [ WCᵢ ] = BFS ( NodeList, root node )

  for each ( node ∈ WCᵢ ) do
    NodeList (node) = WCᵢ
    /* classify node as WCᵢ */

/* STEP 4 - expand clusters */
/* Expand SC components */

for each ( SC ) do
  index = index + 1
  /* a new cluster index */

  for each ( node ∈ SCᵢ ) do
    NodeList (node) = cᵢ
    /* assign node to its cluster cᵢ */

  for each ( WC adjacent only to SCᵢ ) do
    /* add branched WC */

    if ( Size(WCᵢ) < ClMinSize )
      for each ( node ∈ WCᵢ ) do
        NodeList (node) = cᵢ
        /* assign node to cᵢ */

/* Expand WC components */

for each ( WC ) do
  index = index + 1

  for each ( node ∈ WCᵢ ) do
    NodeList (node) = cᵢ
    /* assign node to cᵢ */

  for each ( WCᵢ overlapping WCᵢ )
    /* add overlapping WC */

    for each ( node ∈ WCᵢ ) do
      NodeList (node) = cᵢ
      /* assign node to cᵢ */

/* Append pathway nodes */

for each ( clₙ ) do
  for each ( boundary node ∈ clₙ ) do
    /* find a forest path */

    [ WC ] = BFS ( NodeList, boundary node )

    for each ( node ∈ WC ) do
      if ( node ̸∈ CL )
        NodeList (node) = clₙ
        /* assign node to clₙ */

Fig. 4. Pseudocode of the proposed clustering algorithm
```
a threshold number for a group of nodes to be considered a cluster exists, clusters having a smaller number of nodes will be merged. The occasions in which smaller clusters are assembled are described next.

**Expand a Strongly Connected Cluster by Adding Branches**

The structure of water supply systems comprises of looped and branched subnetworks. A branched subgraph is by definition a weakly connected cluster because the directions of flows in a branched network are constant and do not change in time. If such a branch is connected to a strongly connected cluster then it forms an independent new weakly connected cluster. Such a cluster is called an extended mixed connected cluster (EMCC).

**Couple Weakly Connected Clusters**

Adjacent weakly connected clusters having overlapping nodes can be merged. During the identification of weakly connected clusters a BFS is restarted each time, beginning at a different root node. It is possible that the same group of nodes is reachable from different root nodes and, as previously mentioned, the order of BFS can influence the identification of weakly connected nodes. Hence, instead of assigning the explored nodes to different weakly connected clusters they can be joined together into a single cluster.

**Append Pathway Nodes**

All nodes located on a path between previously identified clusters can be added to their original cluster. Once identifying and extending all strongly connected and weakly connected clusters, the connections between them need to be established. To find a path between any two clusters the BFS is again utilized each time starting from a different boundary node of each cluster. Assuming a path connecting two clusters exists, a connection can be established between them and the rest of the nodes that are not assigned to any cluster can be added to the upward cluster.

**Illustrative Example (Continued)**

The stages of expand a strongly connected cluster through adding branches and couple weakly connected clusters are demonstrated in Fig. 3.

The first three steps of the algorithm are carried out as before [Figs. 2(b)–2(d)]. Next, the BFS starts from root node 3 identifying nodes 4 and 7 as a single weakly connected cluster [i.e., WC2, Fig. 2(e)]. However, if the BFS order was different, starting from node 6 first and then from node 3, then two different weakly connected clusters would have been identified: one with only node 7 and one with only node 4 [i.e., WC2 and WC3, respectively, Fig. 3(a)]. The rest of the procedure remains the same. The new resulting structure comprises six clusters: SC1, SC2, WC1–WC4 [Fig. 3(a)].

The objective is to obtain a coarser partition of clusters. Fig. 3(b) demonstrates the expansion of SC2 into an extended mixed connected cluster (EMCC). Clusters SC2 and WC4 have a hierarchical structure and can be merged because WC4 is a branched cluster connected only to SC2 and having no other entries except from SC2 itself. Hence, SC2 and WC4 are assembled to form an EMCC with nodes 5, 6, 8, and 9. Fig. 3(c) shows the merging of two weakly connected clusters, WC2 and WC3. Because the two clusters are small and overlapping they can be joined together to a single cluster containing nodes 4 and 7 as in the previous clustering formation [Fig. 2(e)]. The resulted representation [Fig. 3(c)] is less detailed [compared with Fig. 2(f)], containing fewer clusters and connecting links. This extension is beneficial in the case of large networks and is demonstrated on the Dover Township system, New Jersey, in the Example Applications section following.

Fig. 4 formally summarizes in a pseudocode description of the proposed clustering algorithm.

![Fig. 5. Example application: Dover Township water system, New Jersey (adapted from Maslia et al. 2000, ©ASCE)](image-url)
Example Application

In this section the methodology is demonstrated on the Dover Township system, New Jersey (Maslia et al. 2000, 2001). The system is subject to a variable demand pattern of 24 h, consists of 16,056 links, 14,945 junctions, 8 ground-level and elevated storage tanks, 8 reservoirs, 12 high-lift or booster pumps, and 20 water-supply wells. Its layout is given in Fig. 5.

Figs. 6–8 describe the cluster partitioning of Dover network for clustering time duration of 5 h starting at midnight. A component having at least 100 nodes was defined as a cluster, any component smaller than that was merged into another cluster. Fig. 6 presents the resulted extended mixed connected clusters (EMCCs) partition, Fig. 7 illustrates an example of an extended mixed connected cluster, and Fig. 8 shows the resulting weakly connected clusters division. Once all the clusters are defined, their connections are constructed based on the simulation mapping (i.e., stage 1 of the methodology). Fig. 9 demonstrates the boundary nodes and edges between clusters W26, S20, and S129. Next, cluster connectivity matrix is formed in which each cluster is represented as a node. The number of rows and columns of the clusters connectivity matrix are equal to the number of clusters, where a “1” entry describes a direct connection between two clusters.
connection between the row (upstream) and the column (downstream) clusters, and a "0" entry, otherwise. Table 1 shows the clusters connectivity matrix for this example. The clusters connectivity matrix can enhance, for example, a response modeling approach in case of a contamination intrusion. This and other possible utilizations of the proposed clustering algorithm are further detailed in the possible applications section. Running time on a Lenovo laptop T61, T7300@2.00 GHz, 2GB of RAM took 4.8 min to form the clusters.

Fig. 10 presents the distribution of the number of nodes [Fig. 10(a)], and clusters [Fig. 10(b)], as a function of the clustering time duration. For small clustering time durations, as expected, the majority of the nodes are weakly connected and only a few are strongly connected (i.e., there are a few small areas in which nodes are fully interconnected). As the clustering time duration increases, new dynamic loading conditions impose flow directions in links to reverse, thus causing existing strongly connected areas to grow and new ones to form. The total number of strongly connected nodes thus increases, and, correspondingly, the number of weakly connected nodes decreases. The number of strongly connected and weakly connected clusters gradually grows with time until the clusters are large enough and can be merged, causing the number of clusters to decrease. It can be seen from Fig. 10 that at approximately 11:00 h the number of the identified clusters and the categorized nodes stabilizes as no new flow patterns are added.
| s20 | s34 | s45 | s52 | s102 | s129 | w7  | w9  | w26 | w97  | w126 | w127 | w161 | w167 | w202 | w203 | w205 | w238 | w271 | w285 | w308 | w320 | w344 | w385 | w395 | w407 | w426 | w442 | w501 |
|-----|-----|-----|-----|-----|------|-----|-----|-----|-----|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|
| 1   | 1   | 1   | 1   | 1   | 1    | 1   | 1   | 1   | 1   | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    |
| 1   | 1   | 1   | 1   | 1   | 1    | 1   | 1   | 1   | 1   | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    |
| 1   | 1   | 1   | 1   | 1   | 1    | 1   | 1   | 1   | 1   | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    |
| 1   | 1   | 1   | 1   | 1   | 1    | 1   | 1   | 1   | 1   | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    |
| 1   | 1   | 1   | 1   | 1   | 1    | 1   | 1   | 1   | 1   | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    |
| 1   | 1   | 1   | 1   | 1   | 1    | 1   | 1   | 1   | 1   | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    |
| 1   | 1   | 1   | 1   | 1   | 1    | 1   | 1   | 1   | 1   | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    |
| 1   | 1   | 1   | 1   | 1   | 1    | 1   | 1   | 1   | 1   | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    |
| 1   | 1   | 1   | 1   | 1   | 1    | 1   | 1   | 1   | 1   | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    |
| 1   | 1   | 1   | 1   | 1   | 1    | 1   | 1   | 1   | 1   | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    |
| 1   | 1   | 1   | 1   | 1   | 1    | 1   | 1   | 1   | 1   | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    |
| 1   | 1   | 1   | 1   | 1   | 1    | 1   | 1   | 1   | 1   | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    |

Note: IDs, for example, are S20 = extended mixed connected cluster (EMCC) 20, W7 = weakly connected cluster 7.
Weakly connected clusters acquired to fully deploy clustering. For some of the problems, the utilization of the proposed algorithm can be beneficial are described in this section.

Possible water-distribution systems analysis problems for which the heuristics of Grayman et al. (2009) can be generalized using the cluster graph can provide the graphical representation (revised from the sensor network design) and all injections in clusters SC3 and WC4 will not be detected by any of the sensors. This corresponds to the results presented by Xu et al. (2010). Area 1 comprises cluster WC3, Area 2 comprises WC5 and SC1, Area 3 comprises WC1 and SC2, and Area 4 comprises SC3 and WC4.

Comparison to Xu et al. (2010) Xu et al. (2010) used a maximum covering model for optimal sensor locations. The model results in location of the sensors and division of the network into areas based on the sensor coverage. The maximum covering model was applied to BWSN1 (i.e., example 1 of the Battle of the Water Sensor Networks, Ostfeld et al. 2008).

As a result, two sensors were placed in the network as shown in Fig. 11, and four areas were identified: Area 1 if a contamination can be identified by Sensor 1 only, Area 2 if contamination identified by Sensor 2 only, Area 3 if contamination identified by both sensors, and Area 4 if contamination identified by neither of the sensors.

From the network cluster structure of the proposed methodology (Fig. 11) the same behavior can be recognized. Sensor 1 located in cluster WC3 can identify a contaminant injected in clusters WC1, SC2, and WC3. Consequently, Sensor 2 located in cluster WC2 can identify a contaminant injected in clusters WC1, SC2, WC5, SC1, and WC2. Injections in clusters SC3 and WC4 will not be detected by any of the sensors. This corresponds to the results presented by Xu et al. (2010). Area 1 comprises cluster WC3, Area 2 comprises WC5 and SC1, Area 3 comprises WC1 and SC2, and Area 4 comprises SC3 and WC4.

This comparison demonstrates a possible utilization of the suggested cluster algorithm for sensor placement and for contamination source detection.

Contamination Source Detection
In this work clusters were detected based on connectivity changes of a system with time, hence the clustering framework can be linked with Bayesian network (BN) for addressing the uncertainty of contaminant source detection and its propagation in the system. The cluster graph can provide the graphical representation (required for the BN) of the conditional probabilities of a set of random variables, then the probabilities of the contaminant source and its propagation can be estimated based on available monitoring data.

Response Modeling
Once a contaminant has been detected by a sensor, the entire zone (cluster) in which the contaminant was discovered can be isolated by closing pipes. The clustering algorithm can be run backward in time for a time duration equal to the time elapsed since the suspected injection time (estimated from the sensor network design) and all upstream clusters can be sequentially blocked. Although eventually most of the system will be closed, the clustering algorithm provides guidance for response modeling in prioritizing which clusters should be first isolated. For example, if a contaminant was detected at 5:00 a.m. at the Dover system (Fig. 5) and the suspected injection time was midnight, then perform clustering from midnight to 5 a.m.

Fig. 10. Example application: (a) distribution of the number of nodes; and (b) clusters as a function of the clustering time duration

Possible Applications
The primary factor that should be considered in clustering and other approximations is the development of a model that is appropriate for its particular applications. What may be 5 an acceptable approximation for one type of application may not be acceptable for a different application. The clustering model presented in this manuscript is an open model, meaning that is appropriate for its particular application. Its basic purpose is to adjust system representation for solution techniques in other applications. This is typically an issue where large distribution systems are concerned. Possible water-distribution systems analysis problems for which the proposed algorithm can be beneficial are described in this section. For some of the problems, the utilization of the clustering method is straightforward; for others, additional research is required to fully deploy clustering.

District Meters Partitioning
Grayman et al. (2009) proposed rehabilitation of a water-distribution system layout to district metering areas (DMAs) to track and control water leakage and to improve water security. Network topology was manually redesigned and evaluated in terms of: (1) fire flow, (2) water age, (3) water security, and (4) reliability.

The heuristics of Grayman et al. (2009) can be generalized using the proposed clustering algorithm with adequate adjustments. The framework for clustering formations can be coupled with additional regulations to minimize the number of connections between clusters with the desire to attain optimal DMAs partition.

Ostfeld and Salomons (2004) presented a methodology for the optimal layout of a detection system comprising two main stages: 1) generating randomized pollution matrix (RPM), in which the j-th row represents the detection domain of a sensor located at node i and the j-th column—the pollution domain of a contaminant injected at node j; 2) finding the optimal sensor location that maximizes the RPM coverage.

As a water-distribution system becomes more complex in terms of its size, loading conditions, controls, etc., so is the difficulty of establishing a representative pollution matrix (i.e., a matrix of reduced size that provides similar results as a full event enumerated contamination matrix). A possible way to cope with this difficulty is to establish a cluster pollution matrix using sets of cluster partitions instead of systems nodes and maximize cluster pollution matrix coverage. The outcome of this process will be cluster locations for sensor placement (i.e., instead of nodes, as in Ostfeld and Salomons 2004). Then an optimization technique can be used (e.g., genetic algorithm) to place the sensors in a given cluster.
(see Figs. 6 and 8). If a sensor detected a contaminant at cluster w9 then there is a need to isolate only cluster w9 (see Table 1 and Fig. 8); if a sensor detected at w26, then w26 and s20 should be first closed, then s34, s129, w97, w127, and w161 (i.e., the column corresponding to s20 in Table 1), then w126, etc.

**Conclusions**

The topology of real water-distribution system is naturally composed of several subnetworks that are hydraulically connected or disconnected. Connectivity properties vary in time as a result of changes in dynamic loading conditions, making it difficult to supervise a systems behavior over time. This work is aimed at introducing a general topological analysis tool for the partition of a water-distribution system as a function of its structural and connectivity properties (i.e., topology and flow patterns) to groups. The model is not based on nodal hydraulic or quality characteristics (e.g., pressures and concentrations), thus it excuses the use of an exact and well-calibrated hydraulic and water quality model of the system. The clustering method provides an improved understanding of the main structure of the system and the connections between its components and can be further utilized for system design or security applications.

**Acknowledgments**

This research was supported by the Institute for Future Defense Technologies Research Named for The Medvedi, Shwartzman, and Gensler families; by the Technion Grand Water Research Institute (GWRI); and by NATO [Science for Peace (SfP) project no. CBD.MD.SFP 981456].

**References**


