



# Sensor placement for fault location identification in water networks: A minimum test cover approach<sup>☆</sup>



Lina Sela Perelman<sup>a</sup>, Waseem Abbas<sup>b</sup>, Xenofon Koutsoukos<sup>b</sup>, Saurabh Amin<sup>a</sup>

<sup>a</sup> Massachusetts Institute of Technology, United States

<sup>b</sup> Vanderbilt University, United States

## ARTICLE INFO

### Article history:

Received 12 May 2015

Received in revised form

21 March 2016

Accepted 22 May 2016

### Keywords:

Fault identification

Minimum test cover

Water networks

## ABSTRACT

This paper focuses on the optimal sensor placement problem for the identification of pipe failure locations in large-scale urban water systems. The problem involves selecting the minimum number of sensors such that every pipe failure can be uniquely localized. This problem can be viewed as a minimum test cover (MTC) problem, which is NP-hard. We consider two approaches to obtain approximate solutions to this problem. In the first approach, we transform the MTC problem to a minimum set cover (MSC) problem and use the greedy algorithm that exploits the submodularity property of the MSC problem to compute the solution to the MTC problem. In the second approach, we develop a new *augmented greedy* algorithm for solving the MTC problem. This approach does not require the transformation of the MTC to MSC. Our augmented greedy algorithm provides in a significant computational improvement while guaranteeing the same approximation ratio as the first approach. We propose several metrics to evaluate the performance of the sensor placement designs. Finally, we present detailed computational experiments for a number of real water distribution networks.

© 2016 Elsevier Ltd. All rights reserved.

## 1. Introduction

Infrastructure deterioration, demand-supply uncertainty, and risk of disruptions pose new challenges in maintaining modern infrastructures. Resilient urban infrastructures including water distribution systems, transportation networks, and electric grids are crucial for societal well-being. *Smart* infrastructure operation driven by sensing and actuation technologies have been identified as one of the primary solutions towards resilient urban systems (Pandharipande, Calabrese, Lim, & Rajagopal, 2014; Zheng & Kleiner, 2014). Through a network of sensors, an individual fault or correlated failures in a system component can be detected and

localized, and restorative actions can be executed in response to these faults. Whereas network observability for a given sensing capability has been widely studied in the context of fault detection, sensor placement for fault isolability, i.e. the ability to distinguish between faults, has not been a commonly studied problem, especially in the context of pipe bursts in water distribution networks.

The goal of this work is to *design a sensor placement configuration for identification of pipe failure locations by using the minimum number of sensors*. The underlying idea behind our approach is to ensure that the sensor placement results in a collective output that is *unique* for each failure event. Specifically, our main contributions are as follows, we:

- Define the *localization* of pipe bursts as the design objective of a sensor network configuration, and using ideas from combinatorial optimization, we formulate the fault location identification problem as a *minimum test cover* (MTC) problem;
- Develop a computationally efficient *augmented greedy* algorithm to solve the minimum test cover problem (resp. identification problem), which is significantly faster in comparison to the previous approaches and therefore, scalable to large-scale networks; and
- Test and evaluate our sensor placement approach on a batch of real-networks of various sizes and parameters using practically relevant performance measures.

<sup>☆</sup> This work was supported in part by FORCES (Foundations Of Resilient Cyber-Physical Systems), which receives support from the National Science Foundation (NSF award numbers CNS-1238959, CNS-1238962, CNS-1239054, CNS-1239166), the AFRL LABLET-Science of Secure and Resilient Cyber-Physical Systems (Contract ID: FA8750-14-2-0180, SUB 2784-018400), and the NSF CAREER award #1453126. The material in this paper was not presented at any conference. This paper was recommended for publication in revised form by Associate Editor Michele Basseville under the direction of Editor Torsten Söderström.

E-mail addresses: [linasela@mit.edu](mailto:linasela@mit.edu) (L. Sela Perelman), [waseem.abbas@vanderbilt.edu](mailto:waseem.abbas@vanderbilt.edu) (W. Abbas), [Xenofon.Koutsoukos@vanderbilt.edu](mailto:Xenofon.Koutsoukos@vanderbilt.edu) (X. Koutsoukos), [amins@mit.edu](mailto:amins@mit.edu) (S. Amin).

### Nomenclature

$C_i^t$	Set of pair-wise link failures detected by the sensor $i$
$C_i$	Set of link failures detected by the sensor $i$
$\mathcal{C}$	Collection of all $C_i$ 's
$\mathcal{C}^t$	Collection of all $C_i^t$ 's
$f_D$	Detection function
$f_I$	Identification function
$h$	Hydraulic head
$I_D$	Normalized detection score
$I_I$	Normalized identification score
$I_L$	Normalized localization score
$I_W$	Number of elements in the largest localization set
$k$	Maximum number of link failures detected by any sensor
$\ell_j$	$j$ th (failure) event
$\ell_{ij}^t$	Unordered pair of (failure) events $\ell_i$ and $\ell_j$
$\mathcal{L}$	Set of all (failure) events
$\mathcal{L}^t$	Set of all pair-wise (failure) events
$L$	Localization set
$m$	Total number of sensors
$\mathcal{M}$	Influence matrix
$\mathcal{M}^t$	Transformed influence matrix
$n$	Total number of events
$p$	Pressure
$q$	Flow
$S_i$	The location of the $i$ th sensor
$\mathcal{S}$	Set of all sensors
$\mathbf{y}_{\mathcal{S}}$	Outputs of sensors in the set $\mathcal{S}$

Our paper is motivated by the need to consider localization of pipe bursts in the deployment phase of new sensing technologies, since this consideration can significantly reduce the response time and overall costs of fault localization to the distribution utilities. We base our work on the use of low-cost, high-rate online sensors measuring water pressure for remote detection of pipe burst using data mining techniques. Real-world examples are the PIPENET in Boston, MA, US (Stoianov, Nachman, Madden, & Tokmouline, 2007) and the WaterWise in Singapore (Allen et al., 2011). The sensor placement problem is not unique to the water sector and can be found in many engineering applications for system operation. We discuss some of the related work in Section 7.

In Section 2, we present the network and the sensing models and formulate the detection and identification problems as the minimum set cover (MSC) and minimum test cover (MTC) problems, respectively. A key aspect of the MTC problem formulation is the choice of the objective function, which is to select the minimum number of tests from a collection of tests such that every event can be uniquely classified in one of the given categories based on selected tests' outcomes (Moret & Shapiro, 1985). In our setup, the set of outcomes of tests comprises of the output vector from sensors, events are pipe failures, and classification categories are the possible locations of the failed pipes. In Section 3, we present a solution approach as in Halldórsson, Halldórsson, and Ravi (2001) and Svård, Nyberg, and Frisk (2013), in which the MTC is first transformed to the MSC and then solved using the greedy approximation (Minoux, 1978).

In Section 4 we present an augmented greedy algorithm for solving the MTC that does not require the complete transformation of the MTC to the equivalent MSC, and directly computes the objective function in a greedy fashion. This algorithm is much faster than the standard greedy approach and considerably improves the scalability of our approach. In Sections 5 and 6, we demonstrate our approach using a benchmark and a batch of twelve real water distribution networks of various sizes and specifications. We

suggest four metrics to evaluate the performance of the design including detection, identification, and localization scores. Although we demonstrate our results in the context of water networks, our algorithm provides an improved solution to the generic test cover problem. Section 8 summarizes our work and proposes future extensions.

## 2. Problem formulation

Consider the problem of placing online sensors measuring hydraulic pressures with high frequency such that the identification of pipe failure locations is maximized. Based on the number of pipes where link failures (i.e., pipe bursts) can happen, we consider  $n$  link failures as a set of failure events, denoted by  $\mathcal{L} = \{\ell_1, \dots, \ell_n\}$ . For the ease of presentation and without the loss of generality, let  $\ell_j$  denote the failure event at the  $j$ th pipe. Moreover, we define a set of sensors that can be placed at  $m$  nodes of the network as  $\mathcal{S} = \{S_1, \dots, S_m\}$ . Here,  $S_i$  denotes the location of the  $i$ th sensor. The outputs from sensors, which are based on the change in pressure induced by the failure event, are denoted by  $\mathbf{y}_{\mathcal{S}}$ .

### 2.1. Network dynamics and sensing model

A water distribution network can be represented by a graph comprising nodes (supply and demand) connected by links (pipes, valves, and pumps). Physical failures of the infrastructure, such as pipe bursts, cause a disturbance in the flow, which moves through the system as a pressure wave known as *water hammer*, or *surge* with very high velocity, varying typically in the range of 600–1500 ( $\frac{m}{s}$ ) (Misiunas, 2005). This implies that the steady state analysis employed by traditional methods such as supervisory control and data acquisition (SCADA) systems are inadequate and that the transient system dynamics between the initial and the final steady state conditions need to be considered.

The transient system state can be typically described by mass and momentum partial differential equations (Wylie, Streeter, & Suo, 1993). The method of characteristics (MOC) is a numerical technique typically used to approximate the solution of the hydraulic transients. The MOC transforms the partial differential equations into ordinary differential equations that evolve along specific characteristic lines of the numerical grid, which are solved explicitly to compute the head and flow,  $h_{i,t+1}$ ,  $q_{i,t+1}$ , at new point in time and space. Here,  $t$  and  $i$  indicate the discrete points of the numerical grid. For a given pipe, the two characteristic equations describing the hydraulic transients are formulated as Misiunas (2005):

$$h_{i,t+1} = \frac{1}{2} [h_{i-1,t} + h_{i+1,t} + b(q_{i-1,t} - q_{i+1,t}) + r(q_{i+1,t}|q_{i+1,t}| - q_{i-1,t}|q_{i-1,t}|)] \quad (1)$$

$$q_{i,t+1} = \frac{1}{b} [h_{i,t+1} - h_{i+1,t} + q_{i+1,t} - r|q_{i+1,t}|], \quad (2)$$

where  $r$  is the resistance coefficient associated with the steady state, and  $b$  is the impedance coefficient associated with the transient state. For  $b = 0$  the set of Eqs. (1), (2) is reduced to the steady state, where the head loss along a pipe occurs only due to friction (Todini & Rossman, 2013). Additional information describing transient dynamics can be found in the supporting information (SI) (Sela Perelman, Abbas, Koutsoukos, & Amin, 2015).

The effect of a pipe burst at location  $i$  can be translated into boundary conditions using the orifice head-flow relation (Wylie et al., 1993). Before the burst occurs, the cross-section area of the orifice is equal to zero and it increases during a burst, hence we can expect a sudden change in the hydraulic head. The relationship between the head and the pressure, measured by the sensors at

location  $i$ , is related to the elevation of the sensor location. If  $z_i$  is the elevation, and  $p_{i,t}$  is the pressure at location  $i$  at any given time  $t$ , then  $p_{i,t} = (h_{i,t} - z_i) \rho g$ , where  $g$  is the gravitational acceleration  $[\frac{m}{s^2}]$  and  $\rho$  is water density  $[\frac{kg}{m^3}]$ . Hence, the disturbance caused by a pipe burst that reaches the sensor location can be detected by sensing the hydraulic pressure. Similar approaches have been suggested in Zan, Lim, Wong, Whittle, and Lee (2014).

The disturbance caused by the pipe burst quickly dissipates with the distance between the burst event  $\ell_j$  and the location of the sensor  $S_i$ . For the purpose of sensor placement, we are interested in obtaining the sensor's output as a result of some event  $\ell_j$ . Let  $y_{S_i}(t, \ell_j) \in \{0, 1\}$  be a discrete state (output) of the sensor  $S_i$  at time  $t$ , where 1 represents a possible detected event and 0 represents otherwise. Let  $\xi$  be a function characterizing the distance between the expected pressure (i.e., when there is no pipe burst), denoted by  $\hat{p}_{i,t}$ , and the measured pressure, denoted by  $p_{i,t}$ . The sensor output can then be formulated as:

$$y_{S_i}(t, \ell_j) = \begin{cases} 1 & \text{if } \xi(p_{i,t} - \hat{p}_{i,t}) \geq \varepsilon \\ 0 & \text{otherwise} \end{cases} \quad (3)$$

where  $\varepsilon$  is a threshold value. A simple detection model would be where the sensor  $S_i$  indicates an event if the change in the pressure is above some threshold value  $\varepsilon$ . We note here that when the failure event  $\ell_j$  occurs during a given time period, then the output of  $S_i$  will be 1 (or 0) independent of the time of the event  $\ell_j$ . Hence, we can neglect the time dependency of the sensor output to detect the event and can restate the output of the sensor as:

$$y_{S_i}(\ell_j) = \begin{cases} 1 & \text{if } y_{S_i}(t, \ell_j) = 1, \text{ for any } t > 0 \\ 0 & \text{otherwise.} \end{cases} \quad (4)$$

Let  $\mathbf{y}_s(\ell_j) = [y_{S_1}(\ell_j), \dots, y_{S_m}(\ell_j)]$  be the fault signature (Cordier et al., 2004) of the failure event  $\ell_j$  represented by a Boolean vector of the outputs of sensors in the set  $\mathcal{S}$ .

Consequently, for a sensor set  $\mathcal{S}$  and the set of events  $\mathcal{L}$ , we can instantiate a Boolean matrix of dimensions  $|\mathcal{L}| \times |\mathcal{S}|$  called the *influence matrix* and denoted by  $\mathcal{M}$ . The  $j$ th row of  $\mathcal{M}$  consists of sensors' outputs in response to the event  $\ell_j$ , i.e.,  $\mathbf{y}_s(\ell_j)$ . Similarly,  $\mathcal{M}_{ij} = 1$  indicates that a sensor  $S_i$  detected the failure at link  $\ell_j$ , and  $\mathcal{M}_{ij} = 0$  means otherwise. Each row of the influence matrix  $\mathcal{M}$  is analogous to the notion of fault signature in the model-based fault diagnosis systems literature (Cordier et al., 2004).

$$\mathcal{M}(\mathcal{L}, \mathcal{S}) = \begin{bmatrix} \mathbf{y}_s(\ell_1) \\ \mathbf{y}_s(\ell_2) \\ \vdots \\ \mathbf{y}_s(\ell_n) \end{bmatrix}. \quad (5)$$

Furthermore, for the set of link failures  $\mathcal{L}$ , and the set of all possible sensor locations  $\mathcal{S}$ , let  $C_i \subseteq \mathcal{L}$  be the set of link failure events detected by the sensor  $S_i$ , i.e.,  $C_i = \{\ell_j \in \mathcal{L} \mid y_{S_i}(\ell_j) = 1\}$ . If  $\mathcal{C}$  is a collection of all such  $C_i$ 's, i.e.,  $\mathcal{C} = \{C_i : \forall i\}$ , then for a given subset of sensors  $S \subseteq \mathcal{S}$ , we define  $\mathcal{C}_S \subseteq \mathcal{C}$  as a set of subsets of failure events, where a subset corresponds to a sensor in  $S$  that detects the failure events in that subset, i.e.,  $\mathcal{C}_S = \{C_i : S_i \in S\}$ .

**Example 1 (Sensing Model).** To illustrate the network dynamics, consider a small network having 8 nodes connected by 10 links as shown in Fig. 1. A pipe burst event is simulated in the middle of pipe  $\ell_1$  and system response at network nodes is recorded. For the ease of notations, we designate the failure events as pipes' ids,  $\ell_j$ . The transient simulations were computed using the HAMMER software (Bentley, 2015). Fig. 2 shows simulated pressure heads and Boolean outputs  $\mathbf{y}_s$ , for sensors located at nodes 2 and 4. Thus for  $S = \{S_2, S_4\}$  the sensors' state is  $\mathbf{y}_s(\ell_1) = [1, 0]$ . If sensors are placed at all nodes of the network, then the sensors'

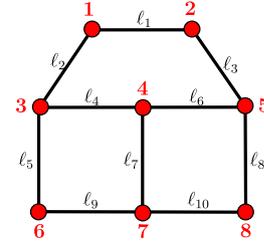


Fig. 1. Illustrative example layout.

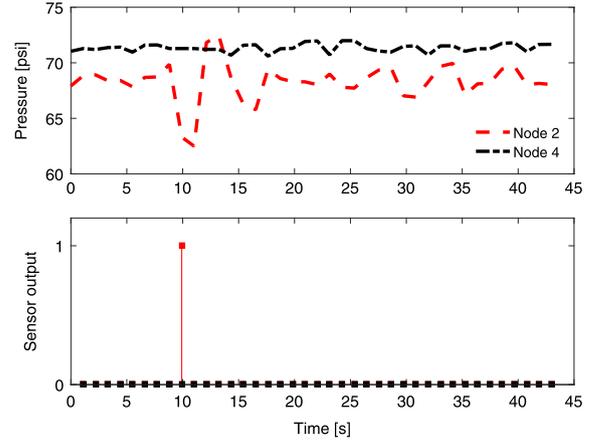


Fig. 2. Failure event generated in pipe  $\ell_1$  in the small example—pressure head [m] and outputs of sensors  $S_2, S_4$ .

state in the case of failure at  $\ell_1$  is  $\mathbf{y}_s(\ell_1) = [1, 1, 1, 0, 1, 0, 0, 0]$ ,  $\mathbf{y}_s(\ell_2) = [1, 1, 1, 1, 0, 1, 0, 0]$ , and so on. The corresponding influence matrix is

$$\mathcal{M}(\mathcal{L}, \mathcal{S}) = \begin{matrix} & S_1 & S_2 & S_3 & S_4 & S_5 & S_6 & S_7 & S_8 \\ \begin{matrix} \ell_1 \\ \ell_2 \\ \ell_3 \\ \ell_4 \\ \ell_5 \\ \ell_6 \\ \ell_7 \\ \ell_8 \\ \ell_9 \\ \ell_{10} \end{matrix} & \begin{pmatrix} 1 & 1 & 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 1 & 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 1 & 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 1 & 0 & 1 & 1 & 0 \\ 0 & 1 & 1 & 1 & 1 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 0 & 1 & 1 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 \end{pmatrix} \end{matrix}.$$

Next, we formulate the detection and identification problems as the minimum set and test cover problems, respectively.

## 2.2. Detection as MSC

For the set of events  $\mathcal{L}$  and the set of sensors  $\mathcal{S}$ , we define a *detectable event* as the one for which there exists at least one sensor in  $\mathcal{S}$  that detects the event. The *detection problem* is to select the minimum number of sensors  $S \subseteq \mathcal{S}$ , such that when a detectable event occurs, at least one sensor in  $S$  detects the event. For a given subset of sensors  $S$ , we define the *detection function*, denoted by  $f_D$ , as follows:

$$f_D(\mathcal{C}_S) = \left| \bigcup_{C_i \in \mathcal{C}_S} C_i \right|. \quad (6)$$

The detection function in (6) gives the number of link failures in  $\mathcal{L}$  that can be detected by the sensors in  $S$ . The detection problem is to select a subset of sensors  $S \subseteq \mathcal{S}$  with the minimum cardinality such that all detectable events are detected, i.e.  $f_D(\mathcal{C}_S) = f_D(\mathcal{C}_\mathcal{S})$ .

The detection performance of a subset of sensors  $S$  is defined as the *normalized detection score*,  $I_D(S)$  and is computed as  $f_D(\mathcal{C}_S)/|\mathcal{L}|$ . The detection problem is equivalent to the *minimum set cover* problem, which could be defined as:

**Definition 2.1** (*Minimum Set Cover (MSC)*). Let  $\mathcal{L}$  be a finite set of elements, and  $\mathcal{C} = \{C_i : C_i \subseteq \mathcal{L}\}$  be the collection of given subsets of  $\mathcal{L}$ . The minimum set cover is to find  $\mathcal{C}_S \subseteq \mathcal{C}$  with the minimum cardinality such that  $\bigcup_{C_i \in \mathcal{C}_S} C_i = \bigcup_{C_j \in \mathcal{C}} C_j$ .

In the above definition, if  $\mathcal{L}$  is the set of link failures and  $\mathcal{C}$  is the collection of  $C_i$ 's corresponding to all the available sensors, then a set cover of minimum size  $\mathcal{C}_S$ , gives the minimum number and locations of sensors that solve the detection problem. Thus, we get the following:

**Proposition 2.1.** *The problem of detection of link failures in a network is equivalent to the minimum set cover problem, and a solution to MSC is therefore, a solution to the detection problem.*

The MSC problem is closely related to the *maximum coverage* problem (Vazirani, 2003), which emerges when the number of sensors that could be used is limited, i.e.,  $|S| \leq B$ . The objective of the maximum coverage problem is to select the sensors such that the number of detectable events is maximized and the constraint  $|S| \leq B$  is satisfied. In Section 3.1 we discuss the *greedy* solution approach, which is very much similar for the MSC and the maximum coverage problems.

### 2.3. Identification as MTC

For the identification of link failures, the goal is to *uniquely* detect the events in  $\mathcal{L}$ , i.e. to distinguish between events using the outputs of sensors. We note that event  $\ell_i \in \mathcal{L}$  can be distinguished from event  $\ell_j \in \mathcal{L}$ , if there exists a sensor in  $\mathcal{S}$  that gives different outputs for  $\ell_i$  and  $\ell_j$ . In such a case, we say that the *pair-wise event*  $\ell_i, \ell_j$  is *detectable* if  $\exists S_p \in \mathcal{S} : \mathbf{y}_{S_p}(\ell_i) \neq \mathbf{y}_{S_p}(\ell_j)$ . In terms of the influence matrix of the network, if a pair-wise event  $\ell_i, \ell_j$  is detectable, then there exists a column with different  $i$  and  $j$  row entries. It follows that an event  $\ell_i$  can be uniquely detected if all pair-wise events  $\ell_i, \ell_j, \forall j \neq i$  are detectable.

The *identification problem* is now defined as follows: for a given  $\mathcal{L}$  and  $\mathcal{S}$ , the *identification problem* is to select a subset of sensors  $S \subseteq \mathcal{S}$  with the minimum cardinality, such that every detectable pair-wise event can be detected by at least one sensor in  $S$ . The *identification function* of  $S$ ,  $f_i(\mathcal{C}_S)$ , is the number of pair-wise events that are detected by a subset of sensors  $S \subseteq \mathcal{S}$ , and will be further discussed in Section 3.2.1. The identification problem is equivalent to the *minimum test cover* problem, which is defined as follows (De Bontridder et al., 2004):

**Definition 2.2** (*Minimum Test Cover (MTC)*). Consider a finite set  $\mathcal{L}$  and a collection of subsets  $\mathcal{C} = \{C_i : C_i \subseteq \mathcal{L}\}$ . The minimum test cover is to find  $\mathcal{C}_t \subseteq \mathcal{C}$  with the minimum cardinality such that if for a pair of elements  $\{\ell_u, \ell_v\} \in \mathcal{L}$ , there exists  $C_i \in \mathcal{C}$  that contains either  $\ell_u$  or  $\ell_v$  but not both, then there exists some  $C_j \in \mathcal{C}_t$  that also contains either  $\ell_u$  or  $\ell_v$ , but not both.

The identification problem is to find a subset  $\mathcal{C}_t \subseteq \mathcal{C}$  of minimum cardinality, or equivalently the corresponding subset of sensors  $S \subseteq \mathcal{S}$ , such that if  $\mathbf{y}_S(\ell_j)$  is unique with respect to the set of all sensors  $\mathcal{S}$ , then  $\mathbf{y}_S(\ell_j)$  is also unique with respect to a subset of sensors  $S$ , which is the MTC problem defined above. Thus, we can state:

**Proposition 2.2.** *The problem of identification of link failures in networks is equivalent to the minimum test cover problem, and therefore, a solution to MTC is also a solution to the identification problem.*

**Example 2** (*Detection vs. Identification*). Following Example 1, consider two sensors placed at nodes 2 and 4,  $S = \{S_2, S_4\}$ . For the detection problem, we note that  $C_2 \cup C_4 = \mathcal{L}$ . That is, at least one of the sensors in  $S$  has an output 1 whenever a link fails. Thus, sensors  $S_2$  and  $S_4$  *cover* (detect) all link failures and solve the detection problem. For the identification problem, sensors 2 and 4 are not sufficient as they generate only three unique states associated with the 10 events, which makes it impossible to distinguish between all link failures. For example, the state  $\{1, 0\}$  is uniquely associated with a failure in link  $\ell_1$ , whereas, the state  $\{1, 1\}$  can be associated with a failure in any of the links  $\ell_2, \ell_3, \ell_6$ , or  $\ell_8$ . However, for the set of sensors  $S^* = \{S_1, S_2, S_3, S_5\}$ , which solves the MTC problem for Example 1, the output is unique for each link failure, i.e. ten distinct indicator vectors, each corresponding to a unique failure event, are obtained.

### 3. Greedy MTC solution

It is well known that both MSC and MTC are NP-hard problems (Garey & Johnson, 1979; Vazirani, 2003). In this section, we first introduce an approximate solution to the MSC, which will be utilized in Section 4 for constructing a computationally efficient solution of the MTC problem.

#### 3.1. Detection solution

MSC has been studied extensively owing to its wide variety of applications in theoretical as well as practical domains. A straightforward way to solve the MSC is by the *greedy approach*. The greedy approach is to select, in each iteration, a sensor that detects the maximum number of undetected link failures, until all link failures are detected, or no further link failure can be detected by any sensor. In the maximum coverage problem, iterations continue until a given number of sensors are selected. If  $n$  is the total number of link failures,  $m$  is the total number of sensors, then greedy algorithm is the  $(\ln n + 1)$  approximation of the MSC. Since MSC cannot be approximated to  $(1 - o(1)) \ln n$  unless  $P = NP$  (Dinur & Steurer, 2014), greedy algorithm gives the optimal approximation ratio. Moreover, if  $k$  is the maximum number of link failures that can be detected by any sensor, then the greedy algorithm has the approximation ratio of  $(\ln k + 1)$  (Vazirani, 2003). In our context,  $k$  depends on the network topology and the sensing model as in (5). Similarly, for the maximum coverage problem, the greedy algorithm gives the approximation ratio of  $(1 - 1/e)$ , which is again the best possible.

Although the greedy approach gives the best known approximation ratio, its straightforward implementation requires a large number of function (as in (6)) evaluations. The running time of greedy approach is a function of the number of sensors and events,  $\mathcal{O}(mn)$ . For large scale systems, in which  $n$  and  $m$  are very large, this simple greedy approach becomes computationally intractable owing to a large number of function evaluations, even if computing a function is not expensive. However, greedy algorithm can be made faster by reducing the number of function evaluations if the *submodularity* property is satisfied (Minoux, 1978). Submodular functions can be defined as follows:

**Definition 3.1** (*Submodularity*). Let  $\mathcal{C}$  be a finite set and  $f$  be a set function,  $f : 2^{\mathcal{C}} \rightarrow \mathbb{R}$ . Moreover,  $\mathcal{C}_s \subseteq \mathcal{C}_r \subseteq \mathcal{C}$ , and  $C_i \in \mathcal{C} \setminus \mathcal{C}_r$ , then  $f$  is submodular whenever

$$f(\mathcal{C}_s \cup \{C_i\}) - f(\mathcal{C}_s) \geq f(\mathcal{C}_r \cup \{C_i\}) - f(\mathcal{C}_r). \quad (7)$$

For the detection problem, this means that as the number of link failures detected by the selected sensors increases, the marginal value of adding a sensor to the cover decreases. It can be shown that the function in (6) is submodular (see Sela Perelman et al., 2015),

and the submodularity of  $f_D$  can be exploited to obtain the *lazy greedy* algorithm as in [Minoux \(1978\)](#). The basic idea behind the lazy greedy approach is to eliminate the redundant computations in each iteration. This can be further explained as follows: For the  $\kappa$ th iteration, let  $F_\kappa(C_i)$  denote the utility of adding a sensor  $i$  to the cover, i.e.  $f_D(\mathcal{C}_s \cup \{C_i\}) - f_D(\mathcal{C}_s)$ , then by the submodularity of  $f_D$ , we know that  $F_{\kappa+1}(C_i) \leq F_\kappa(C_i)$ . Moreover, without the loss of generality, we assume that  $F_\kappa(C_1) \geq F_\kappa(C_2) \geq \dots$ , then  $C_1$  is the greedy choice in the  $\kappa$ th iteration. However, in the next iteration, if  $F_{\kappa+1}(C_2) \geq F_{\kappa+1}(C_3)$ , then  $F_{\kappa+1}(C_2) \geq F_{\kappa+1}(C_j)$ ,  $\forall j \geq 3$ , which means that there is no need to compute  $F_{\kappa+1}(C_j)$ ,  $\forall j \geq 3$ . This saves a large number of potential computations and improves scalability of the solution approach to large scale systems.

### 3.2. Identification solution

One approach to solve the MTC problem is to first transform it to an equivalent MSC problem ([De Bontridder et al., 2004](#)), and then to solve the MSC problem using lazy greedy algorithm, as explained earlier. The greedy approach to solve the MTC yields a  $(2 \ln n + 1)$  approximation ratio algorithm, which is the best possible ([Moret & Shapiro, 1985](#)). A solution of the equivalent MSC is a solution to the original MTC problem. Thus, a straight-forward way to solve the identification problem for link failures is to first obtain an equivalent detection problem, in which each event represents a *pair-wise* link failure, and then utilize the greedy approach to solve the corresponding detection problem. We call this the *transformed lazy greedy (TLG)* and will use it in Section 6.2 to demonstrate the simulation results. Next, we summarize the transformation of the MTC to the MSC problem as outlined in [De Bontridder et al. \(2004\)](#).

#### 3.2.1. Transformation of MTC to MSC

Given an instance of the MTC, i.e.,  $\mathcal{L}$  and  $\mathcal{C} = \{C_i\}$ , where  $C_i \subseteq \mathcal{L}$ , we transform the MTC to the MSC by taking the following two steps:

- *Create a new set of events:*  $\mathcal{L}^t = \{\ell_{12}^t, \dots, \ell_{(n-1)n}^t\}$ . For each unordered pair  $\{\ell_i, \ell_j\}$ , define a new element  $\ell_{ij}^t$ ;  $\mathcal{L}^t$  consists of all such  $\ell_{ij}^t$ 's.
- *Create a new sets of sensors' outputs:*  $\mathcal{C}^t = \{C_1^t, \dots, C_m^t\}$ , where  $C_v^t = \{\ell_{ij}^t : |\{\ell_i, \ell_j\} \cap C_v\} = 1\}$ ,  $\forall v \in \{1, \dots, m\}$ . In other words,  $\ell_{ij}^t \in C_v^t$  if and only if exactly one of  $\ell_i$  or  $\ell_j$  is in  $C_v$ . Moreover, for a subset of sensors  $S \subseteq \mathcal{S}$ , we define  $\mathcal{C}_S^t = \{C_v^t : S_v \in S\}$ .

Hence, we obtain a new identification matrix  $\mathcal{M}^t(\mathcal{L}^t, \mathcal{S})$  of dimensions  $\binom{n}{2} \times m$ , in which each row corresponds to a *pair-wise link failure* and each column represents sensor's output. If a specific row in  $\mathcal{M}^t$  represents a pair  $\ell_i, \ell_j$ , then the  $v$ th column entry of the corresponding row in  $\mathcal{M}^t$  is an *exclusive OR* of the  $(i, v)$ th and  $(j, v)$ th entries of the influence matrix  $\mathcal{M}$ . The above point illustrates the fact that to localize an event  $\ell_i$ , there always exists a sensor that distinguishes  $\ell_i$  from  $\ell_j$  by producing different outputs for  $\ell_i$  and  $\ell_j$  respectively, i.e., if a sensor output is 1 (resp. 0) in case of  $\ell_i$ , then its output for  $\ell_j$  is 0 (resp. 1), for all  $j \neq i$ .

Note that for a given subset of sensors  $S$ , the identification function, which is the number of pair-wise link failures detected by  $S$ , is essentially same as the detection function of  $S$  in the corresponding MSC instance i.e.,

$$f_i(\mathcal{C}_S) = f_D(\mathcal{C}_S^t), \quad (8)$$

where  $f_D$  is defined as in (6). The *normalized identification score*, denoted by  $I_i(S)$ , is computed by dividing  $f_i$  by the total number of pair-wise events,  $|\mathcal{L}^t|$ .

#### 3.2.2. Greedy approach based solution

Once the MTC problem has been transformed to the MSC problem, a straightforward way to obtain a solution is to employ the greedy algorithm, as outlined in Algorithm 1.

### Algorithm 1 Minimum Test Cover – Greedy Algorithm

- 1: **Input:**  $\mathcal{C} = \{C_1, \dots, C_m\}$ ,  $C_i \subseteq \mathcal{L}$
- 2: **Output:** MTC:  $\mathcal{C}^* \subseteq \mathcal{C}$
- 3: **Initialize:**  $\mathcal{C}^* \leftarrow \emptyset$
- 4: **Transform:** the test cover instance to the set cover instance, i.e., from a given  $\mathcal{L}$  and  $\mathcal{C}$ , obtain a corresponding  $\mathcal{L}^t$  and  $\mathcal{C}^t$  (Section 3.2.1).
- 5: **Solve:** using greedy algorithm
  - (a) Select  $C_{i^*}^t \in \mathcal{C}^t$  (i.e., the sensor  $i^*$ ) covering the most uncovered elements in  $\mathcal{L}^t$ .
  - (b) Add to current set  $\mathcal{C}^* \leftarrow \mathcal{C}^* \cup \{C_{i^*}^t\}$ .
  - (c) Repeat until all elements in  $\mathcal{L}^t$  are covered, or no new element in  $\mathcal{L}^t$  can be covered by any  $C_i^t \in \mathcal{C}^t$ .

As in the case of the MSC problem, the lazy greedy approach, which exploits the submodularity property of the set cover problem, can be utilized. However, if there are  $n$  link failures that need to be localized, then the corresponding set cover instance contains  $\binom{n}{2}$  events, and the time complexity of the greedy approach in Algorithm 1 is  $\mathcal{O}(m \binom{n}{2})$ , where  $m$  is the total number of sensors. Even for small-sized networks with a limited number of possible link failures, this approach becomes quite inefficient owing to a large number of computations required. Moreover, employing lazy greedy also achieves desired computational efficiency for realistic size of failure event set. In the next section, we focus on improving the computational time of the solution of the MTC problem.

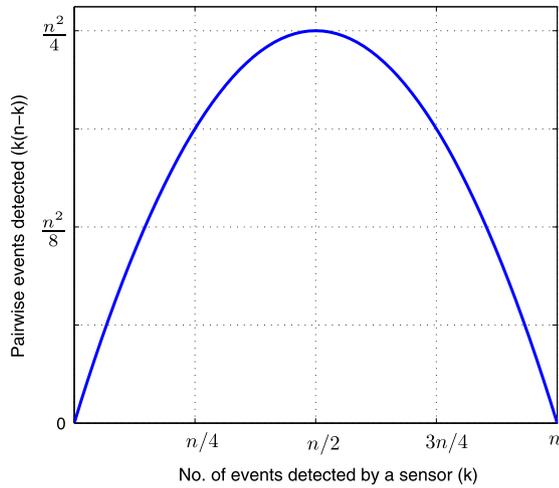
## 4. Augmented greedy MTC solution

The main idea behind the augmented greedy approach is to achieve a computationally efficient approximation algorithm. We do so by avoiding the complete transformation of the MTC to the MSC and directly evaluating the function (8), thus eliminating the need to *pre-compute* the identification matrix  $\mathcal{M}^t(\mathcal{L}^t, \mathcal{S})$ .

In each iteration of the greedy algorithm for the MTC solution, a sensor that covers (detects) the most pair-wise link failures from a total of  $\binom{n}{2}$  pair-wise failures, is selected. Thus  $\mathcal{O}(\binom{n}{2})$  comparisons are made in a single iteration for each potential sensor. In the augmented greedy approach, we avoid this by significantly reducing the number of comparisons made in each step. In fact, for each sensor, the number of comparisons made in a single iteration is always bounded by  $\mathcal{O}(K \binom{k}{2})$ , where  $k$  is the maximum number of link failures that are detected by any sensor, and  $K$  is the number of sensors that are included in the test cover until that iteration. Since  $k$  is typically much smaller than  $n$ , a large number of computations are thus avoided in each iteration.

To explain our approach, we first observe that a sensor  $i$  that detects  $k$  events (i.e.,  $|C_i| = k$ ) can distinguish between  $k$  detected events and  $(n - k)$  undetected events. Thus, such a sensor detects  $k(n - k)$  *pair-wise* events (i.e.,  $|C_i^t| = k(n - k)$ ). Unlike the detection problem, in which a sensor with a large  $k$  is desirable for the detection purposes, a sensor that detects a large number of failures is not always useful for the identification. Fig. 3 shows the number of pair-wise events detected by a (single) sensor as a function of the number of detected events. The maximum number of pair-wise events, which are link failures in our case, are detected when  $k = n/2$ .

Moreover, if a sensor  $i$  included in a test cover and  $\ell_u, \ell_v \in C_i$ , then a distinction between the occurrence of  $\ell_u$  and  $\ell_v$  is not possible through the sensor  $i$ . Thus, a set of sensors that can distinguish between events  $\ell_u, \ell_v \in C_i$ , or equivalently that can detect pair-wise events corresponding to the events in  $C_i$ , also need to be included in the test cover. Based on this observation, we suggest an augmented greedy approach to compute the test cover without computing the  $\binom{n}{2}$  events priori.



**Fig. 3.** The number of pair-wise link detections as a function of the number of detected events.

Let  $C^* \subseteq \mathcal{C}$  be the test cover until the current iteration, and  $\mathcal{C}_{cov}$  be the set of link failures detected by the sensors that are included in the test cover, i.e.,  $\mathcal{C}_{cov} = \bigcup_{C_i \in C^*} C_i$ . Thus, the utility of adding  $C_i$  to  $C^*$  (i.e., adding sensor  $S_i$  to the test cover) in each iteration is based on the following two factors:

- (i) How many pair-wise link failures corresponding to the links which are **not included** in  $\mathcal{C}_{cov}$  can be detected by  $C_i$ ? We define this value as  $x_i$ .
- (ii) How many pair-wise link failures corresponding to the links already **included** in  $\mathcal{C}_{cov}$  can be detected by  $C_i$ ? We define this value as  $y_i$ .

The overall utility of adding sensor  $S_i$  to the test cover, denoted by  $w_i$ , is the sum of  $x_i$  and  $y_i$ . A sensor  $S_{i^*}$  that maximizes this overall utility, let  $w_{i^*}$  denote the maximum utility, will then be included in the test cover, and  $\mathcal{C}_{cov}$  will be updated to  $\mathcal{C}_{cov} \leftarrow \mathcal{C}_{cov} \cup C_{i^*}$ . Now, we state how to compute  $x_i$  and  $y_i$  in the  $j$ th iteration.

- (i) *Computing  $x_i$* —If  $n_j$  is the number of link failures that are not yet included in  $\mathcal{C}_{cov}$ , (i.e.,  $n_j = n - |\mathcal{C}_{cov}|$ ), and  $C_i$  contains  $k_{i,j}$  of such link failures, then  $x_i = k_{i,j}(n_j - k_{i,j})$ . Note that computing  $x_i$  is very straight forward and does not require computing pair-wise link failures from a given set of link failures.
- (ii) *Computing  $y_i$* —If a sensor  $u$  is already included in the test cover, then the pair-wise link failures corresponding to the links in  $C_u$  remain undetected. Thus,  $y_i$  computes how many of such pair-wise link failures can be detected by the inclusion of sensor  $i$  in the test cover. To make it precise, we proceed as follows:

If  $X$  and  $Y$  are two sets, then we define:

$\beta(X) =$  set of all 2-element subsets of  $X$ ,

and

$\alpha(Y, \beta(X)) = \{a \in \beta(X) : |Y \cap a| = 1\}$ .

Here,  $\alpha(Y, \beta(X))$  is a set consisting of such 2-element subsets of  $X$  that have exactly one common element with  $Y$ . For instance, if  $X = \{1, 2, 3\}$  and  $Y = \{1, 3\}$ , then  $\beta(X) = \{\{1, 2\}, \{1, 3\}, \{2, 3\}\}$ , and  $\alpha(Y, \beta(X)) = \{\{1, 2\}, \{2, 3\}\}$ .

To compute  $y_i$ , first we compute the set of link failures common to  $C_i$  and  $\mathcal{C}_{cov}$  and call it as  $Y_i = C_i \cap \mathcal{C}_{cov}$ . Now, if sensor  $u$  is already included in the test cover, and  $G_u \subseteq \beta(X_u)$  is the set of undetected pair-wise link failures corresponding to the links in  $X_u \subseteq C_u$ , then

$$y_i = \sum_{C_u \in C^*} |\alpha(Y_i, G_u)|.$$

The complete algorithm is stated in Algorithm 2.

### Algorithm 2 Minimum Test Cover – Augmented Greedy Algorithm

```

1: Input:  $\mathcal{C} = \{C_1, \dots, C_m\}$ ,  $C_i \subseteq \mathcal{L}$ 
2: Output: MTC:  $\mathcal{C}^* \subseteq \mathcal{C}$ 
3: Initialization:  $\mathcal{C}_{cov} = \emptyset$ ;  $\mathcal{C}^* = \emptyset$ ;  $G_0 = \emptyset$ ;  $j = 1$ ;  $n = |\mathcal{L}|$ ;  $w_{i^*} = 1$ ;
4: while  $w_{i^*} > 0$  do
5:    $n_j \leftarrow n - |\mathcal{C}_{cov}|$ 
6:   for all  $i$  do
7:      $X_i \leftarrow (C_i \setminus \mathcal{C}_{cov})$ ;  $k_{i,j} \leftarrow |X_i|$ 
8:      $x_i \leftarrow k_{i,j}(n_j - k_{i,j})$ 
9:      $Y_i \leftarrow C_i \cap \mathcal{C}_{cov}$ 
10:     $y_i \leftarrow \sum_{t=0}^{j-1} |\alpha(Y_i, G_t)|$ 
11:     $w_i = x_i + y_i$ 
12:  end for
13:   $w_{i^*} \leftarrow \max w_i$ 
14:  if  $w_{i^*} > 0$  then
15:     $\mathcal{C}^* \leftarrow \mathcal{C}^* \cup \{C_{i^*}\}$ 
16:     $\mathcal{C}_{cov} \leftarrow \mathcal{C}_{cov} \cup C_{i^*}$ 
17:     $G_j \leftarrow \beta(X_{i^*})$ 
18:    for  $t = 0$  to  $j - 1$  do
19:       $G_t \leftarrow G_t \setminus \alpha(Y_{i^*}, G_t)$ 
20:    end for
21:     $j \leftarrow j + 1$ 
22:  end if
23: end while

```

**Example 3 (Augmented Greedy).** Consider the network shown in Fig. 1. Let  $k_i$  be the number of failure events detected by the sensor  $i$ , i.e.,  $|C_i| = k_i$ , where  $C_i \subseteq \mathcal{S}$ . In the first iteration ( $j = 1$ ) of the while loop, size of the event space is  $n = 10$ , and  $k_{i,j} = k_i$ ,  $\forall i$ . Then, the number of new pair-wise link failures detected by the sensor  $i$  is given by  $x_i = k_{i,j}(n - k_{i,j})$ . Since there are no sensors in the test cover in the first iteration,  $y_i = 0$  for all the sensors. The maximum value of  $w_i$  is attained for the sensors 1 and 2 with  $w_1 = w_2 = x_1 = x_2 = 5(10 - 5) = 25$ . We include sensor 1 in the test cover, thus  $\mathcal{C}^* = C_1$  after the first iteration of the while loop. The set of all undetected pair-wise events for sensor 1,  $G_1 = \{\{1, 2\}, \{1, 3\}, \dots, \{4, 5\}\}$ , is then updated. Finally, we update the number of covered events as  $\mathcal{C}_{cov} = \{1, 2, 3, 4, 5\}$ . For the second iteration, i.e.,  $j = 2$ , size of the event space is updated as  $n_2 = 5$ . A complete account of the states of variables of the algorithm for the example is provided in (Sela Perelman et al., 2015). The algorithm returns the test cover consisting of sensors  $\{1, 2, 3, 5\}$  that uniquely identify all link failures.

The augmented greedy approach in Algorithm 2 produces the same solution as the greedy approach in Algorithm 1. Thus, Algorithm 2 has the same approximation ratio as the standard greedy algorithm, which has been proven to be the best possible.

Since a large number of computations are avoided in the execution of Algorithm 2, it is more efficient than the simple greedy. In contrast to the  $\mathcal{O}\left(\binom{n}{2}\right)$  comparisons performed in each iteration for a sensor in Algorithm 1,  $\mathcal{O}\left(\sum_i^{m_j} \binom{k_i}{2}\right)$  comparisons are done in each iteration of the Algorithm 2. Here,  $n$  is the total number of link failures,  $k_i$  is the number of link failures detected by the sensor  $i$  (i.e.,  $k_i = |C_i|$ ), and  $m_j$  is the number of sensors included in the test cover until that iteration. Thus, if  $k = \max(k_i)$ , then Algorithm 2 is at least  $n/k$  times faster than the simple greedy approach as shown below. Moreover, typically  $k \ll n$  in the case of link failure detection in water distribution networks, thus,  $n/k$  factor turns out to be a significant improvement.

**Proposition 4.1.** Let  $\sum_i k_i = n$ , and  $k = \max(k_i)$ , then

$$\sum_i \binom{k_i}{2} \leq \frac{k}{n} \binom{n}{2}. \quad (9)$$

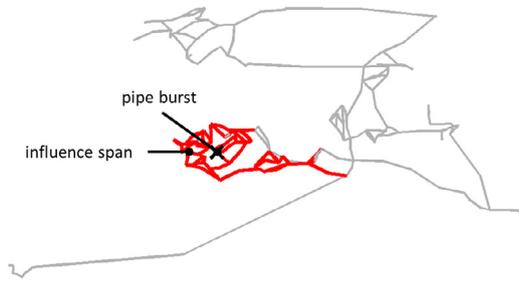


Fig. 4. Layout of Net1 and propagation of failure in LINK-126. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

**Proof.** –

$$\begin{aligned} \sum_i \binom{k_i}{2} &= \frac{1}{2} \left( \sum_i k_i^2 - \sum_i k_i \right) \leq \frac{1}{2} \left( k \sum_i k_i - n \right) \\ &= \frac{1}{2} (kn - n) \leq \frac{1}{2} (kn - k) = \frac{k}{n} \binom{n}{2}. \quad \square \end{aligned}$$

We note that Algorithm 2 is somewhat similar to the two-step greedy algorithm presented in De Bontridder et al. (2004). However, in our approach, both  $x_i$  and  $y_i$  are computed in the same iteration resulting in a more efficient implementation.

## 5. Application to a benchmark network

We first test our approach on a medium-size water network. *Net1* is a benchmark system that has been extensively studied in the context of sensor placement for water quality (Ostfeld et al., 2008). The system consists of 126 nodes, 168 pipes, one reservoir, one pump, and two storage tanks and its layout is shown in Fig. 4. The system supplies a daily demand of  $5.15 \times 10^3$  (m<sup>3</sup>/day) and has a total pipe length of  $37.5 \times 10^3$  (m).

For all our simulations, we consider a single failure event occurring at the center of each pipe and enumerate all possible failure events. For the detection problem, when fully calibrated transient model of the network is not available, we approximate the disturbance propagation using a simple distance based model emulating the dissipation of the pressure wave with the distance from the origin. As in Deshpande, Sarma, Youcef-Toumi, and Mekid (2013), our influence model is based on the shortest distance threshold model, assuming that the disturbance in pressure can be sensed within a specified distance from the location of the burst, i.e.,  $\mathbf{y}_S(\ell_j) = \{1 | d(S_i, \ell_j) \leq \varepsilon\}$ , where  $d$  is the length of the shortest path between two locations  $S_i$  and  $\ell_j$ , and  $\varepsilon$  is some threshold. Fig. 4 shows an example of the influence range (in red) of a burst in LINK-126 of the network for a threshold distance of  $\varepsilon = 1000$  (m), i.e., a sensor located in the red region can detect the pipe failure.

Assuming that a sensor can be placed at any of the 126 network nodes and any of the 168 network pipes can fail, we solve the MTC problem, as described previously in Sections 2.3, 3.2 and 4. Fig. 5 shows the normalized identification score,  $I_I$ , defined in Section 3.2.1, as a function of the number of sensors using the greedy approach. As noted in Section 3.1, we observe that the identification score function exhibits a diminishing return property. The maximum identification score of 0.99 is attained with 48 sensors.

Observing that the identification score of the network is not sufficient to evaluate the quality of the design, since it does not indicate about the number of events that are uniquely identified and, respectively, the number of events that are not uniquely identified. For this reason, we suggest two complementary metrics for evaluating the performance of the sensor network design:

**Localization score**—Let  $L \subseteq \mathcal{L}$  be a subset of all such link failures for which the outputs of sensors in  $S$  are same, i.e.,  $\mathbf{y}_S(\ell_i) = \mathbf{y}_S(\ell_j)$ ,  $\forall \ell_i = \ell_j \in L$ . We call such a subset of link failures  $L$  as a

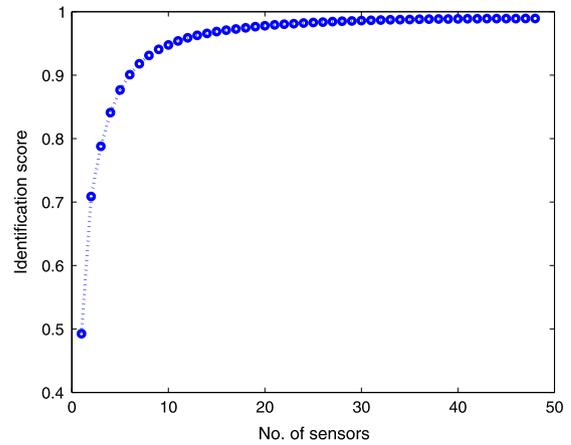


Fig. 5. Identification score for Net1.

**localization set.** A localization can be associated with every unique vector of sensors' outputs. Localization score is the total number of localization sets obtained under the sensor configuration  $S$ . We note that it is not possible to distinguish between the failure events in a localization set by merely observing the outputs of sensors. We define the normalized localization score,  $I_I(S)$ , as the ratio of the total number of localization sets formed under the sensor configuration  $S$  to the total number of event failures. Ideally, the normalized localization score should be equal to 1, indicating that each fault can be uniquely identified.

**Localization size**—is the number of faults associated with a unique output of sensors, or the number of elements in a localization set  $L$ . A localization size of higher value means that it would be difficult to identify the location of the fault, and additional local inspection methods might be needed. We define the worst set size,  $I_W(S)$  as the largest localization set. For complete localization it is required that,  $I_W(S) = 1$ , indicating that all faults could be distinguished from each other, and therefore could be uniquely detected.

**Example 4 (Localization Score).** Continuing Example 2 for the two-sensor design  $S = \{S_2, S_4\}$ , three localization sets are formed, i.e.  $L_1 = \{\ell_1\}$ ,  $L_2 = \{\ell_4, \ell_5, \ell_7, \ell_9, \ell_{10}\}$ ,  $L_3 = \{\ell_2, \ell_3, \ell_6, \ell_8\}$ . The corresponding localization sizes are  $|L_1| = 1$ ,  $|L_2| = 5$ ,  $|L_3| = 4$ . The normalized localization score is thus  $I_I = 3/10$  and the worst localization size is  $I_W = 5$ . It means that if an event is detected, its distinction between three distinct groups is possible, but further distinction within the groups is not possible, with the largest indistinctive group of 5 links. With the four-sensor design,  $S^* = \{S_1, S_2, S_3, S_5\}$ , the optimal normalized localization score and the maximum localization size of 1 are achieved, and we observe ten unique outputs of sensors, each associated with a unique failure event.

Fig. 6(a) shows the normalized localization score as a function of the number of sensors. The highest localization score of 0.65 is achieved when 48 sensors are installed. This result indicates that 110 unique vectors of sensors output are associated with the 168 failure events. Fig. 6(b) shows the worst, median, and minimum localization set sizes as a function of the number of sensors for Net1. We observe that initially sizes of localization sets decrease rapidly with the number of sensors, until the worst localization-set-size reaches a plateau at 20 sensors, and does not improve further. This implies that deploying more sensors might improve local performance, but will not improve the overall network localization performance, making further deployment of sensor unattractive for the water utility from the cost viewpoint.

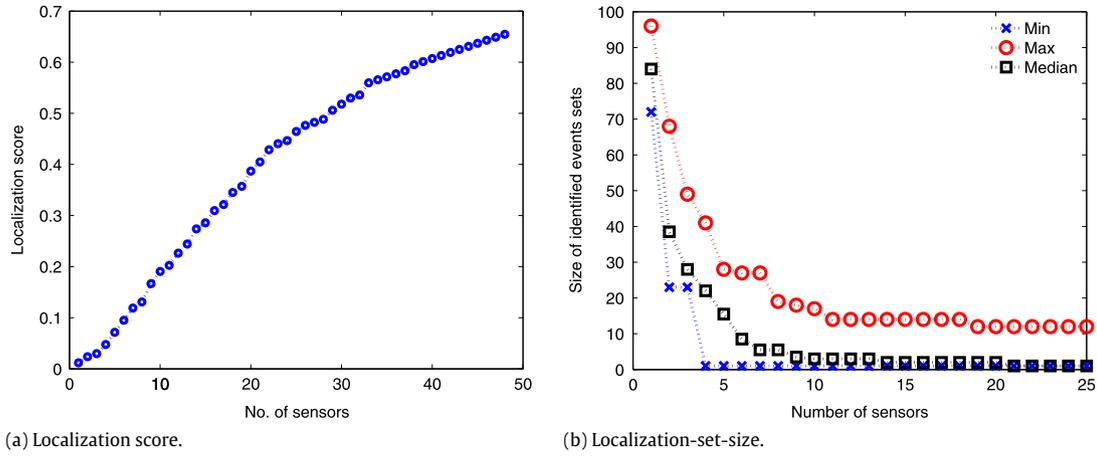


Fig. 6. Localization performance for Net 1.

Table 1

Network data.

Network	Length (km)	Demand $10^3$ (m <sup>3</sup> /day)	No. of pipes	No. of nodes
Net1	37.56	5.15	168	126
Net2	91.29	7.59	366	269
Net3	96.58	8.58	496	420
Net4	137.05	5.78	603	481
Net5	123.20	6.20	644	543
Net6	166.60	5.66	907	791
Net7	153.30	8.93	940	778
Net8	152.25	7.91	1 124	811
Net9	260.24	5.67	1 156	959
Net10	247.34	9.33	1 614	1 325
Net11	760.89	71.88	3 032	1 891
Net12	1844.04	108.80	14 822	12 523

## 6. Application to real networks

We tested our approach on a batch of real water networks. Principal information is listed in Table 1 and the complete data can be obtained from Jolly, Lothes, Bryson, and Ormsbee (2014) for Nets 2–10 and from Centre of Water Systems University of EXETER (2015) for Nets 1, 11, 12. In all our simulations we again assume, that a single failure can occur at each of the network links and that sensors can be placed at each of the network nodes, and set the distance threshold to  $\varepsilon = 1000$  (m).

### 6.1. MSC vs. MTC

First, we compare the sensor placement design for the identification problem obtained from our approach with the design for the detection problem, i.e. MTC vs. MSC (Sections 2.2 and 2.3). We demonstrate our results using Net9, from the Kentucky dataset. Although the system supplies similar daily demand as Net1, it is spatially more distributed with approximately 260 (km) of pipes. Network layout and main features are shown in Fig. 7 and Table 1.

Fig. 7 schematically illustrates the difference between the MTC and MSC problem formulations in the context of Net9. Consider three sensors installed in the network, Fig. 7 demonstrates the seven localization sets corresponding to seven unique sensor states,  $[0, 0, 1], \dots, [1, 1, 1]$ , i.e. we can distinguish between events in different sets but not between events in the same set, and the detection set, being the union of the localization sets. Whereas the detection problem tries to maximize the detection set, the identification problem aims to identify distinct subsets.

Fig. 8 provides a comparison between the detection and localization scores for the MTC (blue circles) and MSC (red squares) designs. For the detection problem, 25 sensors are sufficient to cover the entire system, hence, we also select the first 25 sensors for the identification problem and compare their performance.

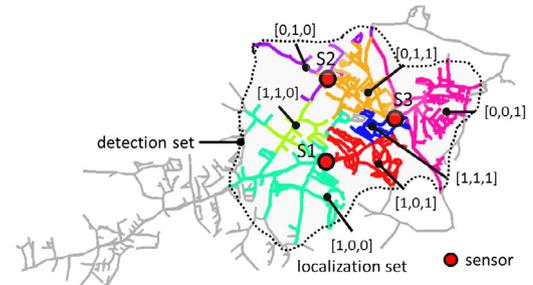
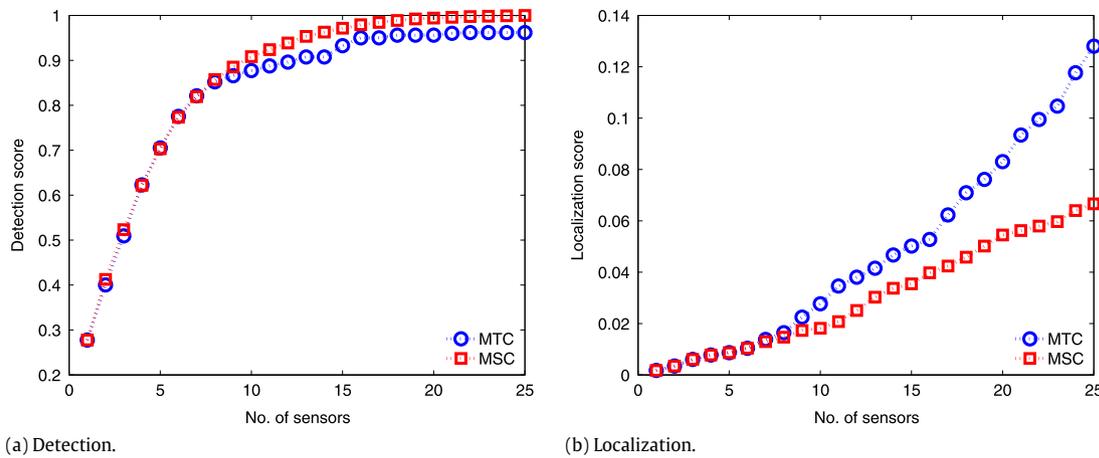


Fig. 7. Layout of Net9 and example of the detection and localization sets for three sensors.

From Fig. 8(a) it can be seen that the two designs overlap for the first 7 sensors and the MSC design only slightly outperforms the MTC design when comparing the detection scores for a higher number of sensors. At the same time, the MTC design significantly outperforms the MSC design when comparing the localization scores as shown in Fig. 8(b). Similar results were attained for the other networks.

### 6.2. Augmented greedy vs. transformed lazy greedy

Next, we compare the solution approach based on the augmented greedy (AG) (Section 4) and the transformed lazy greedy (TLG) (Section 3.2). Table 2 lists the running times (Intel Core i7, 2.9 GHz, 16 GB of RAM) for the augmented greedy and the transformed lazy greedy approaches. For Nets 1–10, the new algorithm is 3–8 times faster than the transformed lazy greedy approach, depending on the maximum number of events detected by any sensor (see Proposition 4.1). The solutions obtained using the two approaches were identical. For Nets 11–12, we were not able to apply the TLG due to the memory requirements and applied only the AG, which further emphasizes the advantage of the AG approach.



**Fig. 8.** MTC vs. MSC performance for Net9. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

**Table 2**

Simulation results.

Network	No. of sensors	$I_D$	$I_I$	$I_L$	$I_W$	TLG (min)	AG (min)
Net1	48	0.99	0.99	0.65	12	0.23	0.08
Net2	98	0.99	1.00	0.86	12	2.39	0.58
Net3	134	0.99	1.00	0.86	7	6.93	1.65
Net4	138	0.99	1.00	0.91	8	11.98	4.93
Net5	164	0.99	1.00	0.86	6	15.58	3.85
Net6	258	1.00	1.00	0.86	8	45.46	6.31
Net7	139	1.00	1.00	0.83	8	49.12	9.31
Net8	195	1.00	1.00	0.70	8	80.55	28.07
Net9	359	1.00	1.00	0.87	6	91.57	11.06
Net10	408	1.00	1.00	0.89	14	257.41	39.48
Net11	717	1.00	1.00	0.69	9	–	50.53
Net12	1000 <sup>a</sup>	1.00	1.00	0.38	17	–	1800

TLG—transformed lazy greedy; AG—augmented greedy;

<sup>a</sup> Terminated after 1000 iterations.

Finally, Table 2 lists the maximum number of sensors and the corresponding four performance scores: normalized detection  $I_D$ , identification  $I_I$ , and localization  $I_L$  scores, and worst localization set size  $I_W$ . For all networks, the layouts and the simulation plots illustrating these metrics as a function of the number of sensors are available in Sela Perelman et al. (2015). These results demonstrate that: (1) The number of sensors required solely for detection purpose is significantly lower than the number of sensors required for localization. (2) Between the two localization measures,  $I_L$  and  $I_W$ , the localization score is more conservative than the worst set size, requiring a larger number of sensors. For example, consider the design for Net9, then to detect 95% of the events, i.e.,  $I_D = 0.95$ , 18 sensors are sufficient, whereas to achieve  $I_L = 0.5$  we require 79 sensors, and 38 to achieve  $I_W = 20$ . This is observed for all tested networks.

## 7. Related work

*Event detection in water networks.* In the urban water sector, majority of previous works focused on the sensor placement for detecting hypothetical contamination events assuming perfect sensors capable of detecting all types of contaminants (Berry, Hart, Phillips, Uber, & Watson, 2006; Eliades & Polycarpou, 2010). In a related work (Krause, Leskovec, Guestrin, Vanbriesen, & Faloutsos, 2008), to detect the presence of contaminants in large water distribution systems, the notion of penalty reduction function was introduced to realize various objective functions such as reduction of detection time and the expected population affected. Submodularity of the penalty reduction function was then used to solve sensor placement problems efficiently and with provable guarantees. Moreover, various data and model-driven techniques

also exist that are applied for system's state estimation and event detection and isolation (Eliades, Lambrou, Panayiotou, & Polycarpou, 2014; Rosich, Frisk, Aslund, Sarrate, & Nejari, 2012). The basic premise in these methods is that once the sensors are in place, data is collected and transmitted in real-time. The difference between measurements, such as pressure (Perez et al., 2014) and flow (Ragot & Maquin, 2006), and their estimated values obtained using the network hydraulic model, is then computed. Model based leakage detection techniques are employed primarily on the operational side with the objective to efficiently utilize available measurements along with the available system model to determine the system faults.

Our approach is somewhat related to Deshpande et al. (2013) and Sarrate, Nejari, and Rosich (2012), which consider pipe bursts as failure events. In Deshpande et al. (2013), detection of events in networks is studied using distance decaying sensing function. The problem is formulated as a continuous  $p$ -median facility location problem and solved using a gradient descent algorithm. However, in contrast to Deshpande et al. (2013), in which only the detection problem is considered, we consider detection as well as location identification of link failures. In Sarrate et al. (2012) both the detection and location identification of failure events are considered in the problem formulation. Their solution approach is based on the depth-first branch and bound search and is tested on a single small network.

In this work, we consider the placement of online high-rate pressure sensors. Additional surface and inline detection techniques include acoustic, umbilical, and autonomous robots. These tools are principally used to verify and pinpoint the location of the burst, their operation is typically time consuming and expensive, and they are not suitable for continuous operation (Zan

et al., 2014). Ideally, flow meters can also be used for detecting and localizing leaks in water networks. However, these are more expensive and can be typically installed on main pipelines only at the inlets of sub-networks (Narayanan, Vasan, Sarangan, & Sivasubramaniam, 2014). Furthermore most flow meters do not react instantaneously to changes in flow, hence are more suitable for persistent leaks (Puust, Kapelan, Savic, & Koppel, 2010).

**Approximation algorithms.** The sensor placement problem is not unique to the water sector and can be found in many engineering applications. Sensor placement is in essence a combinatorial optimization problem, in which a minimum number of sensors are deployed to minimize the uncertainty about the events of interest. The dominant approach is to cast the sensor placement problem as the classical *minimum set cover* (MSC) problem, in which given a set of  $n$  elements and a collection of  $m$  subsets, the goal is to select as few subsets as possible such that their union covers all elements. The MSC problem is known to be NP-hard (Moret & Shapiro, 1985). The *greedy algorithm* guarantees the best possible approximation ratio of  $(\ln n + 1)$ . A key feature in the efficient and practically feasible greedy algorithm is exploiting the submodular property, i.e. decreasing marginal utility of the objective function. Extensive literature exists on the greedy approximation for submodular functions. In Krause, Singh, and Guestrin (2008), a mutual information criterion was proposed to select the most informative sensors to monitor a spatial phenomenon modeled by a Gaussian process. The submodularity property of the criterion, as shown in Nemhauser, Wolsey, and Fisher (1978), was then exploited to obtain a polynomial time algorithm guaranteeing a constant factor approximation of the optimal sensor set.

**Model-based diagnosis.** Fault detection and identification (FDI) and consistency based diagnosis (DX) are two distinct approaches which rely on computing sets of events in a faulty system based on the discrepancies between the observed and predicted system behavior (Cordier et al., 2004). In the FDI community fault diagnosis is captured by localizing faults based on residuals that capture these faults. The problem is then to select a set of residual generators that are sensitive to the set of faults (Krysander & Frisk, 2008; Raghuraj, Bhushan, & Rengaswamy, 1999; Svärd et al., 2013). In the DX community, the diagnosis is derived by computing a set of conflicts that capture the faulty components that explain the observed failures (Abreu & van Gemund, 2009; De Kleer, 2011; Feldman, Provan, & van Gemund, 2008). To compute the minimum set of residual generators or the minimum set of conflicts, the problem often relies on the MSC or the minimum hitting set (MHS) formulation. The MSC problem is equivalent to the MHS, in which given the same input as in the MSC, the goal is to find the smallest subset of elements that *hits* (i.e. has a non empty intersection) every subset (Cordier et al., 2004).

In previous works (Krysander & Frisk, 2008; Raghuraj et al., 1999; Svärd et al., 2013) the isolation solution is obtained by first computing the set of all *pair-wise* faults from a given set of faults, and then using greedy heuristics to solve the MSC or the MHS problems. This is similar to the TLG approach described in Section 3.2. Computing all pair-wise events is the main computational bottleneck, especially when applied to large scale networks. The AG presented in Section 4 is a faster implementation of the greedy approach for the solution of the MTC. Its main feature is avoiding the transformation of the MTC to the MSC/MHS, which makes it more suitable for large-scale distributed systems, as demonstrated for Nets 11–12 in Table 2.

## 8. Conclusions and future work

In this work, we focused on the sensor placement for fault location identification in water networks. We cast the problem as the minimum test cover problem and suggested a fast solution

approach. Additionally, we tested and analyzed the solutions using multiple performance criteria for a suite of real water networks. The outcomes of our approach could provide a better diagnosis of failure events in terms of improved localization and response to failure events in operational mode, and could significantly reduce potential physical losses and service disruptions in water networks. In this work we assumed perfect sensing information, future extension will include sensor placement robust to erroneous and corrupt data.

## References

- Abreu, R., & van Gemund, A.J. (2009). A low-cost approximate minimal hitting set algorithm and its application to model-based diagnosis. In *Proceedings of the annual symposium on applied computing* (pp. 2–9).
- Allen, M., Preis, A., Iqbal, M., Stitangarajan, S., Lim, H. N., Girod, L., et al. (2011). Real time in-network monitoring to improve operational efficiency. *Journal of American Water Works Association*, 103(7), 63–75.
- Bentley, Water Hammer and Transient Analysis Software. <http://www.bentley.com/en-US/Products/HAMMER/>. Accessed: 2015-04-14.
- Berry, J., Hart, W., Phillips, C., Uber, J., & Watson, J. (2006). Sensor placement in municipal water networks with temporal integer programming models. *Journal of Water Resources Planning and Management*, 132(4), 218–224.
- Centre of Water Systems University of EXETER. <http://emps.exeter.ac.uk/engineering/research/cws/downloads/benchmarks/>. Accessed: 2015-04-14.
- Cordier, M.-O., Dague, P., Lévy, F., Montmain, J., Staroswiecki, M., & Trave-Massuyes, L. (2004). Conflicts versus analytical redundancy relations: a comparative analysis of the model based diagnosis approach from the artificial intelligence and automatic control perspectives. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, 34(5), 2163–2177.
- De Bontridder, K., Halldórsson, B. V., Halldórsson, M. M., Hurkens, C., Lenstra, J. K., Ravi, R., et al. (2003). Approximation algorithms for the test cover problem. *Mathematical Programming*, 98(1–3), 477–491.
- De Kleer, J. (2011). Hitting set algorithms for model-based diagnosis. In *22nd international workshop on principles of diagnosis, DX-11*.
- Deshpande, A., Sarma, S. E., Youcef-Toumi, K., & Mekid, S. (2013). Optimal coverage of an infrastructure network using sensors with distance-decaying sensing quality. *Automatica*, 49(11), 3351–3358.
- Dinur, I., & Steurer, D. (2014). Analytical approach to parallel repetition. In *STOC'14, Proceedings of the 46th annual ACM symposium on theory of computing* (pp. 624–633). New York, NY, USA: ACM.
- Eliades, D. G., Lambrou, T. P., Panayiotou, C. G., & Polycarpou, M. M. (2014). Contamination event detection in water distribution systems using a model-based approach. *Procedia Engineering*, 89, 1089–1096.
- Eliades, D. G., & Polycarpou, M. M. (2010). A fault diagnosis and security framework for water systems. *IEEE Transactions on Control Systems Technology*, 18(6), 1254–1265.
- Feldman, A., Provan, G.M., & van Gemund, A. (2008). Computing minimal diagnoses by greedy stochastic search. In *23rd AAAI conference on artificial intelligence* (pp. 911–918).
- Garey, M., & Johnson, D.S. (1979). *Computers and intractability: a guide to the theory of np-completeness*. WH Freeman & Co., San Francisco.
- Halldórsson, B. V., Halldórsson, M. M., & Ravi, R. (2001). On the approximability of the minimum test collection problem. In *Proceedings of the 9th annual European symposium on algorithms* (pp. 158–169). London, UK: Springer-Verlag.
- Jolly, M. D., Lothes, A. D., Bryson, S., & Ormsbee, L. (2014). Research database of water distribution system models. *Journal of Water Resources Planning and Management*, 140(4), 410–416.
- Krause, A., Leskovec, J., Guestrin, C., Vanbriesen, J., & Faloutsos, C. (2008). Efficient sensor placement optimization for securing large water distribution networks. *Journal of Water Resources Planning and Management*, 134(6), 516–526.
- Krause, A., Singh, A., & Guestrin, C. (2008). Near-optimal sensor placements in gaussian processes: Theory, efficient algorithms and empirical studies. *Journal of Machine Learning Research*, 9, 235–284.
- Krysander, M., & Frisk, E. (2008). Sensor placement for fault diagnosis. *IEEE Transactions on Systems, Man & Cybernetics, Part A (Systems & Humans)*, 38(6), 1398–1410.
- Minoux, M. (1978). Accelerated greedy algorithms for maximizing submodular set functions. In *Optimization techniques* (pp. 234–243). Springer.
- Misiunas, D. (2005). *Failure monitoring and asset condition assessment in water supply systems*. (Ph.D. thesis), Sweden: LUND University.
- Moret, B. M. E., & Shapiro, H. D. (1985). On minimizing a set of tests. *SIAM Journal on Scientific and Statistical Computing*, 6(4), 983–1003.
- Narayanan, I., Vasan, A., Sarangan, V., & Sivasubramaniam, A. (2014). One meter to find them all-water network leak localization using a single flow meter. In *Proceedings of the 13th international symposium on information processing in sensor networks* (pp. 47–58). IEEE.
- Nemhauser, G. L., Wolsey, L. A., & Fisher, M. L. (1978). An analysis of approximations for maximizing submodular set functions—I. *Mathematical Programming*, 14(1), 265–294.
- Ostfeld, A., Uber, J. G., Salomons, E., Berry, J. W., Hart, W. E., Phillips, C. A., et al. (2008). The battle of the water sensor networks (BWSN): A design challenge for engineers and algorithms. *Journal of Water Resources Planning and Management*, 134.
- Pandharipande, A., Calabrese, F., Lim, H., & Rajagopal, R. (2014). Guest editorial special issue on sensing technologies for intelligent urban infrastructures. *IEEE Sensors Journal*, 14(12), 4121–4121.

- Perez, R., Sanz, G., Puig, V., Quevedo, J., Cuguero Escofet, M. A., Nejjari, F., et al. (2014). Leak localization in water networks: A model-based methodology using pressure sensors applied to a real network in barcelona. *IEEE Control Systems*, 34(4), 24–36.
- Puust, R., Kapelan, Z., Savic, D. A., & Koppel, T. (2010). A review of methods for leakage management in pipe networks. *Urban Water Journal*, 7(1), 25–45.
- Raghuraj, R., Bhushan, M., & Rengaswamy, R. (1999). Locating sensors in complex chemical plants based on fault diagnostic observability criteria. *AIChE Journal*, 45, 310–322.
- Ragot, J., & Maquin, D. (2006). Fault measurement detection in an urban water supply network. *Journal of Process Control*, 16(9), 887–902.
- Rosich, A., Frisk, E., Aslund, J., Sarrate, R., & Nejjari, F. (2012). Fault diagnosis based on causal computations. *IEEE Transactions on Systems, Man & Cybernetics, Part A (Systems & Humans)*, 42(2), 371–381.
- Sarrate, R., Nejjari, F., & Rosich, A. (2012). Sensor placement for fault diagnosis performance maximization in distribution networks. In *20th mediterranean conference on control automation* (pp. 110–115).
- Sela Perelman, L., Abbas, W., Koutsoukos, X., & Amin, S. (2015). Sensor placement for fault location identification in water networks. arXiv preprint arXiv:1507.07134.
- Stoianov, I., Nachman, L., Madden, S., & Tokmouline, T. (2007). PIPENET a wireless sensor network for pipeline monitoring. In *Proceedings of the 6th international conference on information processing in sensor networks* (pp. 264–273). New York, NY, USA: ACM.
- Svärd, C., Nyberg, M., & Frisk, E. (2013). Realizability constrained selection of residual generators for fault diagnosis with an automotive engine application. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 43(6), 1354–1369.
- Todini, E., & Rossman, L. (2013). Unified framework for deriving simultaneous equation algorithms for water distribution networks. *Journal of Hydraulic Engineering*, 139(5), 511–526.
- Vazirani, V. (2003). *Approximation algorithms*. Springer-Verlag Berlin, Heidelberg.
- Wylie, E. B., Streeter, V. L., & Suo, L. (1993). *Fluid transients in systems*. Prentice Hall.
- Zan, T. T., Lim, H. B., Wong, K., Whittle, A. J., & Lee, B. (2014). Event detection and localization in urban water distribution network. *IEEE Sensors Journal*, 14(12), 4134–4142.
- Zheng, L., & Kleiner, Y. (2014). Computational intelligence for urban infrastructure condition assessment: Water transmission and distribution systems. *IEEE Sensors Journal*, 14(12), 4122–4133.



**Lina Sela Perelman** is a Postdoctoral Research Associate in the Department of Civil and Environmental Engineering, Massachusetts Institute of Technology, where she is affiliated with Resilient Infrastructure Networks Lab and Singapore-MIT Alliance for Science and Technology. She received her Ph.D. in Civil Engineering from the Technion—Israel Institute of Technology in 2011. Her research interests are resilient urban water networks, fault diagnostics and prediction, and systems management under uncertainty.



**Waseem Abbas** received B.Sc. degree in Electrical Engineering from University of Engineering & Technology, Lahore, Pakistan in 2007. He received his M.S. and Ph.D. degrees, both in Electrical and Computer Engineering, from Georgia Institute of Technology, Atlanta, GA in August 2010 and December 2013 respectively, as a Fulbright student. Currently, he is a postdoctoral research scholar at the Institute for Software Integrated Systems, Vanderbilt University, Nashville, TN. His research interests are in the area of network control systems, graph-theoretic methods for large networked systems, and cyber-physical systems.



**Xenofon Koutsoukos** is a Professor in the Department of Electrical Engineering and Computer Science at Vanderbilt University. He is also a Senior Research Scientist in the Institute for Software Integrated Systems (ISIS). Before joining Vanderbilt, Dr. Koutsoukos was a Member of Research Staff in the Xerox Palo Alto Research Center (PARC) (2000–2002), working in the Embedded Collaborative Computing Area. He received his Ph.D. in Electrical Engineering from the University of Notre Dame in 2000. His research work is in the area of cyber-physical systems with emphasis on formal methods, distributed algorithms, security and resilience, diagnosis and fault tolerance, and adaptive resource management. He has published numerous journal and conference papers and he is co-inventor of four US patents. He was the recipient of the NSF Career Award in 2004, the Excellence in Teaching Award in 2009 from the Vanderbilt University School of Engineering, and the 2011 NASA Aeronautics Research Mission Directorate (ARMD) Associate Administrator (AA) Award in Technology and Innovation.



**Saurabh Amin** is Robert N. Noyce Career Development Assistant Professor in the Department of Civil and Environmental Engineering, Massachusetts Institute of Technology (MIT). His research focuses on the design and implementation of high confidence network control algorithms for infrastructure systems. He works on robust diagnostics and control problems that involve using networked systems to facilitate the monitoring and control of large-scale critical infrastructures, including transportation, water, and energy distribution systems. He also studies the effect of security attacks and random faults on the survivability of networked systems, and designs incentive-compatible control mechanisms to reduce network risks. Dr. Amin received his Ph.D. in Systems Engineering from the University of California, Berkeley in 2011. His research is supported by NSF CPS FORCES project, NSF CAREER award, Google Faculty Research award, and Siebel Energy Institute Grant.