January 2015

ADVANCED MODELING AND EFFICIENT OPTIMIZATION METHODS FOR REAL-TIME RESPONSE IN WATER NETWORKS

Arpan Seth
Purdue University

Graduate School Form 30
Updated 1/15/2015

PURDUE UNIVERSITY
GRADUATE SCHOOL
Thesis/Dissertation Acceptance

This is to certify that the thesis/dissertation prepared

By Arpan Seth

Entitled
ADVANCED MODELING AND EFFICIENT OPTIMIZATION METHODS FOR REAL-TIME RESPONSE IN WATER NETWORKS

For the degree of Doctor of Philosophy

Is approved by the final examining committee:

Carl D. Laird
Chair

Dulcy M. Abraham

Joseph F. Pekny

Zoltan K. Nagy

To the best of my knowledge and as understood by the student in the Thesis/Dissertation Agreement, Publication Delay, and Certification Disclaimer (Graduate School Form 32), this thesis/dissertation adheres to the provisions of Purdue University's "Policy of Integrity in Research" and the use of copyright material.

Approved by Major Professor(s): Carl D. Laird

Approved by: John A. Morgan 12/03/2015

Head of the Departmental Graduate Program Date

ADVANCED MODELING AND EFFICIENT OPTIMIZATION METHODS FOR REAL-TIME RESPONSE IN WATER NETWORKS

A Dissertation

Submitted to the Faculty

of

Purdue University

by

Arpan Seth

In Partial Fulfillment of the

Requirements for the Degree

of

Doctor of Philosophy

December 2015

Purdue University

West Lafayette, Indiana

To my family for their unconditional love and support.

ACKNOWLEDGMENTS

One of the earliest childhood memories I have is from when I was in 2nd grade. I remember receiving my final grades at the end of the year. I had performed disappointingly and was ranked 37th out of the 45 students in my class. I walked out of the class where my dad was eagerly waiting to hear the good news. With my head held low, I told my dad my rank in the class. What came next is what I believe had a big impact on my life. My dad was genuinely happy and excited that I had passed all my exams. He didn't care where I stood in my class. That was the easy part and I would expect many parents to say that to show encouragement. But he then went on to tell everyone in my extended family, his friends, and our neighbors how he was so happy and proud of me. From that day onwards, I have moved forward in my life knowing that the unconditional love and support of my parents, Mr Praveen Kumar Seth and Mrs. Chhavi Seth, will always be there to guide me. They were always very supportive of me during my years in graduate school and I will forever be thankful for it.

A similar moment came 16 year later in my 2nd year of graduate school. I was the teaching assistant (TA) for a course taught by my advisor Dr. Carl Laird. By the end of the semester, I felt like I had done an average job as a TA, making several mistakes during the way. But I remember the last day of the course when we had a class competition and the winners were given prizes. Dr. Laird gave me a prize too along with a card saying that he appreciated my efforts during the past semester. It was a simple gesture that meant so much to me. In that moment, I remember thinking about 2nd grade and my dad. Words cannot describe how thankful I am to Dr. Laird for everything he has given me. He has gone out of his way to be a

patient and encouraging advisor and a caring and thoughtful friend. I will forever be indebted to him for not only the knowledge he has imparted to me but also for making me a better person.

I would like to thank my committee members, Dr. Joseph Pekny, Dr. Zoltan Nagy, and Dr. Dulcy Abraham for their valuable comments and suggestions on my research. I am extremely grateful to all my co-authors and collaborators from Sandia National Laboratories (Katherine Klise and John Siirola) and the U.S. Environmental Protection Agency (Terra Haxton and Regan Murray) for their contributions to my projects. I am especially thankful to Gabriel Hackebeil for his invaluable inputs to this work.

Every member of the Laird Research Group, former or present, has been a good friend and has played an important part in my journey through graduate school. I thank Angelica Mann, Daniel Word, Sean Legg, and Jia Kang for making me feel welcome and answering all my questions when I joined the group. I also thank Yankai Cao, Shawn McGee, Alberto Benavides, Jianfeng Liu, Michael Bynum, Jose Rodriguez, and Todd Zhen not only for being the best colleagues one could hope for, but also for being great friends and making everyday fun. I wish all of you the best in your lives.

Finally, I would like to thank every member of my family for making me the person who I am today. My grandfather, Dr. Jagmohan Seth, for all his wisdom and love throughout my childhood; my sister, Anvita Seth, for her belief in me; my aunt and uncle, Renu and Alok Seth, for being my parents in the new world; and all my cousins, Prateek and Shaily Mehrotra, Samarth and Tanu Mehrotra, Amit and Saumya Arora, Rajat and Divya Diwan, Sagar and Tanu Seth, Shubham Seth and Shivam Seth, for making life worth living.

LIST OF PUBLICATIONS AND PRESENTATIONS

Publications

- Seth A., Klise, K.A., Siirola, J.D., Haxton, T., Laird, C.D., **Testing Contamination Source Identification Methods for Water Distribution Networks**, accepted by Journal of Water Resources Planning and Management, 2015.

- Seth, A., Hackebeil, G.A., Klise, K.A., Haxton, T., Murray, R., and Laird, C.D., **Efficient Reduction of Optimal Disinfectant Booster Station Placement Formulations for Security of Large-Scale Water Distribution Networks**, submitted to Computational Optimization and Applications, 2015.

- Seth, A., Hackebeil, G.A., Haxton, T., Murray, R., Laird, C.D., and Klise, K.A., **Evaluation of Chlorine Booster Station Placement for Water Security**, submitted to Journal of Water Resources Planning and Management, American Society of Civil Engineers, 2015.

- Klise, K.A., Siirola, J.D., Hart, D.B., Hart, W.E., Phillips, C.A., Haxton, T., Murray, R., Janke, R., Taxon, T., Laird, C.D., Seth, A., Hackebeil, G., McGee, S., Mann, A.V., **Water Security Toolkit User Manual Version 1.1**, Sandia National Laboratories, SAND2013-8346P, 2013.

- Hackebeil, G.A., Laird, C.D., Seth A., Watson, J.P, Woodruff, D.L., **A methodology for determining the appropriate number of contamination events in sensor placement optimization for contamination warning system design**, submitted to Operations Research, 2015.

- Cao, Y., Seth, A., and Laird, C.D., **A parallel augmented lagrangian interior-point approach for large-scale NLP problems on graphics processing units**, accepted by Computers and Chemical Engineering, 2015.

Conference Presentations

- Seth, A., Hackebeil, G.A., Klise, K.A., Haxton, T., Murray, R., and Laird, C.D., **Solution of large scale stochastic programming problems for optimal placement of booster stations in water networks** AIChE Annual Meeting, Salt Lake City, UT, November, 2015.

- Seth, A., and Laird, C.D., **Optimization application in water distribution networks using embedded hydraulic models**, AIChE Midwest Regional Conference, Chicago, IL, March, 2015.

- Seth, A., and Laird, C.D., **Solution of mixed-integer nonlinear programming problems in hydraulics of water distribution networks**, AIChE Annual Meeting, Atlanta, GA, November, 2014.

- Seth, A., McGee, S., McKenna, S., Hart, D.B., and Laird, C.D., **An application for real-time response to contamination events in a large-scale public water network**, AIChE Annual Meeting, San Francisco, CA, November, 2013.

- Seth, A., Word, D.P., Kang, J., Cummings, D., Laird, C.D., **NLP approaches for estimation of seasonal transmission parameters in childhood infectious diseases**, INFORMS Annual Meeting, Minneapolis, MN, October 8, 2013.

- Seth, A., Mann, A., Klise, K., Haxton, T., and Laird, C.D., **Development of a testbed for contaminant source inversion methods**, American Society of Civil Engineers, EWRI, Cincinnati, OH, May, 2013.

- Seth, A., Mann, A., McGee, S., and Laird, C.D., **A real-time source inversion and optimal sampling strategy for large-scale drinking water distribution systems**, 13th INFORMS Computing Society Conference, Santa Fe, NM, January, 2013.

- Klise, K.A., Seth, A., Laird, C.D., and Murray, R., **Resilience evaluation for water distribution networks**, EWRI 2015, Austin, TX, May, 2015.

- Klise, K., Laird, C.D., Hackebeil, G., Seth, A., Murray, R., and Haxton, T. , **Evaluation of booster station placement for water quality and water security concerns**, EWRI 2013, Cincinnati, OH, May, 2013.

TABLE OF CONTENTS

LIST OF TABLES

Table Page

LIST OF FIGURES

Figure Page

Figure Page

ABSTRACT

Seth, Arpan PhD, Purdue University, December 2015. Advanced Modeling and Efficient Optimization Methods for Real-Time Response in Water Networks . Major Professor: Carl D. Laird.

In response to a contamination incident in water distribution networks, effective mitigation procedures must be planned. Disinfectant booster stations can be used to neutralize a variety of contaminant and protect the public. In this thesis, two methods are proposed for the optimal placement of booster stations. Since the contaminant species is unknown a priori, these two methods differ in how they model the unknown reaction between the contaminant and the disinfectant. Both methods employ Mixed-Integer Linear Programming to minimize the expected impact over a large set of potential contamination scenarios that consider the uncertainty in the location and time of the incident. To make the optimal booster placement problem tractable for realistic large-scale networks, we exploit the symmetry in the problem structure to drastically reduce the problem size. The results highlight the effectiveness of booster stations in reducing the overall impact on the population, which is measured using two different metrics - mass of contaminant consumed, and population dosed above a cumulative mass threshold. Additionally, we also study the importance of various factors that influence the performance of disinfectant booster stations (e.g., sensor placement, contaminant reactivity and toxicity, etc.).

The booster station placement is performed at the planning stage. Once a contamination incident has taken place, knowledge of the contamination source location is important to inform the control and cleanup operations. Since this source identification problem needs to be solved in real time, computational speed on large-

scale networks is of utmost importance. With this in mind, we propose a Bayesian probability-based method for source identification and a greedy algorithm for selecting manual grab sample locations. Measurements obtained from the selected manual sampling location can be used by the source identification method to further narrow the possible set of source locations. Indeed, the case study performed on a large-scale network (with over 12,000 nodes) highlights the computational speed of the proposed techniques, where both the source identification and sampling location calculations can be performed within seconds.

Various source identification strategies that have been developed by researchers differ in their underlying assumptions and solution techniques. In this work, we present a systematic procedure for testing and evaluating source identification methods. The performance of these source identification methods is affected by various factors including: size of water distribution network model, measurement error, modeling error, time and number of contaminant injections, and time and number of measurements. This work includes test cases that vary these factors and evaluates the proposed Bayesian probability-based source identification method along with two other methods from the literature. The tests are used to review and compare these different source identification methods, highlighting their strengths in handling various identification scenarios.

1. INTRODUCTION [1]

Public water distribution networks are critical infrastructures in the modern world. According to the Organization for Economic Co-operation and Development's environmental outlook report (OECD, 2012), by year 2050, the global demand for water is expected to increase by 130% for domestic use, 140% for electricity use, and 400% for manufacturing use. Simultaneously, the fast growth of large urban centers is going to require significant expansion in the existing public water distribution networks. As these networks become larger and more complex, advanced modeling and efficient optimization techniques are necessary to help design and operate these networks, and secure them against harmful contamination incidents.

Water distribution networks are large complex systems with many access points, leading to the potential for accidental or intentional contamination. Rapid response and mitigation of contamination incidents requires a three-part approach. First, improve security at network interface points (e.g., physical security at treatment plants and storage tanks, and backflow preventers at customer interfaces). Second, implement an event detection system (EDS) that includes contamination sensors to rapidly alert system operators to the presence of contamination. Third, develop response

[1]Part of this section is reprinted with permission from "Testing Contamination Source Identification Methods for Water Distribution Networks" by Seth, A., Klise, K.A., Siirola, J.D., Haxton, T., and Laird, C.D., 2015. to appear in Journal of Water Resources Planning and Management, Copyright 2015 by American Society of Civil Engineers.
Part of this section is reprinted from "Efficient Reduction of Optimal Disinfectant Booster Station Placement Formulations for Security of Large-Scale Water Distribution Networks" by Seth, A., Hackebeil, G.A., Klise, K.A., Haxton, T., Murray, R., and Laird, C.D., 2015. Submitted to Computational Optimization and Applications.
Part of this section is reprinted from "Evaluation of Chlorine Booster Station Placement for Water Security" by Seth, A., Hackebeil, G.A., Haxton, T., Murray, R., Laird, C.D., and Klise, K.A., 2015. Submitted to Journal of Water Resources Planning and Management, American Society of Civil Engineers.

plans with a goal to rapidly contain and remove contamination from the system using actions like closing isolation valves, flushing network pipes, or injecting disinfecting agents. This work focuses on developing modeling and optimization methods for planning real-time response strategies to contamination incidents in water distribution networks.

1.1 Real-time response to contamination incidents

Early-warning detection systems can be used to identify the presence of contaminant using a fixed grid of sensors throughout the network. Berry et al. (2005b); Ostfeld and Salomons (2004a); Murray et al. (2010b) have extensively studied the problem of optimal sensor layouts within these drinking water distribution systems. However, adequate emergency response mechanisms must also be developed. In this work, we study two important response actions that can be critical in reducing the impact of potential contamination incidents:

- A typical response to a detected contaminant from an early warning system includes laboratory confirmation. A manual water sample will be drawn and sent for laboratory analysis. Following analysis (which can take several hours or more), a positive confirmation of contaminant will likely result in a no-drink order. However, during the time between the first detection and the laboratory confirmation, contaminant continues to travel and spread through the network. Disinfectant booster stations can help mitigate the effect of potential contamination by injecting additional (but safe) amounts of disinfectant immediately following the initial warning (Parks and VanBriesen, 2009). Moreover, intelligent placement of these booster stations can help in efficiently providing incident response.

- Once a contamination incident has occurred, real-time response strategies can include closing valves to isolate contaminated parts of the network, and opening selected fire hydrants to flush the contaminated water out of the network. For

these type of response actions to be more effective, an accurate understanding of the extent of the contamination plume within the WDN is necessary; while estimating the plume extent requires having an accurate real-time model of the network and knowledge of the contamination source. Therefore, the accuracy of a real-time source identification method can be crucial for efficient response actions.

In this work, we propose modeling and optimization techniques to efficiently solve the above two problems. Optimization provides a great tool for system design, operation, and real-time response planning problems. Hence, a large body of work is dedicated to using modeling and optimization techniques to solve a wide variety of problems related to water distribution networks. To understand the challenges associated with the two problems addressed in this work, we first provide a brief overview the landscape of modeling and optimization problems related to water distribution networks.

1.2 Modeling and optimization problems in water distribution networks

Water distribution systems are typically modeled as a network of nodes and links where nodes include reservoirs, storage tanks, and junctions, while links include pipes, pumps, and valves. Mass and energy conservations laws are then used to derived first-principles models. Typically, the injection and flow of chemical or biological species can be assumed to have negligible impact on the water flow rates and pressures in the network, and therefore, the chemical/biological species mass balances can be decoupled from the conservation laws describing the flow of water. The set of equation describing the water flow rates and system pressures are referred to as the *Hydraulic Model*. Information calculated from the hydraulic model can be used as input parameters to write species mass balances that make up the so called *Water Quality Model*. Next, we present the equation that describe these two models, highlighting their

key characteristics that need to be considered when using them in a mathematical programming framework.

1.2.1 Hydraulic Model

The hydraulic model is primarily composed of mass balances at nodes, pressure drop equations in pipes, pressure gain equation in pumps, and level dynamics in tanks. The mass balances at all junctions and tanks are given by

$$\sum_{p \in N_{in}} Q_{p,t} - \sum_{p \in N_{out}} Q_{p,t} = \mathbf{D}_{n,t}, \qquad \forall\, n \in JN,\, t \in T, \tag{1.1}$$

$$\sum_{p \in N_{in}} Q_{p,t} - \sum_{p \in N_{out}} Q_{p,t} = Q^{IN}_{n,t}, \qquad \forall\, n \in TN,\, t \in T, \tag{1.2}$$

where JN, TN, and T are set of junction nodes, tank nodes, and time steps being considered in the model. $Q_{p,t}$ represents the volumetric flow rate of water in a link p at a time step t. Link p can belong to a predetermined set of input links, N_{in}, or output links N_{out} from a node n. $\mathbf{D}_{n,t}$ represents consumer demands at junctions that are known inputs to the model. The net volumetric flow rate into a tank is denoted by the variable $Q^{IN}_{n,t}$, which is used to calculate the change in pressure head (or level) in the tank by using explicit Euler discretization of tank dynamics, $AdH_n/dt = Q^{IN}$, as follows:

$$H_{n,t} - H_{n,t-1} = \frac{1}{A} Q^{IN}_{n,t-1}, \qquad \forall\, n \in TN,\, t \in \hat{T} \tag{1.3}$$

where $H_{n,t}$ is the water head in tank n at time t. For simplicity, here we assume a constant cross-sectional area for the tanks, A. $\hat{T}$ represents the set of all time steps excluding the first time step.

Next, the head (or pressure) loss inside the pipes due to friction from the pipe walls is typically modeled using one of three different formulas proposed in the literature:

1. Hazen-Williams formula

2. Darcy-Weisbach formula

3. Chezy-Manning formula

The general equation to calculate the head loss inside pipes is given by

$$H_{n_s,t} - H_{n_e,t} = KQ_{p,t}^{C}, \qquad \forall\, p \in P,\, t \in T \tag{1.4}$$

where n_s and n_e represent the start and end node of pipe p respectively. Similarly, $H_{n_s,t}$ and $H_{n_e,t}$ represents the head at the start and end node of the pipe p. K is called the resistance coefficient and C is called the flow exponent, and they can be calculated using any of the thee head loss formulas from above. The resistance coefficient depends on the material, length, diameter, and friction factor of the pipe along with the type of flow regime. The flow exponent is either 1.852 (Hazen-Williams) or 2.0 (Darcy-Weisbach or Chezy-Manning). Note that the above head loss equation is one of the major sources of nonlinearity in the hydraulic model.

Pumps are often used to provide additional hydraulic head that is necessary to fill storage tanks. They can either be constant energy devices or have variable speed settings. The equations describing the head gain provided by pump are typically nonlinear. A common form of the head gain equation is given by

$$H_{n_s,t} - H_{n_e,t} = \alpha - \beta Q_{pu,t}^{\gamma}, \qquad \forall\, pu \in PU,\, t \in T \tag{1.5}$$

where n_s and n_e represent the start and end node of pump pu respectively. Similarly, $H_{n_s,t}$ and $H_{n_e,t}$ represents the head at the start and end node of the pump pu. α, β, and γ are characteristic parameters for a particular pump.

Additionally, various types of valves can also be included in the hydraulic model that can vary in their modeling complexity from being simple on/off switches like Shutoff or Check Valves to more complicated pressure reducing or general purpose valve that can have a nonlinear flow-head relationship.

To perform a hydraulic simulation, the inputs to the hydraulic model typically include: network connectivity structure, time varying nodal demands, initial tank and reservoir heads, and pipe, pump, and tank parameters. The nonlinear set of equations described above can then be used to calculate flow rates in all links and hydraulic heads at all nodes over a simulation duration. One caveat in running hydraulic simulation is that often discrete decisions need to be taken at certain time of day or when certain pressure of flow conditions are reached. These decisions are referred to as "controls." For example, pumps providing hydraulic head to a tank have to be turned off when a maximum level in the tank is reached. These controls are typically handled in an event based simulation environments.

1.2.2 Water Quality Model

Dynamic water quality models are used to track the flow of a chemical or biological species through the water distribution network. These models can be classified as either Eulerian or Lagrangian (Rossman and Boulos, 1996). Eulerian models divide the pipes into spacial elements of fixed size and track concentration changes inside and at their boundaries over time. Lagrangian models track discrete packets or parcels of water and their concentrations as they move through the pipes. The water quality model used throughout this work is based on a Lagrangian approach that was originally proposed by (Laird et al., 2005) and later extended by (Mann et al., 2012a).

The first set of equations in the water quality model are the species mass balances at the junctions and tanks:

$$c_{n,t} = \frac{\sum_{p \in N_{out}} Q_{p,t} \hat{c}^{O}_{p,t} - \sum_{p \in N_{in}} Q_{p,t} \hat{c}^{I}_{p,t} + m_{n,t}}{\sum_{p \in N_{out}} Q_{p,t} - \sum_{p \in N_{in}} Q_{p,t} + Q^{ext}_{n,t}}, \qquad \forall\, n \in JN,\, t \in T \tag{1.6}$$

$$V_{n,t}\frac{dc_{n,t}}{dt} = \sum_{p\in N_{out}} Q_{p,t}\hat{c}^{O}_{p,t} - \sum_{p\in N_{in}} Q_{p,t}\hat{c}^{I}_{p,t} + m_{n,t} - \left[\sum_{p\in N_{out}} Q_{p,t} - \sum_{p\in N_{in}} Q_{p,t} + Q^{ext}_{n,t}\right] c_{n,t}, \qquad \forall\, n \in TN,\, t \in T \tag{1.7}$$

where $c_{n,t}$ is the species concentration at the junction or tank node n at time t. $\hat{c}^{O}_{p,t}$ and $\hat{c}^{I}_{p,t}$ are the species concentrations at the outlet and inlet of pipe p at time t respectively. $m_{n,t}$ is the mass of species entering node n at time t from and external source (i.e., a mass injection). Similarly, $Q^{ext}_{n,t}$ is an external volumetric flow rate of water entering the node. $V_{n,t}$ is the volume of tank n at time t. All flow rates and tank volumes calculated using the hydraulic model can be assumed to be constant over a time step and then used as inputs to the above equations. Therefore, Equations 1.6 and 1.7 are linear in terms of the concentration variables $c_{n,t}$, $\hat{c}^{O}_{p,t}$, and $\hat{c}^{I}_{p,t}$.

The remaining equations in the water quality model describe the species concentration gradient inside pipes. Assuming plug flow with instantaneous cross-sectional mixing and negligible longitudinal dispersion, species concentration inside a pipe is described by the following partial differential equation:

$$\frac{\delta\hat{c}_p(x,t)}{dt} + u_p(t)\frac{\delta\hat{c}_p(x,t)}{dx} = 0, \qquad \forall\, p \in P \tag{1.8}$$

where $\hat{c}_p(x,t)$ represents species concentration along the pipe p at displacement x and time t. u_p is the longitudinal velocity of water inside pipe p, which can also be calculated from the hydraulic simulations.

Discretizing Equations 1.6 and 1.7, and using the origin tracking algorithm proposed by (Laird et al., 2005; Mann et al., 2012a) to replace Equation 1.8, the water quality model can be described as a set of linear equations that provide an input-

output relationship between species mass injections at all nodes and time steps to species concentrations at all nodes and time steps:

$$\mathbf{Gc} = \mathbf{Dm} \tag{1.9}$$

where $\mathbf{G}$ and $\mathbf{D}$ are coefficient matrices. $\mathbf{c}$=[...$c_{n,t}$...], $\forall n \in N, t \in T$ is the vector of concentrations at all nodes in the set N and all time steps in the set T. $\mathbf{m}$=[...$m_{n,t}$...], $\forall n \in N, t \in T$ is the vector of mass injections at all nodes and time steps.

The water quality model in Equation 1.9 is available in U.S EPA's Water Security Toolkit (WST) (EPA, 2014) under the Merlion package, and is used extensively in the modeling and optimization formulations proposed in this work. Apart from the assumptions already discussed, the following simplifications are also made. Pumps and valves are modeled as zero-length pipes with inlet and outlet concentrations that are the same. Mixing at all nodes is assumed to be complete and instantaneous. Although, not included in the above equations, Merlion can also support first-order linear decay.

1.2.3 Classification of Optimization Problems

Water utilities can use accurate water network models as a valuable tool for many applications that facilitate safe and efficient delivery of clean drinking water to the public. These applications can be divided into three major categories: (1) Design, (2) Operations, and (3) Safety and security. Optimization methods have been widely used at the planning stage to design networks that are both cost-effective and robust (Eusuff and Lansey, 2003; Geem, 2009; Cunha and Sousa, 1999; Vasan and Simonovic, 2010; Zecchin et al., 2007). Minimizing operating cost associated with maintaining pressure and water quality requirements has also been a major area of study (Jowitt and Germanopoulos, 1992; Mackle et al., 1995; Yu et al., 1994; Van Zyl et al., 2004; Constans et al., 2003; Munavalli and Kumar, 2003; Boccelli et al., 1998). The com-

plexity of optimization problems arising in water distribution networks depend on three major factors: (1) Modeling requirements, (2) Scaling with network size, and (3) Type of problem. Table 1.1 categorizes a variety of optimization problems based on these three factors. It should be pointed out that in this table, we are categorizing mathematical programming formulations of these problems that embed the hydraulic or water quality model directly into the formulation. We do not consider methods that use a simulation engine as a black-box linked to an external optimization routine.

Table 1.1: Classification of optimization problems in water distribution networks

Example Problems	Modeling Requirements	Scaling	Common Problem Class
Sensor Placement, Booster Placement	Pre-simulation of Scenarios	Node × Time	MILP
Source Identification, Booster Placement	Embedded Water Quality Model	Node × Space × Time	MILP NLP
Pump Scheduling, Pressure Management Infrastructure Sizing	Embedded Hydraulic Model	Node × Time	NLP MINLP
Hydrant Flushing, Contaminant Control	Embedded Hydraulic & Water Quality Model	Node × Space × Time	NLP MINLP

The first category is composed of problems like optimal sensor placement where we are placing water quality sensors to detect contamination incidents and minimize impact over a large set of possible contamination scenarios. In these type of problems, the decision variables (e.g., sensor locations) have no impact on the scenarios, and therefore, we can pre-simulate the scenarios to generate data, which can then be used in an optimization formulation. The second category is composed of problems like source identification, where we can use historical hydraulic information as input to build a water quality model, which can then be embedded into an optimization

formulation. On the other hand, there are problems that only deal with the hydraulics of the network, and these include operational problems like pump scheduling to minimize electricity cost or design problems like valve placement to manage pressure requirements. Finally, there are more challenging problems like hydrant flushing that requires selection of hydrant locations to flush contaminated water out of the network as efficiently as possible. These type of problems require manipulation of the network hydraulics in order to improve the water quality. Therefore, for such problems we need to embed both the hydraulic and water quality model into the optimization formulation.

The second factor that influences the problem complexity is how the problem scales with the size of a network. The optimization formulations that have an embedded water quality model, involve tracking concentrations at not only the nodes and time steps, but can also have a concentration gradient inside the pipes. And therefore, these problems can have an extra spacial component that can grow significantly with network size.

The class of mathematical programming problem that needs to be solved plays a major part in problem tractability, especially for large-scale network models. Formulations with an embedded hydraulic model typically require solving Nonlinear Programming (NLP) or Mixed-Integer Nonlinear (MINLP) problems due to nonlinear pressure-flow relationships in pipes and pumps (Equation 1.4 and 1.5). Integer variables in these formulations arise naturally from discrete decisions like available pipe diameters for the network design problem or selected hydrant flushing nodes for the hydrant flushing problem. Since the water quality can be modeled as a linear system (Equation 1.9), most problems that only require an embedded water quality model can be formulated as Mixed-Integer Linear Programming (MILP) or NLP problems. Keeping the current solver technology in mind, in general MINLP problems are significantly more challenging to solve compared to MILP or NLP problems even for small-scale networks.

The two problems that we address in this work - optimal booster station placement and contaminant source identification - can be formulated as MILP problems that may or may not require an embedded water quality model. As we will show through the case studies on a range of network sizes, these MILP problems can very easily become intractably large. In this work, we propose efficient solution methods that take advantage of the problem structure to make these problems tractable for large-scale networks. The proposed techniques set a precedence for custom solution methods that will need to be explored for the more challenging NLP or MINLP problems in the future.

1.3 Booster chlorination for incidence response

Chlorine booster stations are commonly used in water distribution networks to maintain drinking water standards because chlorine degrade as it reacts with microbes and other chemicals as it moves through the system. Booster stations are designed to inject chlorine at strategic locations, helping to maintain residual levels that can prevent pathogen re-growth. Chlorine booster stations are typically installed at pump stations or other facilities but could also be added throughout the water distribution system. Several optimization methods have been suggested to place booster stations and to schedule booster operations for water quality objectives (Boccelli et al., 1998; Kang and Lansey, 2010; Lansey et al., 2007; Munavalli and Kumar, 2003; Ostfeld and Salomons, 2006; Ozdemir and Ucaner, 2005; Prasad et al., 2004; Propato and Uber, 2004a,b; Tryby et al., 2002; Uber et al., 1998).

Disinfectant booster stations can also be used as a first line of response to a contamination incident. In the event of a contamination incident, an effective emergency response plan could include injecting chlorine at fixed booster locations to inactivate or destroy a potentially harmful contaminant. Unlike booster station placement for water quality objectives, optimal booster station placement for water security should take into account a wide range of possible contamination injection scenarios. Another

major source of uncertainty associated with this problem is the unknown reaction between the chlorine and a contaminant species that is not known a priori.

In one optimal booster placement approach proposed by Ostfeld and Salomons (2006), two different objective functions are recommended. The first objective, Min Cost, minimizes the overall cost of pumping and disinfection. This objective is designed to solve the residual maintenance problem. The second objective, Max Protection, maximizes the disinfectant concentration at all consumption nodes while maintaining acceptable upper bounds. The authors note that this objective can be used as a response to a contamination incident, however, the uncertainty in the contamination location and time is not considered.

One of the biggest challenges associated with the booster placement problem for water security is modeling the reaction kinetics between chlorine and an unknown contaminant species. Booster stations are only effective for response to water contamination incidents if the contaminants ability to cause harm can be reduced by chlorine. Many biological contaminants are inactivated in the presence of sufficient chlorine; meaning that they are killed or damaged to the extent they cannot cause human disease or death. Some chemical contaminants are oxidized in the presence of chlorine, reducing the toxicity of the contaminant. However, dangerous byproducts might be formed in reactions with chlorine. For example, chlorine can react with some organophosphate pesticides to form oxons, which might be more toxic than the original compound. Understanding these complex reactions is critical in order to estimate the benefits of booster stations in the context of water security. However, with limited knowledge at the planning stage, reasonable assumptions can be made to approximate the unknown reaction kinetics. Additionally, during a real-world contamination incident, the contaminant species is typically unknown at the time of detection. Current contamination detection technologies rely on standard water quality parameters (i.e., pH, turbidity, residual chlorine), which do not indicate the type of contaminant in the network. For this reason, the exact reaction kinetics between the contaminant and chlorine cannot generally be modeled at the time of response. In

this work, we propose two different MILP formulations to place booster stations for the security problem. The formulations differ how they model the unknown chlorine-contaminant reaction. These modeling assumptions have a major impact on the size of the networks that are tractable for these formulations. Both formulations consider uncertainty in the location and time of a contamination incident and our results show that the optimal booster placement obtained can significantly reduce the expected impact.

1.4 Contamination source identification

Identifying the source of a contamination incident is a critical step towards planning the cleanup and control operations. The source identification problem is typically formulated as an inverse problem with the objective to find the source location of a contamination incident using the limited measurement data available from a sparse set of water quality sensors. Several researchers have proposed different methods to solve this problem.

Early work assumed the availability of contaminant concentration measurements from water quality sensors (Shang et al., 2002a; Laird et al., 2005, 2006; Preis and Ostfeld, 2006). Since the contaminant species is not known *a priori* (chemical or biological), recent developments in contamination detection technology utilize fault detection approaches by monitoring standard water quality measures (e.g., pH, free chlorine, turbidity, conductivity) to provide a binary yes/no indication of the presence or absence of contamination in the network (EPA, 2010a). Therefore, recent source identification methods proposed in the literature incorporate these type of measurements (De Sanctis et al., 2008; Zechman and Ranjithan, 2009; Mann et al., 2012). Additionally, a variety of statistical approaches have also been proposed that consider measurement error (Liu et al., 2011; Perelman and Ostfeld, 2012; Wagner and Neupauer, 2013; Wang and Harrison, 2012).

In a real-time response scenario, on-line computational efficiency of a source identification method is crucial to facilitate quick response actions. With this goal in

mind, we propose a Bayesian probability-based method that takes advantage of the Merlion water quality modeling framework (Mann et al., 2012a) to perform source identification on large-scale water networks within seconds. Additionally, we also propose a fast greedy algorithm for the selection of manual sampling locations to further assist in the source identification.

Given the diversity of source identification methods proposed in the literature, there is the need for a common set of tests to evaluate their performance. Thus, in this work we present a testing methodology for source identification techniques. This methodology includes a comprehensive set of potential contamination scenarios designed to cover a wide variety of factors that impact the effectiveness of source identification techniques. Using this testing methodology, the proposed Bayesian probability-based source identification method is compared to two other techniques from the literature.

1.5 Dissertation outline

The outline of this thesis is as follows. In Chapter 2, we introduce the optimal booster station placement problem and provide a background of different methods proposed in the literature. In Chapter 3, we propose a modeling technique for the disinfectant booster placement problem that simplifies the reaction kinetics between chlorine and an unknown contaminant. We assume that the chlorine instantaneously and completely neutralizes a contaminant on contact. This assumption gives us the ability to pre-simulate a large number of contamination and chlorine injection scenarios and use the resulting data in an MILP formulation. For large-scale network the original formulation is intractably large, and therefore, we propose three reductions that decrease the size of these problem by up to five orders of magnitude. This modeling and optimization technique for placing booster stations is referred to as the *Neutralization method.*

In Chapter 4 we propose another formulation for the booster station placement problem that lets us model different levels of contaminant reactivities by embedding

the water quality model directly into an MILP formulation. This optimal booster placement method is referred to as the *Limiting Reagent method.* Similar to the Neutralization method, it assumes that the chlorine reacts instantaneously with the contaminant, however, the reaction happens with respect to a stoichiometric ratio. This chapter also provide several case studies that evaluate the performance of the booster placements obtained using the two methods.

Booster chlorination can be used as a first line of defense to protect the public against potential contamination. However, as a contamination incident unfolds, a more targeted response requires identification of the source of the contamination as quickly as possible. In Chapter 5, we define the source identification problem and review the different classes of source identification methods proposed in the literature. In Chapter 6, we propose a Bayesian probability-based source identification method that identifies probable contamination source location upstream from the sensor locations. This method takes advantage of fast water quality simulations using Merlion and several code optimizations that result in accurate source identification within seconds for large-scale networks. Additionally, a greedy algorithm for selecting manual sampling locations is proposed that is based on the optimization technique presented by Wong et al. (2010).

Chapter 7 provides a testing methodology to compare the performance of source identification methods under realistic scenarios (e.g., measurement and modeling error). In Chapter 8, the proposed methodology is used to compare three source identification methods highlighting the advantages and disadvantages of each.

Finally, Chapter 9 concludes this thesis with a summary and future research directions.

2. BOOSTER CHLORINATION IN WATER DISTRIBUTION NETWORKS

One of the first responses to a detected contamination incident from an early warning system is laboratory confirmation. A manual water sample would be drawn and sent for laboratory analysis. Following analysis (which can take several hours or more), a positive confirmation of contaminant would likely result in a "Do Not Drink" order. However, during the time between the first detection and the laboratory confirmation, the contaminant would continue to travel and spread through the network.

Disinfectant booster stations can help mitigate the effect of some type of contamination by injecting additional (yet within acceptable range) amounts of disinfectant into the water distribution network immediately following the initial warning (Parks and VanBriesen, 2009). Moreover, intelligent placement of these booster stations can improve the efficiency of incident response. Optimal booster station placement can be used to meet two primary objectives. First, following initial disinfection at the main treatment facility, as the water flows through the water distribution network, booster stations can be used to maintain specified disinfectant levels in the water. Second, booster stations can also be used to raise the disinfectant concentration in the water (within acceptable range) in response to a contamination incident.

2.1 Booster placement for chlorine maintenance

Typically, booster disinfection is used by utilities to reintroduce disinfectant into the water distribution network in order to maintain acceptable disinfectant residual levels in remote parts of the network. Most utilities in the United States use free chlorine as their disinfectant while a few of them also use other alternatives like chloramine, ozonation, and Ultraviolet light (Uber et al., 2003; Ellison, 2003). In this

thesis, the terms “disinfectant booster stations” and “chlorine booster stations” are used interchangeably. A booster station generally is generally composed of a disinfectant storage tank, a small injector pump, and some type of control and safety unit that could either be manually operated or automatically controlled via a Supervisory Control Data Acquisition (SCADA) system (Isovitsch and VanBriesen, 2007). For a large water distribution network serving over 1 million customers, the total capital cost of installing a fixed booster disinfection station, which includes the equipment cost as well as the physical building and installations cost, can be up to $50,000 (EPA, 2010b).

The majority of existing research on the optimal booster station placement problem focuses on the first objective of maintaining a safe and consistent disinfectant residual throughout a water network. A number of different techniques have been proposed to solve this residual maintenance problem. Boccelli et al. (1998)present a Mixed Integer Linear Programming (MILP) formulation to determine optimal scheduling and location of booster stations that minimize the total amount of disinfectant needed to maintain specified disinfectant residual levels. The MILP formulation presented by Tryby et al. (2002) has a similar objective of minimizing the average dosage needed. Uber et al. (1998) propose another approach that aims at decoupling the influence of each booster while maximizing the overall node-time coverage. While these approaches maintain the residual concentration within acceptable bounds, they do not explicitly tackle the residual variability. Propato and Uber (2004b) address this issue by presenting a linear least-squares formulation that solves for the optimal injection schedule by explicitly minimizing the deviation of residual concentration from a required target. They later extend this approach by incorporating booster station locations as decision variables leading to a Mixed Integer Quadratic Programming (MIQP) formulation (Propato and Uber, 2004a). Ozdemir and Ucaner (2005) use a Genetic Algorithm (GA) linked with water network simulation software (EPANET) (Rossman, 2000) to optimize booster station locations and schedule. Lansey et al. (2007) propose a two level approach where the booster location problem is solved

at the top level (using GA, Branch & Bound, or enumeration), followed by a Linear Programming (LP) scheduling problem at the lower level.

2.2 Booster placement for incident response

While considerable research has focused on solving the residual maintenance problem, the optimal booster station location and scheduling problem for emergency response has been relatively less explored. This problem requires modeling the interaction between the contaminant and disinfectant. Moreover, since there is stochasticity associated with the location and time of a contamination, a multi-scenario approach is necessary. Parks and VanBriesen (2009) evaluate the effectiveness of disinfectant booster stations in intrusion mitigation by performing extensive contamination incident simulations using EPANET. These simulations are done over a range of reaction rate constants and a pre-selected set of possible booster locations (based on high reachability and low-residual criteria). The volume of contaminated water consumed (i.e., removed from the network for customer use) is used to gauge the impact of a particular contamination scenario along with the corresponding booster injection(s). This study also considers 5 different levels of response mechanisms: a combination of no response, booster response at first or second detection, and a do not consume order at first or second detection. The results show that using booster stations as the first level response (while waiting for further confirmation in case of a false positive) to a contamination incident can significantly reduce the overall impact (for contaminant susceptible to disinfectant). More importantly, the authors conclude that the location of booster stations can play a crucial role, and, therefore, an optimization-based approach is needed.

In one optimization-based approach proposed by Ostfeld and Salomons (2006), two different objective functions are recommended. The first objective, Min Cost, minimizes the overall cost of pumping and disinfection. This objective is designed to solve the residual maintenance problem. The second objective, Max Protection, maximizes the disinfectant concentration at all consumption nodes while maintaining

acceptable upper bounds. The authors note that this objective can be used as a response to a contamination incident. Since the disinfectant concentration at consumer nodes is dependent on the hydraulics in the network, both objectives consider four types of decision variables: scheduling of existing pump stations, tuning of valves, tuning injection rates at existing boosters stations, and the location and tuning of new booster stations. Therefore, the fact that this formulation does not assume precalculated hydraulics (scheduling of existing pumps and tuning of valves is a decision variable), leads to a large Mixed-Integer Nonlinear Programming Problem (MINLP), and the authors tackle this problem with a simulation-optimization based approach that couples a GA with EPANET. The authors recognize the limitations of using a GA that include computational cost and non-provable optimality. Also, the formulation presented does not explicitly maximize the impact of booster disinfectants in the case of a contamination incident and does not consider the stochasticity associated with the location and time of these incidents. In contrast, in this thesis we present stochastic programming formulations that considers interaction of contaminant and disinfectant (albeit approximately), and provides an optimal booster station placement that minimizes mass consumed (i.e., the mass of contaminant in the water that is removed for customer use) over a large set of potential contamination scenarios. Where the mass consumed at a node is defined as the demand times the contaminant concentration at that node. In order to solve the booster placement problem for large-scale realistic networks, we make several simplifying assumptions that will be discussed in the next two chapters. For example, in contrast to Ostfeld and Salomons (2006), additional benefit obtained from scheduling existing pumps are not considered.

3. THE NEUTRALIZATION FORMULATION FOR OPTIMAL BOOSTER STATION PLACEMENT ON LARGE-SCALE NETWORKS [1]

In this chapter, we address the optimal placement of fixed disinfectant booster stations to mitigate the effect of contamination incidents. This is a particularly challenging problem for two reasons. First, nonlinear reaction kinetics are required in order to accurately describe the interaction between the contaminant and the disinfectant. Additionally, the nonlinear interaction is specific to the contaminant-disinfectant pair, and the contaminant is likely unknown until after the laboratory analysis. Second, as the water network itself is large, and the time and location of the contamination incident is not known a priori, considering potential contamination incidents from every network node and all possible time steps leads to an extremely large number of potential contamination scenarios needed as inputs to the optimization problem.

Here, we assume a simplified contaminant-disinfectant interaction that allows us to precompute the effect of disinfectant booster stations and contaminant injection scenarios by independent simulation, thereby removing the need to embed a large-scale water quality reaction model within the optimization problem formulation. These simulations provide input data to two large mixed-integer linear programming formulations with hundreds of thousands of scenarios and discrete decision variables corresponding to the placement of booster stations within the network. The two proposed formulations use two separate objectives - mass of contaminant consumed as demand from nodes, and number of people that ingested the contaminant above a mass threshold. While these initial formulations are intractably large, we show a

[1]Part of this section is reprinted from "Efficient Reduction of Optimal Disinfectant Booster Station Placement Formulations for Security of Large-Scale Water Distribution Networks" by Seth, A., Hackebeil, G.A., Klise, K.A., Haxton, T., Murray, R., and Laird, C.D., 2015. Submitted to Computational Optimization and Applications.

series of reductions that significantly decrease the problem size and yield an exact mathematical transformation of the original stochastic programming problem. With these techniques, we demonstrate effective optimal booster station placement using real water network models containing more than 3,000 nodes. We originally proposed the first formulation discussed in this chapter (Mass Consumed formulation) along with the problem size reductions in a short proceedings document (Hackebeil et al., 2012). In this chapter, another formulation is proposed (Population Dosed formulation), the modeling technique for unknown contaminant-disinfectant reaction dynamics is described, and three case studies are performed analyzing the impact of network size on both the scalability of the formulations and the effectiveness of the booster placement.

3.1 Simplified modeling of unknown contaminant-chlorine reaction

Here, we discuss some of the modeling challenges associated with optimal placement of booster stations for response to contamination incidents. A water distribution system is typically modeled as a network of nodes and links, where the nodes include junctions, tanks, or reservoirs, and the links include pipes, pumps, and valves.

The optimal booster placement problem is discrete in nature: binary variables indicate whether or not a booster station is located at the corresponding network node. Realistic water network models can have thousands to hundreds of thousands of nodes, which means the number of binary variables in any problem formulation could be large. Inherent uncertainty also exists in the location and time of potential contamination incidents. Because of this uncertainty, it is necessary to consider potential contamination sources from different nodes and at different times. Here, we consider individual contamination incidents from every node and every time during a typical daily cycle of the water distribution network. In the absence of contaminant-disinfectant interaction, and with reasonable assumptions on flow properties within the water network, contaminant transport could be modeled as a linear system of equations (Mann et al., 2012b; Shang et al., 2002b). However, for network models

consisting of thousands of nodes, the water quality reaction model necessary to track the transport of a single species through the network can require hundreds of thousands to millions of variables and constraints depending on the time discretization (time steps) and simulation period used. In this work, we use the Merlion water quality modeling framework proposed by Mann et al. (2010), which provides a linear input-output relationship between species injection at all nodes and all time steps to the species concentration at all nodes and all time steps.

In the response problem, the specific contaminant and the reaction kinetics between the contaminant and the disinfectant will not be known at the time of detection. For the design problem of booster station placement, this causes significant uncertainty in the kinetic model form and the kinetic parameters. Furthermore, addressing the uncertainty and including these kinetic expressions in the water quality reaction model will give rise to a large-scale MINLP problem in which finding an optimal solution is intractable using existing tools.

We propose a method to overcome these challenges by using simplifying assumptions about the reaction between the contaminant and disinfectant which eliminates the complexities associated with modeling the reaction kinetics. These assumptions also allow contamination and booster simulations to be precomputed so that the optimization formulation does not have the water quality reaction model embedded. The resulting formulation is a stochastic MILP problem. Next, we list in detail the simplifying assumptions made to set up the optimal booster placement problem:

- The injection of contaminant or disinfectant into the network is assumed to not have an impact on the water flow rates. Therefore, the water quality reaction model assumes that the hydraulics are know inputs.

- We assume that no existing disinfectant is present in the network. In other words we assume that the existing disinfectant is only involved in maintaining water quality under normal operation and has no impact during a contamination incident.

- In case of a biological contaminant, the disinfectant reacts with it to either kill or damage the biological species to an extent that it cannot cause human disease or death. In case of a chemical contaminant, the disinfectant neutralizes it so that neither the contaminant nor its byproducts can cause harm.

- To remove the complexities associated with modeling the disinfection reaction, the disinfectant concentration is assumed to be high enough to completely neutralize the contaminant if they come into contact with a negligible change in the disinfectant concentration. Furthermore, when the contaminant comes into contact with the disinfectant at a particular node, the reaction proceeds to completion quickly enough (or at a timescale much smaller than the water quality time step) so that the contaminant does not get consumed from that node and does not travel to any downstream nodes.

- The booster stations start injecting disinfectants as soon as a contamination is detected.

Figure 3.1a illustrates the true behavior of a contaminant-disinfectant mixture while Figure 3.1b shows the impact of these assumptions on modeling the interaction between the disinfectant and the contaminant. Once a contaminant comes into contact with a disinfectant at node C, it is assumed to be completely and instantly neutralized while the excess disinfectant flows to the downstream nodes D, E, and F.

The resulting booster placement problem has many important modeling advantages which make the proposed formulation tractable for use with large networks. Under these assumptions, the problem can be formulated independent of any specific contaminant and disinfectant species, thereby removing the need to include nonlinear reaction kinetic equations. The other important advantage is the ability to superimpose individual simulations to determine the nodes that are neutralized. In Figure 3.2, we show how the results of two independent simulations of boosters at different locations can be superimposed over the contaminant simulation to obtain the overall neutralization effect of both boosters combined. This superposition principle holds at

any point in time as long as the time points of individual simulations are the same. As we will show, this property is important to formulate and solve stochastic programs considering a large number of contamination scenarios.

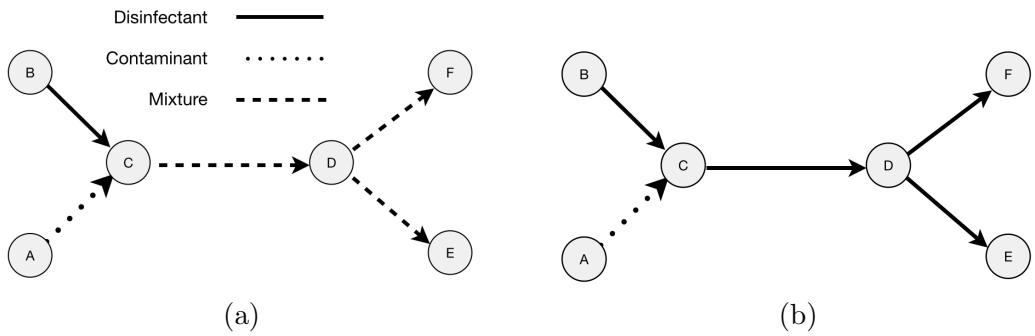
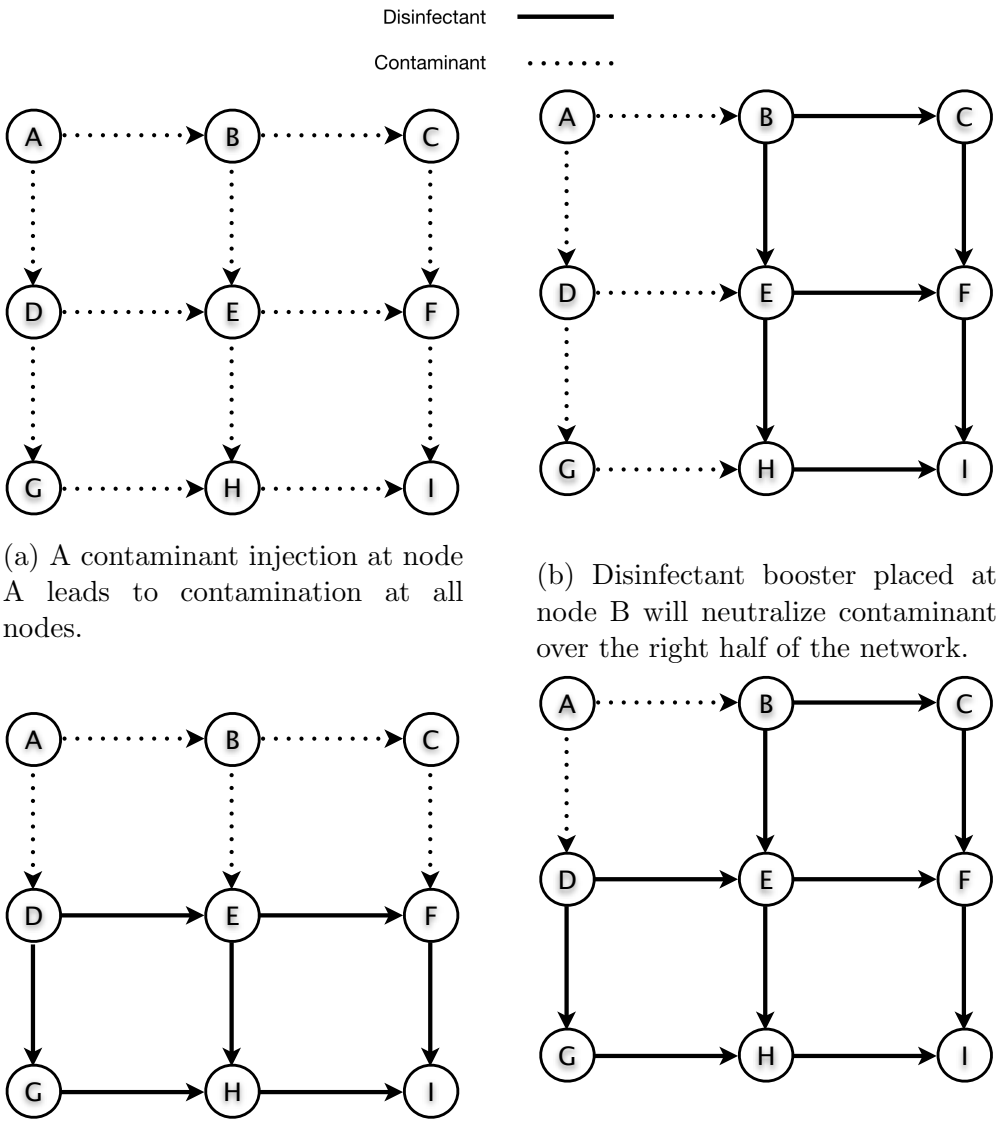


Figure 3.1.: Circles and arrows represent network nodes and links respectively. (a) Simple schematic showing that separation of species never occurs after mixing of contaminant and disinfectant streams, (b) Idealized reaction assumptions showing the complete and instant neutralization of the contaminant while there is sufficient amount of disinfectant to continue neutralizing downstream nodes.



(a) A contaminant injection at node A leads to contamination at all nodes.

(b) Disinfectant booster placed at node B will neutralize contaminant over the right half of the network.

(c) Disinfectant booster placed at node D will neutralize contaminant over the lower half of the network.

(d) The effect of contaminant injection at A and disinfectant booster at B and D is simply an overlap of booster impacts in Figure (b) and (c).

Figure 3.2.: An illustration showing that multiple disinfectant boosters can be simulated individually and then their effects can be superimposed to get the overall neutralization effect.

3.2 MILP formulations for optimal booster placement

In this section we propose two stochastic MILP formulation for optimal placement of fixed disinfectant booster stations. The first formulation is referred to as the "mass consumed formulation" or the "MC formulation" and it minimizes the mass of contaminant consumed in the form of demand from nodes. The second formulation is referred to as the "population dosed formulation" or the "PD formulation" and it minimizes the number of people that ingest the contaminant above a mass threshold. These objectives have been commonly used as metrics to assess the threat of contamination incidents and for optimal placement of water quality sensors (Murray et al., 2010a).

3.2.1 Mass Consumed (MC) Formulation

Let c_{nts} be a parameter that gives the concentration of contamination (in grams per cubic meter) present at node n and time t resulting from contaminant scenario s. This parameter is calculated by performing contamination simulations for each scenario. Now let δ_{nts} be a variable that is set to 1 if and only if the current booster station placement does not provide disinfectant to node n at time t for scenario s. Under this notation, the expected mass consumed (in grams) over all scenarios, where the mass consumed for each scenario is summed over all nodes and time steps can be written as:

$$E = \sum_{s \in S} \alpha_s \sum_{n \in N} \sum_{t \in T} \delta_{nts} v_{nt} c_{nts}, \tag{3.1}$$

where v_{nt} is the volumetric water demand (in cubic meter) consumed from node n during time step t (the same for each scenario) and α_s is the probability of scenario s. Here, N represents the set of nodes in the network, T represents the set of time steps resulting from the discretization, and S represents the set of all contamination scenarios. Let y_b be a binary variable that is 1 if a booster station is installed at node b and 0 otherwise. In later sections, we refer to B as the set of booster station

candidate nodes which could be a subset of N. For any particular booster station placement, the value of δ_{nts} can be constrained by

$$\delta_{nts} \geq 1 - \sum_{b \in D_{nts}} y_b, \tag{3.2}$$

where D_{nts} is the set of all booster station locations that supply disinfectant to node n at time t for scenario s. Note that D_{nts} depends on the detection time of scenarios s since additional disinfectant might not be added until contamination is suspected.

Given a known fixed sensor layout, we can compute the initial detection time for each contamination scenario and we will assume that booster stations begin injecting disinfectant at this detection time. Therefore, with knowledge of the sensor layout, we can find the list of detection times over all scenarios as part of the contamination simulations used to compute c_{nts}. For each of these detection times, we perform a disinfectant simulation from every candidate booster node. Using these simulation results, we can determine set D_{nts}. This requires a disinfectant simulation for every candidate booster node and every unique detection time (which is, at most, every time step in the simulation). The booster station placement problem for the MC objective can then be formulated as

$$\min \quad \sum_{s \in S} \alpha_s \sum_{n \in N} \sum_{t \in T} \delta_{nts} v_{nt} c_{nts} \tag{3.3}$$

$$\text{s.t.} \quad \delta_{nts} \geq 1 - \sum_{b \in D_{nts}} y_b \qquad \forall\, n \in N,\ \forall\, t \in T,\ \forall\, s \in S \tag{3.4}$$

$$\sum_{b} y_b \leq \mathbf{B}_{max} \tag{3.5}$$

$$0 \leq \delta_{nts} \leq 1 \qquad \forall\, n \in N, t \in T, s \in S \tag{3.6}$$

$$y_b \in \{0, 1\} \qquad \forall\, b \in B \tag{3.7}$$

Constraint 3.5 restricts the number of booster stations to be no more than $\mathbf{B}_{max}$. Although δ_{nts} is given as a continuous variable, since the objective function exerts pressure to minimize these variables, each δ_{nts} is guaranteed to have a value of 0

or 1 at the solution as long as $v_{nt}c_{nts} > 0$ (Berry et al., 2006). Otherwise, if the volumetric demand v_{nt} for the node is 0 or the concentration of contaminant c_{nts} is 0, those corresponding δ_{nts} variables will have no effect on the problem. The fact that we can avoid using discrete variables for δ_{nts}, which is defined over all scenario at all nodes and at all time steps, helps us significantly in being able to solve the proposed MILP problem efficiently. The formulation given in Equations 3.3-3.7 is equivalent to the weighted maximum coverage problem (Hochbaum, 1996) where we have to select a maximum of $\mathbf{B}_{max}$ sets from all D_{nts} sets in order to find the union that covers maximum number of δ_{nts} (weighted by $v_{nt}c_{nts}$).

3.2.2 Population Dosed (PD) Formulation

The booster station placement problem for the PD objective includes two additional constraints, and can be similarly formulated as

$$\min \quad \sum_{s\in S} \alpha_s \sum_{n\in N} z_{ns} pop_n \tag{3.8}$$

$$s.t. \quad \delta_{nts} \geq 1 - \sum_{b\in D_{nts}} y_b \qquad \forall\, n \in N, t \in T, s \in S \tag{3.9}$$

$$d_{ns} = \sum_{t\in T} \delta_{nts} I_{nts} \qquad \forall\, n \in N, s \in S \tag{3.10}$$

$$d_{ns} \leq z_{ns}(M - \tau) + \tau \qquad \forall\, n \in N, s \in S \tag{3.11}$$

$$\sum_{b\in B} y_b \leq \mathbf{B}_{max} \tag{3.12}$$

$$0 \leq \delta_{nts} \leq 1 \qquad \forall\, n \in N, t \in T, s \in S \tag{3.13}$$

$$y_b \in \{0, 1\} \qquad \forall\, b \in B \tag{3.14}$$

$$z_{ns} \in \{0, 1\} \qquad \forall\, n \in N, s \in S \tag{3.15}$$

The PD formulation is similar to the MC formulation in that Equations 3.9, 3.12, 3.13, and 3.14 are the same as Equations 3.4, 3.5, 3.6, and 3.7 respectively. The objective function in Equation 3.8 minimizes the population dosed across all nodes

and for every scenario. Each scenario s has probability α_s. Binary variable z_{ns} is used to indicate whether the total dosage at node n for scenario s is above a user specified dose threshold τ. The total population at a node is given by pop_n. Equation 3.10 calculates the mass dosed by the population at node n for scenario s. The parameter I_{nts} represents the mass ingested by the population at node n for scenario s, over the time step t. This parameter is also calculated from precomputed injection scenario simulations. Equation 3.11 is the big-M constraint used to switch the binary variable z_{ns} to 1 when the total mass dosed at node n for scenario s is above the threshold τ. Equations 3.13, 3.14, and 3.15 limit the range for δ_{nts} and state that booster placement, y_b, and dose above threshold, z_{ns}, are binary decision variables.

3.3 Structure-based problem size reductions

For large water network models, the full MILP formulations can still be intractable. For example, in the case of a 3,000 node network with 72,000 possible contamination scenarios (all nodes and all hours in a 24 hour cycle) and 100 water quality time steps, the MC formulation results in a problem with over 20 billion variables and constraints. Fortunately, a number of reductions can significantly decrease the problem size, while still providing an exact mathematical transformation of the full problem. In the next subsections, we describe these reductions for the two formulations.

3.3.1 MC Formulation

First, we outline the reductions for the MC formulation:

1. All variables and constraints corresponding to nodes where the mass consumed is 0 (i.e., $v_{nt}c_{nts} = 0$) can be eliminated from the problem. In these cases, we can remove δ_{nts} and its corresponding constraint from Equation 3.4. This reduction can eliminate a significant portion of the problem space that needs to be considered in the formulation. The reduction is particularly notable when

contamination scenarios occur in the edge of the network with a small number of downstream nodes, which means a large number of variables and constraints corresponding to the rest of the network can be eliminated. This reduction will also apply to all time steps prior to the start of the contamination.

2. The booster station simulations required to build D_{nts} also provide information about the set of (nts) indices which will never be impacted by any of the candidate booster nodes. To further reduce the size of the problem, all corresponding δ_{nts} variables can be replaced with a 1 in the objective function, and all corresponding constraints from Equation 3.4 can be eliminated. This is equivalent to the situation where $D_{nts} = \emptyset$. At a minimum, this reduction is applicable to all nodes and all times before the detection time for that scenario, t_s^d. In practice, this reduction might also apply to some nodes and times after t_s^d in the case where $B \subset N$ (i.e., all nodes are not candidate booster stations).

3. Because this problem formulation is derived from a network flow model, there is a tremendous amount of symmetry which occurs for the constraints in Equation 3.4 across different nodes, times, and scenarios. In particular, we consider the case where two booster impact sets are equal, $D_{n_1t_1s_1}$=$D_{n_2t_2s_2}$. Here, the corresponding $\delta_{n_1t_1s_1}$ and $\delta_{n_2t_2s_2}$ variables can be aggregated into one, removing one of the variables and the corresponding constraint from Equation 3.4 and aggregating the coefficients in the objective function. Note that this reduction is substantial, and it allows us to dramatically reduce the number of contamination simulations required, in addition to reducing the size of the problem. If two contamination scenarios, j and k, have the same detection time (i.e., $t_j^d = t_k^d$) then the booster impact sets will be the same for all nodes and times. That is, $D_{ntj} = D_{ntk}, \quad \forall\, n \in N,\ t \in T$. This allows us to aggregate all variables and constraints corresponding to these two contamination scenarios. Furthermore, the new coefficient in the objective function is a sum of all the aggregated terms, and because the water quality reaction model is linear, we can simply compute

this aggregated coefficient using a single contamination simulation where the contaminant injection is a probability (α_s) weighted sum of the individual contaminant injections. This reduction can be applied across booster impact sets until the final formulation consists entirely of unique sets D_p and disinfection indicator variables δ_p for $p \in Q$. Here Q is an indexing for the reduced problem with cardinality much smaller than the original number of constraints in Equation 3.4 (i.e, $|N| \times |T| \times |S|$).

These reductions not only help make the size of the MC formulation tractable, but they also reduce the number of contamination simulations required to generate the necessary data. The numerical results presented in the next section show that a significant reduction in the number of required contamination simulation can be obtained from the contamination scenario aggregation performed in reduction (3).

Without reduction (3), an estimate of the number of booster and contamination simulations required would be $|S| \times (|B| + 1)$ where S is the set of contamination scenarios and B is the set of booster station candidate nodes. This expression is explained by noting that, for each contamination scenario, we require one simulation for the contaminant and a simulation for each candidate booster node ($|B|$). To insure a high quality solution that accounts for uncertainty in the time and location of a contamination source, it is reasonable to assume a contamination scenario is needed for every node and at every hour during a demand cycle. In this case the number of simulations required is $|N| * |T| * (|B| + 1)$, where N is the set of nodes in the network and T is the set of hourly time steps over a typical demand cycle. In the case where the set of candidate booster stations B is nearly the same as the entire set of nodes N, we have that the proposed mixed integer program requires $O(|N|^2)$ contamination and booster simulations. As an example, consider the water network used in this study which has roughly 3,000 nodes. Formulating an optimization problem considering potential contamination from every node and at every hour over a 24 hour period requires roughly 3000*24*(3000+1)= 216 million simulations. However, by using the reductions discussed above, and assuming the water quality sensors sample every 15

minutes, the required simulations can be reduced to be no more than $|T_s| * (|B| + 1)$, which for the current example is roughly 96*(3000+1)=288,096 required simulations. Where T_s is the set of unique detection times over a 24 hour period.

3.3.2 PD Formulation

For the PD formulation, a similar set of reductions can be made:

1. All δ_{nts} variables and constraints (Equation 3.9) corresponding to 0 mass ingested (i.e., $I_{nts} = 0$) can be eliminated from the problem. Similarly, all variables and constraints corresponding to nodes with zero population can be removed from the problem. This includes variables z_{ns} and d_{ns} and the relevant constraints. Similar to the MC formulation, this reduction is particularly notable when contamination scenarios occur in the edge of the network with a small number of downstream nodes, which means a large number of variables and constraints corresponding to the rest of the network can be eliminated. This reduction will also apply to all time steps prior to the start of the contamination.

2. Similar to the reduction proposed for the MC formulation, all δ_{nts} variables corresponding to (nts) indicies that will never be impacted by any of the candidate booster nodes are replaced by 1. Consequently, all corresponding constraints from Equation 3.9 can be eliminated. This is equivalent to the situation where $D_{nts} = \emptyset$. Ideally, this reduction is applicable to all nodes and all times before the detection time for that scenario, t_s^d. In practice, this reduction might also apply to some nodes and times after t_s^d in the case where $B \subset N$ (i.e., all nodes are not candidate booster stations). Additionally, a similar reduction can be made by replacing binary variable z_{ns} by 1, for all nodes in a scenario where the cumulative dose, d_{ns}, is already greater than the dose threshold, τ, before the detection time. This reduction also results in the elimination of all corresponding δ_{nts} variables and constraints from Equations 3.9, 3.10, and 3.11.

Note that unlike the MC formulation, for the PD formulation we cannot combine two variables $\delta_{n_1 t_1 s_1}$ and $\delta_{n_2 t_2 s_2}$ in the general case where the two booster impact sets are equal, $D_{n_1 t_1 s_1}$=$D_{n_2 t_2 s_2}$. This is because we need to calculate the cumulative dose, d_{ns}, separately for all nodes and all scenarios. Therefore, the benefits of the third reduction proposed for the MC formulation, that include reducing the number of required simulations, are not applicable in the case of the PD formulation.

3.4 Numerical results and discussions

Given a particular water quality sensor layout, we can calculate the detection time, t_s^d, of a contamination scenario. At this time, the booster stations begins injecting additional (but within acceptable range of) disinfectant into the network, and a manual grab sample is drawn for lab testing and confirmation. The lab analysis can take Δt^{lab} time to obtain the results, while the booster stations continue to inject additional disinfectant. If lab results are negative, the booster stations cease injecting additional disinfectant. If lab results are positive, further response actions are required. To formulate the booster station placement problem, we are concerned with finding placements that provide as much benefit as possible while waiting for lab results.

The case studies performed in this manuscript assume random contamination scenario detection times that are uniformly distributed between 2 to 8 hours following the injections. Given the detection time t_s^d for each contamination scenario, we can simulate an injection from each individual booster station location, starting at t_s^d and ending at $t_s^d + \Delta t^{\text{lab}}$. With these booster simulation results, we can collect all booster station locations that affect a particular node and time and build the sets D_{nts} for each $n \in N,\ t \in T, s \in S$.

We examined the optimal placement of booster stations using the MC formulation on three water distribution networks of different sizes and the PD formulation on the smallest network. Results presented later in this section show drastic reduction in problem size for the MC formulation due to the third reduction presented in Section

3.3. Since the PD formulation does not benefit from this reduction, the larger network problems remain intractable. For each network, we altered the number of booster stations being placed and show their effectiveness in reducing the impact over a large set of contamination scenarios. We also present the optimization problem size statistics and the required computation time for each network.

Table 3.1 shows network size and contamination scenario statistics for each network (EPANET Example Network 3 (Rossman, 2000), Micropolis (Brumbelow et al., 2007), and Net6 (Watson et al., 2009)). For a particular network, the contamination scenario set contained contaminant injections from every junction and starting at every demand pattern time step during the first 24 hours of the simulation duration. For example, the contamination scenario set for the Micropolis network contained contaminant injections at all 1,574 junctions starting at each hour of the first 24 hours ($24 \times 1,574 = 37,776$). The duration of all contaminant injections was assumed to be 6 hours and all the contamination simulation durations were for 24 hours past the detection time ($\Delta t^{lab} = 24$ hours). We used random contamination scenario detection times that are uniformly distributed between 2 to 8 hours following the injection time (with hourly frequency). Therefore, for example, an injection taking place at 2 AM would have a detection time randomly assigned at any hour between 4 AM and 10 AM. Our assumptions of both injection duration and detection times impacted the results presented in the following subsections and they will be discussed therein. Due to the assumptions made about the contaminant-disinfectant reaction, the actual concentration of the contamination and booster injections has no impact on the following results.

3.4.1 Problem Size Reductions

Figure 3.3 shows the size of the MC formulation (MILP problem) that needs to be solved for each network (vertical axis is logarithmic scale). The size of the network has a clear impact on the size of the optimization problem. The large number of contamination scenarios also results in problem sizes that are prohibitively large.

Table 3.1: Scenario setup for each network

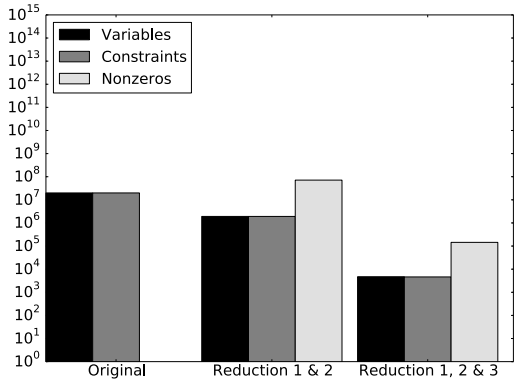
	Net3	Micropolis	Net6
Junctions	92	1,574	3,323
Links	119	1,619	3,892
Hydraulic Time Step (min)	15	15	15
Quality Time Step (min)	15	15	15
Pattern Time Step (min)	60	60	60
Contamination Scenarios	2,208	37,776	79,752
Booster Duration Δt^{lab} (hrs)	24	24	24

For example, the original problem for Net6 had close to 25 billion variables and constraints. Following the application of reductions (1) and (2), the problem size for each network was reduced by more than an order of magnitude. However, the problem size was still fairly large. For instance, in the case of Net6, we required approximately 8 terabytes to store the nonzeros (as 8-byte doubles) in the constraints. By applying reduction (3), the problem size was reduced another three to four orders of magnitude, giving reasonably sized problems with approximately $5,000$ variables for Net3, $100,000$ variables for Micropolis, and 1×10^6 variables for Net6. Figure 3.4 shows the size of the PD formulation that needs to be solved for Net3 before and after reduction (1) and (2). The final problem size in this case is approximately $200,000$ variables. Note that reductions (1) and (2) have a bigger impact on the PD formulation as compared to the MC formulation (Figure 3.3a and 3.4). This is because for the PD formulation, when the cumulative dose for a particular node and scenario, $d_{n,s}$, goes above the dose threshold, τ, before the scenario detection time, reduction (2) replaces $\delta_{n,t,s}$ with 1 for all time steps for that node and scenario. However, in the case of the MC formulation, only the $\delta_{n,t,s}$ variables before the detection time are replaced by 1.

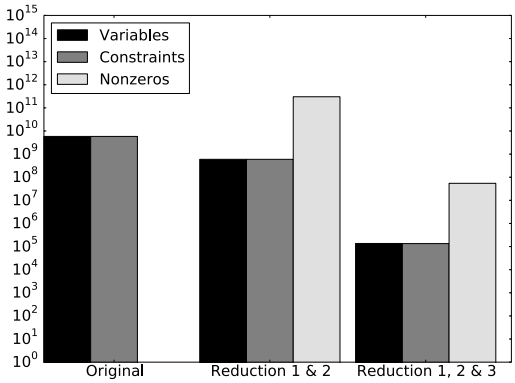
All problems were solved using CPLEX 12.5 on a machine with 64 AMD Opteron(TM) processors (6278 @ 2.4GHz). For each network, we solved multiple problems with the number of booster stations ranging from 0 to 10.

Table 3.2 provides the original and reduced number of contamination and booster simulations required, the mean simulation time, the mean solve time, and the peak memory usage for each network using the MC formulation. For the PD formulation on Net3, the original number of simulations were performed (204,972) that took 0.2 minutes and the MILP problems were solved with a mean solve time of 0.3 minutes and a peak memory usage of 0.8 GB. All of the contamination and booster simulations used the Merlion water quality reaction model (Mann et al., 2010). A preprocessing step was performed to discard a small set of contamination scenarios that injected at nodes with stagnant flow, since these scenarios had no impact.

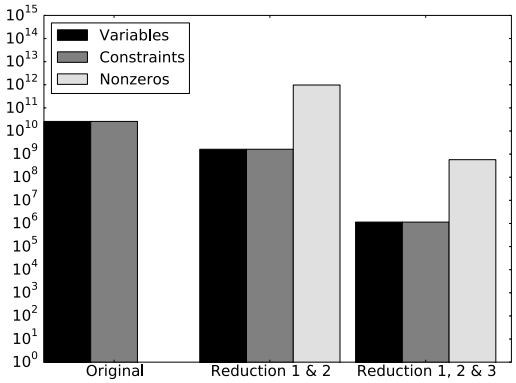
It should be noted that, while commercial optimization and modeling softwares like CPLEX have a presolve phase, the original problem was far too large to even fit in memory on a reasonable workstation. Furthermore, even with smaller test problems, we did not see significant reduction in the problem size using the presolve in CPLEX 12.5.



(a) Net3



(b) Micropolis



(c) Net6

Figure 3.3.: Problem size (log scale) of the original full space MC formulation, the problem size following reductions (1) and (2), and the problem size following reductions (1), (2) and (3).

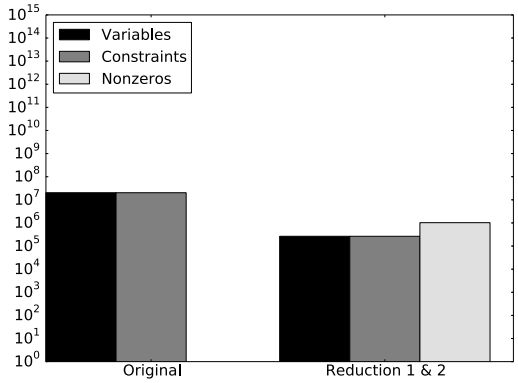


Figure 3.4.: Problem size (log scale) for Net3 of the original full space PD formulation and the problem size following reductions (1) and (2).

Table 3.2: Number of simulations required for the MC formulation along with timing and memory usage statistics for solving the fully-reduced problem.

Network	Original Simulations	Reduced Simulations	Simulation Time (min)	Mean Solve Time (min)	Peak Memory Usage (GB)
Net3	204,972	2,790	0.01	0.0025	0.018
Micropolis	57,229,200	47,250	0.97	4.4	4.8
Net6	264,783,192	99,720	19.5	19.2	50

As mentioned earlier, the problem sizes depended not only on network size and contamination scenario set size, but also on the assumptions of contaminant injection length and detection time within the contamination scenario set. Longer injections would generally mean that more nodes have nonzero concentrations over the simulation period and therefore fewer variables and constraints can be eliminated using reduction (1) (nodes where the mass consumed is zero for the MC formulation i.e., $d_{nt}c_{nts} = 0$ or mass ingested is zero for the PD formulation i.e., $I_{nts} = 0$). On the other hand, shorter injection lengths generally mean that reduction (1) can eliminate

more variables and constraints. Similarly, in the case of detection times, we assumed that detection takes place randomly at any hour in a 2 to 8 hour window following the injection. This assumption implied that for our contamination scenario set containing injections at every node and starting at every hour during the first 24 hours, there could be 31 unique detection times (detection time could range from 2 to 32 hours as injection times ranged from 0 to 24 hours). Therefore, during reduction (3) when we aggregated constraints and variables for contamination scenarios with the same detection times, a different number of unique detection times could have impacted the size of the reduced set D_{nts}.

3.4.2 Impact of Optimal Booster Placement

The effectiveness of optimal booster placement in reducing the expected mass consumed over the large set of contamination scenarios is shown in Figure 3.5. Figures 3.5a, 3.5b, and 3.5c illustrate the results for Net3, Micropolis, and Net6, respectively. The horizontal axis in each plot represents the number of booster stations being placed while the vertical axis represents the expected mass consumed over all contamination scenarios normalized with respect to the overall expected mass consumed when no booster is placed. The horizontal dashed line in each plot represents the normalized expected mass consumed before detection, or in other words, the amount which cannot be reduced. This line signifies the best possible performance that could be achieved by placing boosters to completely neutralize all contamination as soon as an incident is detected. All three plots in Figure 3.5 show that as we increase the number of booster stations being placed, the normalized expected mass consumed asymptotically progresses towards the best possible performance represented by the dashed line. Similarly, Figure 3.6 shows the effectiveness of optimal booster placement in reducing the expected population dosed over a set of contamination scenarios on Net3.

Again the length of contaminant injections and their detection times used to build our contamination scenario set plays an important role in determining the booster performance shown in Figures 3.5 and 3.6. For instance, consider the position of the

dashed lines in these two figures. Increasing the injection lengths while keeping the detection times the same would mean that the fraction of impact before detection decreases and, therefore, the dashed line would be lower on the plot. Likewise, an increase in detection times would correspond to a larger fraction of impact before detection and, therefore, the dashed line would be higher on the plot.

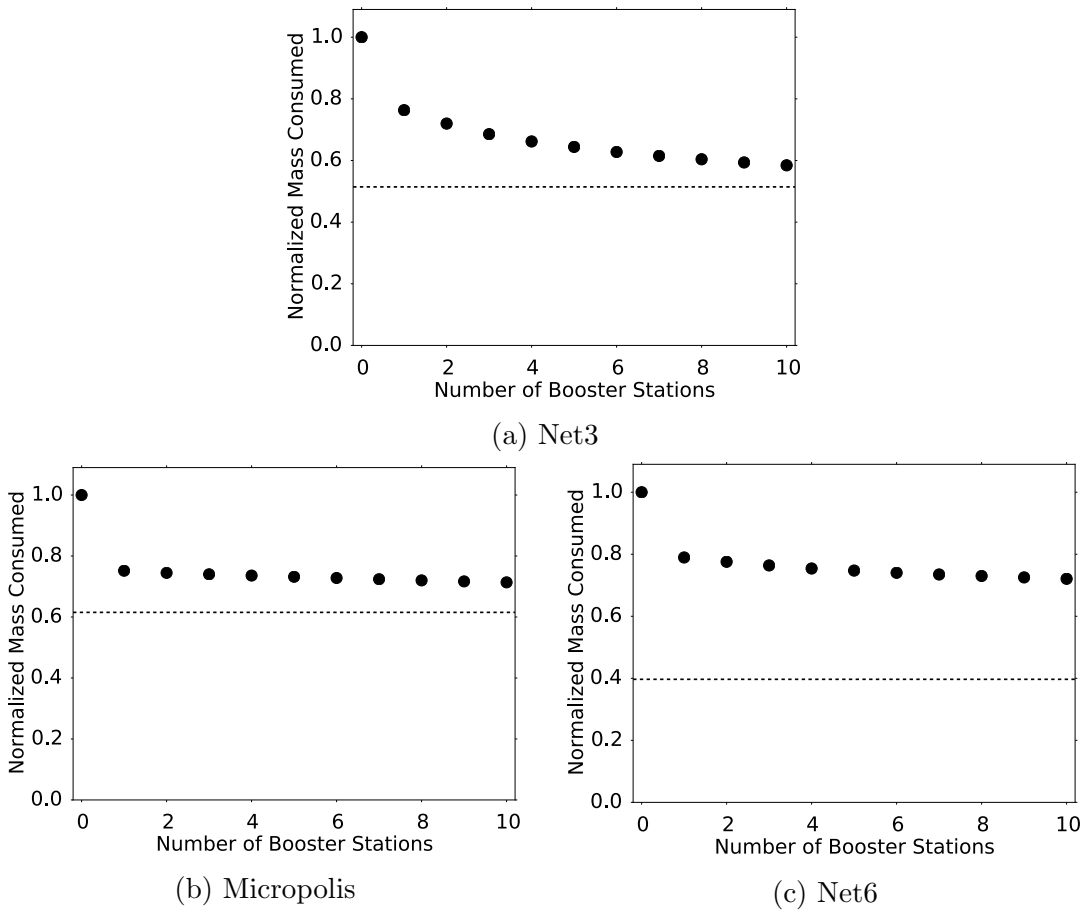


Figure 3.5.: The impact of optimal booster station placement on normalized expected mass consumed. The horizontal dashed line represents the normalized expected mass consumed before detection.

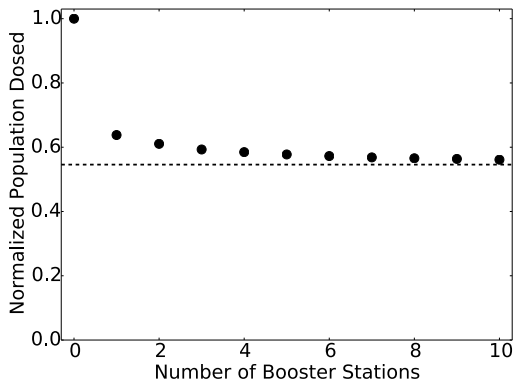


Figure 3.6.: The impact of optimal booster station placement on normalized expected population dosed on Net3. The horizontal dashed line represents the normalized expected population dosed before detection.

All things considered, for all three networks, the optimal booster placement is able to notably reduce the overall impact. For example, Figure 3.5c shows that, 40% of the mass consumed is before detection. However, placing 10 boosters helps reduce 49% of the remaining 60% mass consumed after detection. The population dosed metric shows a more drastic reduction in Figure 3.6, even with a small number of optimally placed booster stations. These results imply that booster stations can be used as an effective response strategy to reduce the impact of potential contamination incidents in water distribution networks.

4. THE LIMITING REAGENT FORMULATION FOR OPTIMAL BOOSTER PLACEMENT [1]

In Chapter 3 we proposed two MILP formulations to identify booster station locations that minimized two different objectives (mass consumed or the population dosed). The MILP formulations were able to find optimal locations for boosters in large networks by considering a large ensemble of contaminant scenarios, but the Neutralization method greatly simplified the reaction by assuming that the chlorine instantly and completely inactivates the contaminant when it comes in contact with chlorine. In this chapter, we propose a new booster station optimization method that is referred to as the "Limiting reagent method" and evaluate both the methods to compare their results.

Since the contaminant species is unknown at the planning stage, both methods need to make assumptions in order to approximately model the contaminant-chlorine reaction. Simplifying assumptions also aid in keeping the optimization problems tractable for large scale networks. The optimization formulations proposed in both methods includes stochasticity in the location and time of the contamination incident. The major difference between the two methods lies in the fact that the Neutralization method assumes that chlorine is always in stoichiometric excess as it neutralizes contaminant through the network, while the Limiting reagent formulation proposed in this chapter allows for modeling different stoichiometric ratios between the contaminant and chlorine. Different levels of contaminant-chlorine reactivities are approximated using the proposed formulation and their impact is studied on both the

[1]Part of this section is reprinted from "Evaluation of Chlorine Booster Station Placement for Water Security" by Seth, A., Hackebeil, G.A., Haxton, T., Murray, R., Laird, C.D., and Klise, K.A., 2015. Submitted to Journal of Water Resources Planning and Management, American Society of Civil Engineers.

booster placement layout and the effectiveness of the booster stations in reducing impact on the population. Additionally, we also investigate how sensor placement can influence the performance of disinfectant booster stations. The booster station placement case studies presented here considers all the different uncertainties associated with the problem (location and time of incident, contaminant-chlorine reactivities, and incident detection time) as realistically as possible.

4.1 Formulations based on different modeling techniques

The Neutralization and Limiting Reagent methods approximate the unknown reaction between a contaminant and chlorine. For both methods, the contaminant and chlorine concentrations are calculated using water network hydraulic and water quality models. For the results shown in this work, EPANET 2.0 (Rossman, 2000) is used to perform the hydraulic simulations and Merlion (Mann et al., 2012b; Wong et al., 2010) is used for the water quality calculations. Although not used in this work, our implementation of the Neutralization method does support the use of EPANET as the hydraulic and water quality simulator. However, as we will show in the next section, the Limiting reagent method requires us to use the linear water quality equations from Merlion. Both the Neutralization and the Limiting Reagent methods are included in US EPA's Water Security Toolkit (WST), a suite of software tools designed to help evaluate and plan response strategies in the case of a contamination incident. Additional information on these methods and on human health impact models used to compute the population dosed can be found in the WST User Manual (EPA, 2014).

4.1.1 Neutralization method

The optimization formulation described in Equations 3.8-3.15 (PD formulation) for the Neutralization method is studied in this chapter. Here, we reiterate some of the major assumptions that are made. The Neutralization method assumes that the chlorine completely and quickly inactivates all of the contaminant on contact. The

Neutralization method takes advantage of several simplifying assumptions to model the unknown contaminant-chlorine reaction. Firstly, it is assumed that the chlorine remains in stoichiometric excess, and therefore the contaminant-chlorine reaction does not effect the chlorine transport in the network. Secondly, both the contaminant and the chlorine are assumed to behave like tracers that do not decay as they flow through the network. Finally, this method ignores residual chlorine already injected into the network from water treatment facilities and only the chlorine injected from the booster stations in considered. The advantage of these assumptions is that we no longer need to embed a reaction model, which can be non-linear, within the problem formulation. Another advantage of these assumptions is that now the booster chlorine injections and contaminations injections can be pre-simulated and the resulting data can be used to formulate an MILP problem for optimal booster placement.

4.1.2 Limiting Reagent Method

Similar to the Neutralization method, the Limiting reagent method also assumes that the contaminant-chlorine reaction happens at a fast rate. However, unlike the Neutralization method, the Limiting reagent method allows for the reaction to happen with respect to a stoichiometric ratio. In this work, we define the stoichiometric ratio as the mass of chlorine removed per the mass (mg) (if chemical) or colony-forming units (CFU) (if biological) of contaminant rendered harmless after reacting with chlorine. To illustrate the difference between the Limiting Reagent and Neutralization method, two examples are shown in Figure 4.1. For both examples, a stoichiometric ratio of 1 mg chlorine/mg contaminant is used for the Limiting Reagent method. In Example A, 100 mg of chlorine comes in contact with 80 mg of contaminant at a pipe junction. Using the Neutralization method, all of the contaminant is inactivated and the amount of chlorine remains unchanged. Using the Limiting Reagent method, 80 mg of chlorine is used to inactivate all of the contaminant. In this case, the contaminant is the limiting reagent and 20 mg of chlorine remains. Example B illustrates a case where chlorine is the limiting reagent. In this case, 100 mg of chlorine comes

in contact with 120 mg of contaminant. Results using the Neutralization method are unchanged. Using the Limiting Reagent method, 100 mg of chlorine can inactivate 100 mg of the contaminant with 20 mg of contaminant remaining.

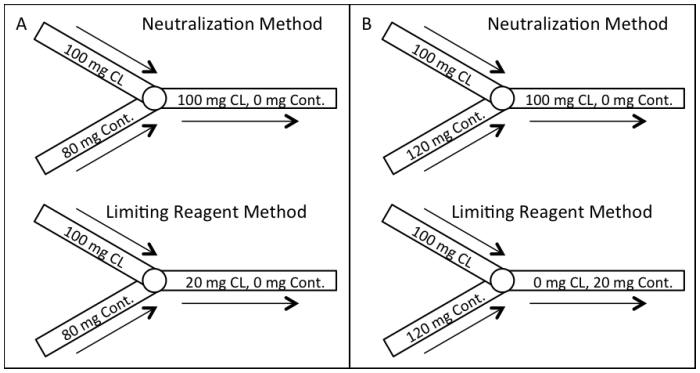


Figure 4.1.: Neutralization and Limiting Reagent methods Example A and Example B. Both examples assume a stoichiometric ratio of 1 mg chlorine (CL)/mg contaminant (Cont.)

The Limiting Reagent method assumes that the contaminant-chlorine reaction proceeds to completion at a fast rate until the limiting reagent, which can either be the chlorine or the contaminant, is exhausted. As the stoichiometric ratio approaches zero, the Limiting Reagent method is equivalent to the Neutralization method. The Limiting Reagent method also ignores the residual chlorine already present in the network and that only chlorine injected from booster stations reacts with the contaminant. However, in order to model the contaminant-chlorine reaction, the Limiting reagent model explicitly embeds the water quality model directly into the optimization formulation. Using the linear water quality model introduced in Section 1.2.2, the Limiting reagent method formulates the optimal booster placement problem as an MILP as follows:

$$\min \quad \sum_{s \in S} P(s) \sum_{n \in N} z_{ns} pop_n \tag{4.1}$$

$$s.t. \quad \mathbf{G}\mathbf{c}_s^{con} = \mathbf{D}(\mathbf{m}_s^{con} - \mathbf{r}_s^{con}) \qquad \forall\, s \in S \tag{4.2}$$

$$\mathbf{G}\mathbf{c}_s^{dis} = \mathbf{D}(\mathbf{m}_s^{dis} - \rho \mathbf{r}_s^{con}) \qquad \forall\, s \in S \tag{4.3}$$

$$m_{bts}^{dis} = y_b L_{bts} \qquad \forall\, b \in B, t \in T, s \in S \tag{4.4}$$

$$m_{nts}^{dis} = 0 \qquad \forall\, n \in N \backslash B, t \in T, s \in S \tag{4.5}$$

$$d_{ns} = \sum_{t \in T} c_{nst}^{con} v_{nst} \qquad \forall\, n \in N, s \in S \tag{4.6}$$

$$d_{ns} \leq z_{ns}(M - \tau) + \tau \qquad \forall\, n \in N, s \in S \tag{4.7}$$

$$\sum_{b \in B} y_b \leq B_{max} \tag{4.8}$$

$$y_b \in \{0, 1\} \qquad \forall\, b \in B \tag{4.9}$$

$$z_{ns} \in \{0, 1\} \qquad \forall\, n \in N, s \in S \tag{4.10}$$

$$c_{nts}^{con}, c_{nts}^{dis}, r_{nts}^{con} \geq 0 \qquad \forall\, n \in N, t \in T, s \in S \tag{4.11}$$

$$\tag{4.12}$$

where S, N, T and B represent the sets of contamination scenarios, network nodes, time steps, and potential booster station locations, respectively. The objective function in Equation 4.1 minimizes the population dosed at all nodes for every contamination scenario in the simulation. Each scenario s has probability $P(s)$. Binary variable z_{ns} is used to indicate whether the total dosage at node n for scenario s is above a user specified dose threshold τ. The total population at a node is represented by pop_n. Because the water quality model is formulated as a set of linear equations, the forward tracing simulations can be included directly within the MILP. The concentration of the contaminant and chlorine, c_{nts}^{con} and c_{nts}^{dis}, respectively, are defined for each node n, time step t, and contamination scenario s. The variables m_{nts}^{con} and m_{nts}^{dis} are the mass injections for the contaminant and chlorine, respectively,

at node n and time step t for contamination scenario s. Equations 4.2 and 4.3 include the embedded linear water quality model, Merlion, as stored in the $\mathbf{G}$ and $\mathbf{D}$ matrices. The $\mathbf{G}$ and $\mathbf{D}$ matrices map the contaminant and chlorine mass injected at all nodes and time steps for a scenario s (vectors $\mathbf{m}_s^{con}$ and $\mathbf{m}_s^{dis}$) to contaminant and chlorine concentration at all nodes and time steps for each scenario s (vectors $\mathbf{c}_s^{con}$ and $\mathbf{c}_s^{dis}$). The contaminant mass removed at all nodes and time steps for contamination scenario s, based on the reaction between the contaminant and chlorine, is given by the vector $\mathbf{r}_s^{con}$. The stoichiometric ratio, ρ, defines the mass of chlorine removed per mass of contaminant removed. Equations 4.4 and 4.5 set the booster injection amount. The amount is L_{bts} if a booster station is placed at node b, otherwise the injection amount is zero. The binary variable y_b is 1 if node b is selected as a booster station location and 0 otherwise. Equation 4.6 calculates the mass dosed by the population,d_{ns}, at node n for scenario s. The parameter v_{nst} represents the volume of water ingested by the population at node n for scenario s, over the time step t. Equation 4.7 is the big-M constraint used to switch the binary variable z_{ns} to 1 when the total mass dosed at node n for scenario s is above the threshold τ. Equation 4.8 restricts the number of booster stations to be less than or equal to B_{max}. Equation 4.9 and 4.10 define y_b and z_{ns} as a binary variable respectively. Equation 4.11 indicates that the contaminant and chlorine concentrations and the contaminant mass removed are greater than or equal to zero.

4.2 Evaluation of the Neutralization and the Limiting reagent method

The case studies presented in this section cover several factors that can influence the effectiveness of booster chlorination as in incident response action. These include the water network layout, the ensemble of potential contamination scenarios, the sensor placement layout, and the parameters related to the booster station operation. We first define the range of these parameters used in this case study, followed by a discussion of the results.

4.2.1 Case study design

This case study uses two example water distribution networks from the literature - (1) The Example Network 3 distributed with EPANET 2.0 (Figure 4.2), which we refer to as "Network 1," and (2) A larger network, referred to as "Network 2" (Figure 4.3) (Watson et al., 2009). Network 1 is composed of 92 junctions, 3 tanks, and 2 reservoirs, and serves water to approximately 62,000 customers. Network 2 has 407 junctions, 2 tanks, and 1 reservoir, and serves water to approximately 6,400 customers.

The time delay between the contamination incident and the start of booster chlorination depends on the incident detection time and the time it takes to activate the boosters. The ability of a sensor placement layout to quickly detect a contamination incident has a great impact on the effectiveness of all mitigation actions including booster chlorination. A late detection can result in the majority of the damage being done before the booster can activate. To study the effect of scenario detection time, a range of sensor layout were studied in this work. Sensor locations were identified using the *sp* (sensor placement) module in WST. For Network 1, three sensor placement layouts were optimized to place 2, 5, and 10 sensors. For Network 2, the studies were performed using one sensor placement layout with 5 optimally placed sensors. Examples of the chosen sensor locations are shown in Figures 4.2 and 4.3.

The optimal sensor placement and the optimal booster placement are two independent problems that require defining a set of possible contamination scenarios. For simplicity, the same set of scenarios was used for the optimum placement of sensors and booster stations in these case studies. The sensor placement was performed to minimize the detection time. One contamination scenario was simulated from each non-zero demand (NZD) node in the network. NZD nodes are defined as nodes with positive customer demands. Network 1 has 59 NZD nodes and Network 2 has 105 NZD nodes. For all case studies performed in this chapter, injection strength and dosage threshold values studied by Davis et al. (2014) were used. For each scenario,

0.5 kg of contaminant was injected into the network, starting at midnight on the second day of the simulation and ending an hour later. The impact of each contamination scenario was calculated for 8 hours following the detection time. To calculate population dosed, it was assumed that each person ingested 2 liters of water uniformly throughout the day. Two dosage thresholds (τ) were used to evaluate the population dosed metric: 0.0001 and 0.1 mg. Although these threshold values can go lower for certain contaminants, for the purposes of this work, a dose threshold of 0.0001 mg represents high toxicity, while a dose threshold of 0.1 mg represents low toxicity. It should be noted that the mass injection rate and the dose threshold are relative and can be scaled as described in Davis et al. (2014).

The chlorine injected at the booster station was assumed to be at a concentration of 4 mg/L (the MCL for chlorine) and in the injections were assumed to continue for 8 hours. Only the chlorine supplied by the booster stations was considered by the optimization methods while the residual chlorine was ignored. The set of NZD nodes were used as feasible booster station locations for both networks. In order to cover a wide range of contaminants, the following stoichiometric ratios were used to approximate a strong to weak reaction with chlorine: 0 mg CL/mg contaminant, 1 mg CL/mg contaminant, 10 mg CL/mg contaminant, and 100 mg CL/mg contaminant. When the stoichiometric ratio is set to 0 mg CL/mg contaminant, the Neutralization method is used to place boosters in the network.

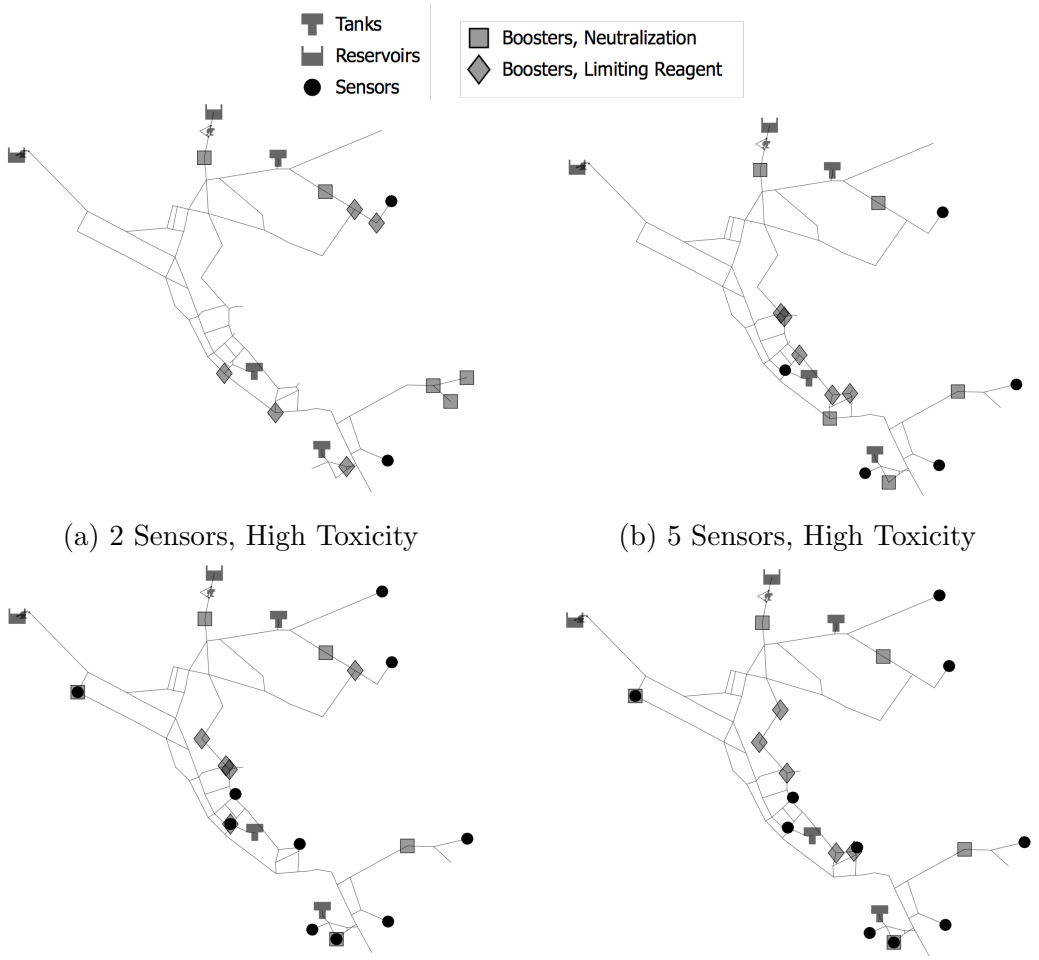
4.2.2 Case Study Results

The following results compare the effectiveness of booster station response to a set of possible contamination scenarios in two networks given a range of detection times (i.e., range of sensor layouts), contaminant toxicities, and stoichiometric ratios. Since booster stations would not be turned on until after detection, it is important to understand how a particular sensor layout impacts the population dosed at the time of detection. For Network 1, 3 sensor layouts (2, 5, and 10 sensors) were used to detect the possible contamination scenarios, while for Network 2 a single network

design with 5 sensors was tested. Ideally, all scenarios would be detected by a given sensor placement layout, but this is not the case (unless a sensor can be placed at every node). The number of detected scenarios in Network 1 increased from 72% with a 2-sensor layout to 85% with a 5-sensor layout to 93% with a 10-sensor layout. For the detected scenarios, the corresponding average detection time decreased from 8.0 hours to 1.9 hours to 1.6 hour. The average population dosed at the time of detection was 16,537, using a 2-sensor layout, 2,999 using a 5-sensor layout and 1,123 using a 10-sensor layout, assuming a highly toxic contaminant (dose threshold of 0.0001 mg). This impact cannot be reduced by adding chlorine at booster stations because the boosters are not initiated until after detection. The percent of scenarios detected in Network 2 was lower than Network 1, the average detection time and population dosed at detection was also higher in Network 2. Based on a 5-sensor layout, 78% of contamination scenarios was detected, the average detection time was 4.6 hours, and the population dosed at time of detection was 162 assuming a highly toxic contaminant (dose threshold of 0.0001 mg). Using a 10-sensor layout, 86% of contamination scenarios was detected, the average detection time was 2.9 hours, and the population dosed at time of detection was 132, assuming a highly toxic contaminant (dose threshold of 0.0001 mg).

The Neutralization and Limiting Reagent methods were used to optimally locate 1 to 10 booster station locations in both networks. Figures 4.2 and 4.3 illustrate five optimally placed booster stations in Network 1 and Network 2 using two different sets of parameters that include the number of sensors, the stoichiometric ratio (Neutralization = 0, Limiting Reagent = 100), and the dose threshold. Both figures show significant variability in the optimal booster station locations depending on these parameters. Figures 4.2a and 4.2b show that going from 2 sensors to 5 sensors can result in very different booster placements for both Neutralization and Limiting Reagent methods. However, the Neutralization method shows no or very little sensitivity to dose threshold as evident from comparing Figures 4.2c and 4.2d or Figures 4.3a and 4.3b. Overall, for both networks (Figures 4.2 and 4.3), the Limiting Reagent

method (with a high stoichiometric ratio) resulted in the boosters being placed more centrally in the network, while the Neutralization method resulted in the boosters being placed more towards the edges of the network.



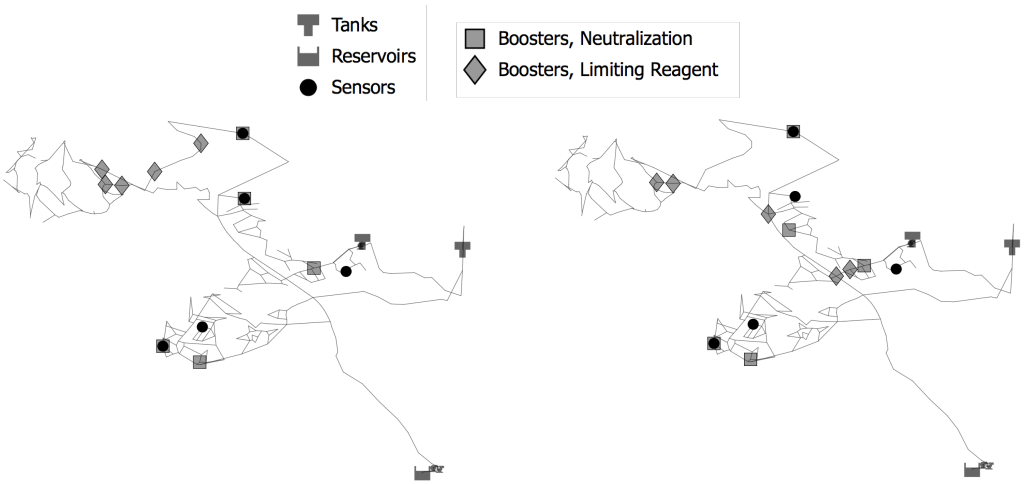
(a) 2 Sensors, High Toxicity

(b) 5 Sensors, High Toxicity

(c) 10 Sensors, Low Toxicity

(d) 10 Sensors, High Toxicity

Figure 4.2.: Example booster station placement for Network 1 with (a) a 2-sensor layout and high toxicity contaminant, (b) a 5-sensor layout and a high toxicity contaminant, (c) a 10-sensor layout and a low toxicity contaminant, and (d) a 10-sensor layout and a high toxicity contaminant. Five booster stations are placed using the Neutralization method and the Limiting Reagent method with ρ=100.



(a) Low Toxicity (b) High Toxicity

Figure 4.3.: Example booster station placement for Network 2 with (a) a 5-sensor layout and a low toxicity contaminant and (b) a 5-sensor layout and high toxicity contaminant. Five booster stations are placed using the Neutralization method and the Limiting Reagent method with ρ=100.

Table 4.1 provides mean problem size and solution time statistics for all the optimization problems solved using the Neutralization and Limiting Reagent methods. These results highlight the biggest advantage of the Neutralization method. While the Limiting Reagent formulation can take several hours to solve for Network 2, the Neutralization formulation solves within seconds. Instead of embedding the water quality model into the formulation, the Neutralization method benefits from performing simulations outside of the optimization that take less than a 10 seconds for both networks.

Table 4.1: Mean number of variables and solution times for the Neutralization and Limiting Reagent formulations on Network 1 and Network 2.

		Neutralization Method		Limiting Reagent Method	
Network	Junctions	Variables	Solve Time (Hrs)	Variables	Solve Time (Hrs)
Network 1	92	6,114	0.002	461,928	1.23
Network 2	407	18,253	0.01	3,583,786	7.32

Figure 4.4 quantifies the impact of the number of sensors, the stoichiometric ratio, and the contaminant toxicity on the performance of the optimally placed booster stations on Network 1. The following set of observations can be made from Figure 4.4:

- With an increase in the number of booster stations, the population dosed asymptotically approached a minimum value.
- This minimum value was a function of the number of sensors, the stoichiometric ratio, and the contaminant toxicity.
- A three orders of magnitude difference in the dose threshold resulted in about an order of magnitude difference in the number of population dosed. For example, the population dosed for the 10 optimally placed booster obtained using the Neutralization method (ρ=0) with 2 sensors and a dose threshold of 0.0001 mg (high toxicity) was 17,498, while the population dosed for the same set of parameters using a dose threshold of 0.1 mg (low toxicity) was 1,224.
- Going from 2 sensors to 5 sensors had a bigger impact in the performance of the booster stations as compared to going from 5 sensors to 10 sensors. This was because of a larger reduction in mean detection time going from 2 sensors to 5

sensors (8.0 hours to 1.9 hours) as compared to the reduction in mean detection time going from 5 sensors to 10 sensors (1.9 hours to 1.6 hours).

- The effect of the stoichiometric ratio on the expected population dosed was more significant in the presence of higher number of sensors. For example, in Figure 4.4f (10 sensors, high toxicity), placing 10 boosters resulted in the population dosed going from 1,433 for a stochiometric ratio of 1 to 7,021 for a stochiometric ratio of 100. On the other hand, in Figure 4.4d (5 sensors, high toxicity), placing 10 boosters resulted in the population dosed going from 3,467 for a stochiometric ratio of 1 to 7,239 for a stochiometric ratio of 100.

Figure 4.5 shows the impact of the stoichiometric ratio and the contaminant toxicity on the performance of the optimally placed booster stations on Network 2 with five optimally placed sensors. For Network 2, the effect of contaminant toxicity was not as significant as observed in Network 1. It is conjectured that this behavior is due to the fact that Network 2 has a much smaller population that is spread out over a larger number of nodes, and, therefore the expected population dosed did not show a big variation with respect to the contaminant toxicity even in the absence of booster stations.

If booster stations are to be used as a part of water utilities response action plan, then a single booster station placement would be used without knowing the specific contaminant toxicity or its reaction with chlorine. The physical locations of booster stations placed using the Neutralization and Limiting Reagent methods can be evaluated given contamination scenarios with different toxicities and stoichiometric ratios of reaction with chlorine. For example, if the water utility assumes the worst, and places booster stations assuming that the contaminant is of high toxicity and does not react strongly with chlorine, then the booster stations placed using the Limiting Reagent method with a high stoichiometric ratio and low population dosed threshold (high toxicity) can be used to evaluate other types of scenarios. Tables 4.2 and 4.3 list the expected population dosed given the optimal placement of 10

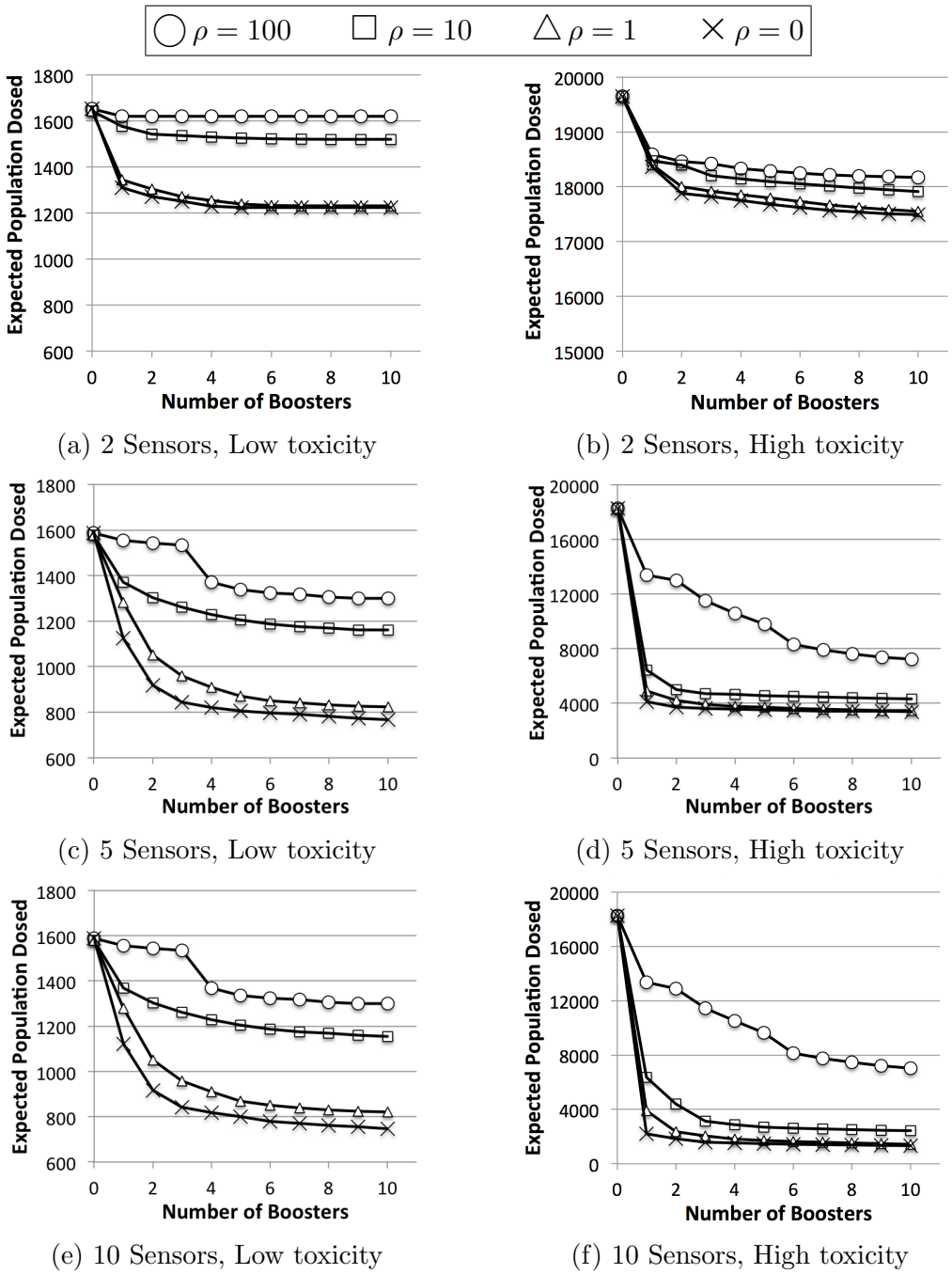
booster stations for each toxicity level and stoichiometric ratio on Networks 1 and 2, respectively. The performance of the optimal placement was then used to evaluate the population dosed under the same range of contaminant toxicities and stoichiometric ratios used for the optimal booster station placement. Each case used the 5-sensor layout to detect contamination in the network. The last column in each table is the mean of the expected population dosed over a row; the mean explains the average number of people dosed if all 8 of the scenarios occurred and the specific set of booster stations were installed. As expected, for a particular contaminant toxicity and stoichiometric ratio, the optimal placement always gave a lower objective as compared to the evaluation of all other placements on the same contaminant toxicity and stoichiometric ratio. For instance, in Table 4.2 the expected population dosed for the optimal placement considering high contaminant toxicity and a stoichiometric ratio of 1 was 3,455. This value was lower than all the evaluated objective values for high contaminant toxicity and stoichiometric ratio of 1 (4th column). Tables 4.2 and 4.3 also show that the overall performance of a booster placement, represented as the mean of population dosed values over different levels of contaminant toxicity and stoichiometric ratio (last column), improved as the stoichiometric ratio and the contaminant toxicity increased. For Network 1 (Table 4.2), the lowest mean expected population dosed was evaluated at 3,061 using contamination scenarios of high toxicity and high stoichiometric ratio (ρ=100). On the other hand, the largest mean expected population dosed was evaluated at 4,320 using contamination scenarios of high toxicity and ρ=0 (Neutralization method). Comparing these two numbers (minimum and maximum mean) resulted in a difference of 1,251 in the mean expected population dosed on Network 1. Similarly, for Network 2 (Table 4.3), the lowest mean expected population dosed was evaluated at 366 using contamination scenarios of high toxicity and high stoichiometric ratio (ρ=100). These results imply that performing optimal booster placement for the worst case scenario (high contaminant toxicity and high stoichiometric ratio) resulted in a booster station placement that gave the best overall performance measured in terms of expected population dosed.

Table 4.2: Trade-off analysis of optimal booster placements using Network 1 in terms of population dosed (number of people). Each row represents the evaluation of 10 optimally placed booster stations selected based on a particular stoichiometric ratio (ρ) and contaminant toxicity (first column) against seven other contamination scenarios with different ratios and toxicities (column 3-10). Second column provides the optimal objective value for the evaluated booster placement. ρ=0 represents Neutralization method.

Booster Design		Evaluation Scenarios								
		High	High	High	High	Low	Low	Low	Low	
Toxicity, ρ	Opt.	ρ=0	ρ=1	ρ=10	ρ=100	ρ=0	ρ=1	ρ=10	ρ=100	Mean
High, $\rho = 0$	3,397	–	4,884	8,522	12,897	792	1,061	1,465	1,540	4,320
High, $\rho = 1$	3,455	3,445	–	4,529	12,670	821	868	1,199	1,540	3,566
High, $\rho = 10$	4,319	3,510	3,615	–	10,711	822	899	1,204	1,537	3,327
High, $\rho = 100$	7,239	3,555	4,147	4,666	–	892	1,290	1,327	1,370	3,061
Low, $\rho = 0$	767	3,611	4,365	8,739	12,879	–	1,067	1,480	1,547	4,307
Low, $\rho = 1$	823	3,692	4,166	8,569	12,287	820	–	1,396	1,531	4,161
Low, $\rho = 10$	1,161	3,821	4,273	5,378	12,248	912	949	–	1,528	3,784
Low, $\rho = 100$	1,300	3,838	3,919	4,717	10,694	919	974	1,244	–	3,451

Table 4.3: Trade-off analysis of optimal booster placements using Network 2 in terms of population dosed (number of people). Each row represents the evaluation of 10 optimally placed booster stations selected based on a particular stoichiometric ratio (ρ) and contaminant toxicity (first column) against seven other contamination scenarios with different ratios and toxicities (column 3-10). Second column provides the optimal objective value for the evaluated booster placement. ρ=0 represents Neutralization method.

Booster Design		Evaluation Scenarios								
		High	High	High	High	Low	Low	Low	Low	
Toxicity, ρ	Opt.	ρ=0	ρ=1	ρ=10	ρ=100	ρ=0	ρ=1	ρ=10	ρ=100	Mean
High, $\rho = 0$	268	–	374	787	1,041	100	209	461	694	492
High, $\rho = 1$	297	290	–	632	1021	117	168	386	683	449
High, $\rho = 10$	351	285	337	–	924	118	161	238	587	375
High, $\rho = 100$	667	325	363	416	–	131	183	255	586	366
Low, $\rho = 0$	99	274	571	1,036	1,080	–	257	561	702	573
Low, $\rho = 1$	137	281	310	481	968	111	–	311	651	406
Low, $\rho = 10$	221	284	336	404	929	115	167	–	585	380
Low, $\rho = 100$	568	284	336	399	868	115	167	222	–	370



(a) 2 Sensors, Low toxicity

(b) 2 Sensors, High toxicity

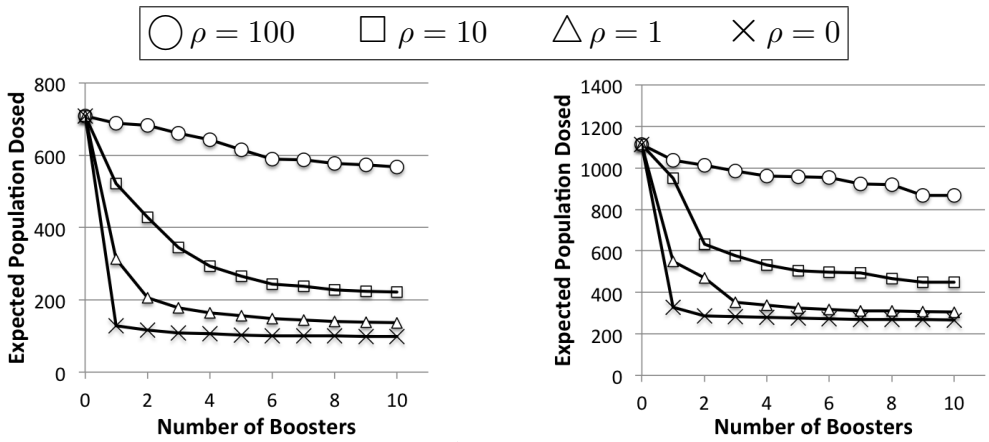
(c) 5 Sensors, Low toxicity

(d) 5 Sensors, High toxicity

(e) 10 Sensors, Low toxicity

(f) 10 Sensors, High toxicity

Figure 4.4.: Reduction in expected population dosed on Network 1, Left column: PD dose threshold (τ) of 0.0001 (high toxicity), Right column: PD dose threshold (τ) of 0.01 (low toxicity), Top row: 2 sensors, Middle row: 5 sensors, Bottom row: 10 sensors. ρ=0 represents Neutralization method.



(a) Low toxicity (b) High toxicity

Figure 4.5.: Reduction in expected population dosed on Network 2 with 5 sensors, Left: PD dose threshold (τ) of 0.0001 (high toxicity), Right: PD dose threshold (τ) of 0.01 (low toxicity). ρ=0 represents Neutralization method.

5. A REVIEW OF CONTAMINATION SOURCE IDENTIFICATION METHODS [1]

The optimal booster station placement problem discussed in the previous chapters is solved at the planning stage. The next aspect of the water network security pertains to devising a fast response system once a contamination event has been detected. Therefore, different levels of event response techniques have been proposed that include: (a) curtailing the spread of contaminant by isolating parts of the network, and (b) optimal flushing schemes to quickly and efficiently remove the contaminated water from the network. The effectiveness of these response techniques hinges on the information available about the source and extent of a contamination incident. Therefore, identification of the source of a contamination incident is a critical step to stop further ingress of the contaminant and begin control and cleanup.

The source identification problem is typically formulated as an inverse problem of finding the source of a contamination incident using the limited measurement data available from a sparse set of water quality sensors. Several researchers have proposed different methods to solve this problem. Almost all of these methods can be broadly categorized based on the following characteristics:

- Modeling Approach: There are variety of techniques used to model the water quality or input-output behavior in the water distribution network. These can consist of explicit model equations embedded directly into the problem formulation, use of existing simulator as a black-box model (e.g. EPANET (Rossman, 2000)), or surrogate models like binary trees or neural networks.

[1]Part of this section is reprinted with permission from "Testing Contamination Source Identification Methods for Water Distribution Networks" by Seth, A., Klise, K.A., Siirola, J.D., Haxton, T., and Laird, C.D., 2015. to appear in Journal of Water Resources Planning and Management, Copyright 2015 by American Society of Civil Engineers.

- Formulation Framework: Various solution strategies and theoretical frameworks can be used to formulate and solve the inverse problem. These can be very diverse including optimization based methods, probability based methods, and data-mining and pattern matching techniques.

- Underlying Assumptions: Methods may also differ on the basis of various underlying assumptions they make. These can include whether they assume single or multiple simultaneous injections during a contamination scenario, type of measurement data available, length of candidate injections, reaction rate of the contaminant, etc.

A significant body of research exists describing different approaches for the source identification problem. Early work proposed optimization based methods that assumed the availability of concentration measurements form water quality sensors (Laird et al., 2005, 2006; Preis and Ostfeld, 2006). Simulation-optimization approaches that use a water quality simulator (e.g., EPANET) linked to a pattern-search method or a Genetic Algorithm (GA) have also been proposed (Preis and Ostfeld, 2007, 2008; Guan et al., 2006). Shang et al. (2002a) present a water quality modeling framework called the Particle Backtracking Algorithm (PBA) and suggest its application in the identification of unknown contamination sources in water distribution networks.

Laird et al. (2005) present a least-squares formulation that seeks to find the contamination source profile that minimizes the sum of squares of the difference between calculated and measured contaminant concentrations observed at water quality sensors. A major challenge associated with contaminant source identification is dealing with the non-uniqueness of the solution inherent in such kind of inverse problems. Given limited measurement information, there may be many nodes and contamination profiles that are able to reproduce the observed measurements. The authors later extend this approach to identify multiple contamination sources (Laird et al., 2006)

and to deal with the non-uniqueness inherent in the inverse problem (Laird et al., 2005).

An alternate approach is introduced in Preis and Ostfeld (2006), where a large number of contamination simulations are performed using EPANET to build an approximate model in the form of hybrid model trees. Once contamination has been detected, the source identification algorithm involves "climbing backwards" in the model tree and solving a Linear Programming (LP) problem at each step. Although this technique is shown to be accurate for source identification on small (10 Node) networks, the number of simulations required to build adequate model trees can become extremely large for realistic networks (e.g. 10,000 EPANET simulations required for 10 Node network).

There are also a number of simulation-optimization approaches where a water quality simulator (e.g. EPANET) is used as a black-box to perform contaminant source identification. Pattern-search methods, or Genetic Algorithms are common approaches for solution of these black-box problems. Evolutionary Algorithms (EAs) or Genetic Algorithms (GAs) are combinatorial search heuristics that operate on Darwin's evolutionary principles of selection, crossover and mutation. In the context of contaminant source identification, a candidate set of injection scenarios with different characteristics or genes (node, start time, duration, strength) is selected, which then undergoes a mixing and re-selection process based on the concept of survival of the fittest. Preis and Ostfeld (2007) demonstrate a method to perform source identification which links EPANET with a GA. The fitness function used for the selection process is sum of squares error between measured and modeled concentrations at sensor nodes. The real-time efficiency of this algorithm is highly dependent on the initial candidate set and therefore the authors later extend their work by building an input-output relationship matrix through an offline simulation process (Preis and Ostfeld, 2008). This matrix can then be used to get much better starting populations for the GA. In the later work, the authors also analyze the accuracy of the algorithm in the presence of imperfect sensors. Alternatively, Guan et al. (2006) provide an

online simulation-optimization approach where EPANET is employed as a black-box, however gradient information is provided using finite difference approximation, that can then be used to solve a least-squares optimization problem.

The ability of an source identification technique to correctly determine the true injection scenario is significantly limited by the accuracy, reliability, and placement of sensors in the contamination warning system (Tryby et al., 2010). Due to the lack of prior knowledge about the contaminant (chemical or biological), recent developments in contamination detection technology utilize fault detection approaches by monitoring standard water quality measures (e.g., pH, free chlorine, turbidity, conductivity) to provide a binary yes/no indication of the presence or absence of contamination in the network (EPA, 2010a; Oliker and Ostfeld, 2014; Zhao et al., 2014). Unlike most of the source identification methods discussed earlier in this chapter that assume availability of accurate contaminant concentration data, more realistic techniques that handle these limitations are necessary. The aforementioned EPANET-GA based algorithm of Preis and Ostfeld (2008) does consider three types of measurements - concentration, fuzzy (low, medium, high), and binary (yes/no) - and concludes that finding unique solutions to the source identification problem becomes more difficult going from complete concentration information to only binary information. Cristo and Leopardi (2008) provide a input-output model based source identification technique where the model is built by running large number of EPANET water quality simulations (the use of PBA is also suggested). Although this work assumes the availability of concentration information, the adverse effect of measurement error is considered to verify the robustness of the overall algorithm. Liu et al. (2011) extend the work of Zechman and Ranjithan (2009) by a presenting an adaptive evolutionary strategy linked with EPANET that considers binary measurements in the form of detection thresholds. This work also introduces hot-start capabilities in an EA to perform source identification for a real-time response application. The Contaminant Status Algorithm (CSA) of De Sanctis et al. (2009) utilizes binary measurement data and the input-output model generated from PBA to identify possible candidate in-

jection nodes as being safe, unsafe or unknown. CSA is one of the algorithms tested using the framework presented in this manuscript and the details of this algorithm are discussed in Chapter 8. Another source identification method tested in Chapter 8 was presented by Mann et al. (2012). This method uses an MILP formulation for performing source identification that incorporates a detection threshold to model discrete measurements.

Apart from the challenges associated with contaminant measurement, network modeling errors can also be introduced due to demand variability, inaccurate estimation of pipe friction factors, and contaminant reaction dynamics. In order to address these uncertainties, various researchers have proposed statistical approaches to this problem. Given a prior probability of a node being a contamination node, the Bayesian methodology developed by Propato et al. (2009) is designed to calculate the corresponding posterior probability by minimizing an entropy function that represent the amount of information available. Using a reduced version of the linear input-output model generated from PBA as constraints to the entropy minimization problem, the authors provide an analytical solution along with confidence intervals on posterior probabilities that capture the uncertainty and non-uniqueness associated with source identification. Liu et al. (2011) propose a similar entropy minimization approach that builds a Logistic Regression Model by running a large number of contamination simulations and then use this model to calculate probability values required to evaluate the entropy function. The technique demonstrated by Perelman and Ostfeld (2012) uses network clustering to represent a water distribution network as an acyclic graph. This simplified representation of the network requires smaller number of water quality simulations to calculate detection probabilities (probability that a particular sensor will detect a candidate contamination scenario) than are required in the minimization of an entropy function. See Wagner et al. (2015), Wagner and Neupauer (2013), and Wang and Harrison (2012) for more recent advancements in probabilistic approaches for contamination source identification.

A data mining approach that requires building a large database containing historical or simulated contamination scenario characteristics is proposed by Huang and McBean (2009). This database can be used in a real-time situation to estimate detection probabilities that are then used in a likelihood maximization method to identify the contamination source(s). Shen and McBean (2011) extend this work by implementing the simulation-based data mining process on a large-scale parallel computing architecture to perform thousands of Monte Carlo simulations that account for measurement and model uncertainties.

The majority of the source identification methodologies proposed in the literature assume measurement information coming from sensors placed at fixed locations around the network. Instead, using mobile sensors or manual sampling teams to dynamically choose measurement locations during a response to a contamination event has shown promising results (Mann et al., 2012; Eliades and Polycarpou, 2011). Mann et al. (2012) present two Mixed Integer Linear Programming (MILP) formulations where one performs source identification and the other is used in selection of manual sampling nodes that improve the performance of source identification.

5.1 Source identification problem definition

Here, we define the source identification problem being considered in this work. Measurement data is assumed to be available from a fixed number of sensors located at specific nodes in a network. An Event Detection System (EDS) provides discrete yes/no measurements that indicate the presence or absence of contamination in the water. One probable response to the initial detection can be to obtain additional grab sample measurements to confirm the contamination. We assume that these grab sample measurements are also discrete (yes/no). Therefore, using measurements from fixed continuous sensors and manual grab samples, the goal of performing source identification is to identify the candidate locations where contamination could have taken place. This inverse problem is solved considering a fixed historical time period called the *time horizon*, providing a measure of likeliness for all nodes. This measure

Table 5.1: Measurement data from the 5 sensor locations for the example source identification problem shown in Figure 5.1. Dots show continuous incoming measurements to EDS.

Time (HH:MM)	Node 15	Node 35	Node 109	Node 219	Node 253
23:45	.	.	.	.	.
24:00	0	0	0	0	0
24:15	0	0	0	0	0
24:30	0	0	1	0	0
24:45	0	0	1	0	0
25:00	0	0	1	0	0
25:15	.	.	.	.	.

is used to provide a ranking of all nodes where a higher value indicates a greater chance of being the contamination source. It is assumed that contaminant ingress can take place at any node (junctions, tanks, and reservoirs) in the entire network.

An example of how we formulate a typical source identification problem is demonstrated using the Net3 distribution network shown in Figure 5.1 (an example network from EPANET (Rossman, 2000)). This example network has 92 junctions, 3 tanks, and 2 reservoirs. The example shows a scenario where a contamination injection takes place at node 111 (at time 24:00). The EDS, which has 5 fixed water quality sensors (marked as squares in Figure 5.1) gathers binary measurements at a 15 minute frequency as shown in Table 5.1. A positive measurements at sensor node 109 indicate a contamination incident. Given the measurement information in Table 5.1 and a candidate injection time horizon of 10 hours, we define the goal of source identification to determine the list of possible source locations and values by which to rank their likeliness. For instance, the probability-based source identification method described in the next chapter produces the results shown in Table 5.2. The table contains a list of possible injection nodes sorted by their corresponding probability of being the true injection node.

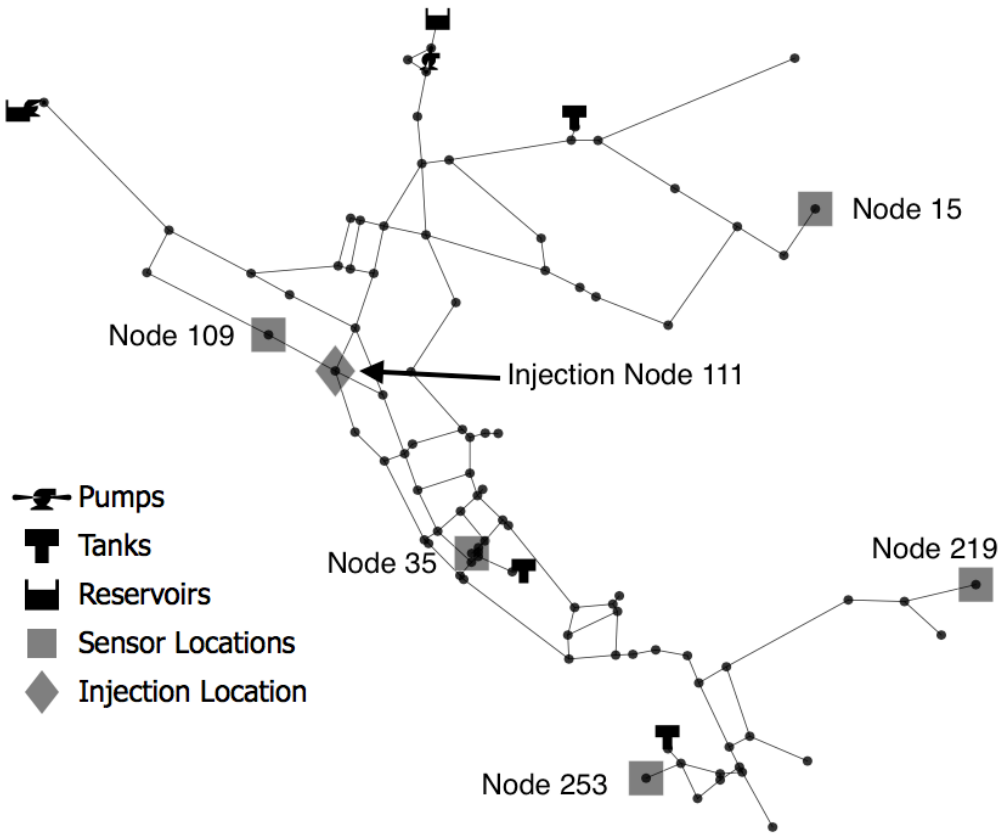


Figure 5.1.: An example of a typical source identification problem using EPANET Net3.

Table 5.2: A typical example of source identification results obtained for the example problem shown in Figure 5.1.

Node	Probability
109	0.17
111	0.17
113	0.17
115	0.17
117	0.17
120	1E-8
193	1E-8
195	1E-8
.	.
.	.

6. BAYESIAN PROBABILITY-BASED SOURCE IDENTIFICATION METHOD AND OPTIMAL SAMPLING

Unlike the optimal booster station placement problem, which is solved at the planning stage, source identification needs to be performed in real-time as an incident unfolds. Therefore, computational speed of a source identification technique is critical. Keeping this in mind, in this chapter we propose a Bayesian probability-based source identification method that takes advantage of the fast water quality simulation framework proposed by Mann et al. (2012a). A case study performed on a large-scale network with over 12,000 nodes highlights the computational speed of the Bayesian-probability based method. The performance of a source identification method depends on many factors including size of network model, measurement error, modeling error, time and number of contaminant injections, and sensor density and placement. Therefore, in the next chapter (Chapter 7), we propose a testing methodology to compare the performance of the Bayesian probability-based method and two other methods from the literature.

6.1 Bayesian probability-based method

This method operates by simulating all candidate contaminant injections and then calculating the probability of each injection based on how well the simulated measurement profile matches the true measurements obtained from the sensors. The probability calculations are performed using Bayes theorem:

$$P(i|m) = \frac{P(m|i)P(i)}{P(m)} \qquad \forall\, i \in \mathbf{C} \tag{6.1}$$

Where $\mathbf{C}$ is a set of all possible contamination incidents. An incident can start at any node and any time step, and is assumed to continue for the complete simulation duration (i.e., continuous injections are assumed). $P(i|m)$ is the probability of incident i given a vector of measurements m. $P(i)$ is the prior probability of contamination incident i. This formulation assumes that only a single injection incident is possible, and therefore $P(i)$ is set to a uniform prior that is the inverse of the cardinality of $\mathbf{C}$ given by $1/|\mathbf{C}|$. Since an estimate of $P(m)$ (the prior probability of the observed measurement) is generally not available, it is common to replace this calculation by normalizing the calculated values of $P(i|m)$ so that they sum up to one. Finally, $P(m|i)$ is the probability of measurement m given injection incident i. It is calculated using the following equation:

$$P(m|i) = (1 - p_f)^{match(i)} p_f^{num\_meas - match(i)} \tag{6.2}$$

Where, p_f is a user specified estimate of measurement failure probability (false positive or false negative), num_meas is the total number of available measurements, and $match(i)$ is the number of actual discrete measurements that match the discrete measurements obtained by simulating incident i.

The overall algorithm for this method is as follows. Following detection, hydraulic simulations are performed (using EPANET) for a specified time window preceding the detection time and the flow data is used to build the linear input-output water quality model, Merlion (Equation 1.9). Next, the set of candidate injections, $\mathbf{C}$, is populated by analyzing the input-output model and choosing injection node-time pairs that are hydraulically connected only to the positive measurements. Next, all the candidate injections are simulated using Merlion to obtain simulated measurement profiles that are then compared to the actual measurement profile to get the number of matches. The posterior probability of each injection node-time pair is calculated using Equations 6.1 and 6.2. Finally, the probability of each node being the injection node is chosen as the highest posterior probability value over all injection node-time

pairs containing that particular node. This posterior probability is used as a measure of likeliness of a node being the true injection node.

The majority of the computational time in the above algorithm is spent on performing simulations of the candidate injections in set C. We harness the fast simulation capabilities of Merlion water quality modeling framework proposed by Mann et al. (2012a) to perform these simulations. The framework provides a custom linear solver that is optimized for performing a large ensemble of water quality simulations.

6.2 Greedy grab sampling algorithm

Immediately following the initial detection of a contamination incident, measurement information is typically very limited. Given the limited measurement information, the source identification problem is often non-unique, with many possible solutions. Since the source identification problem is ill-posed, using limited measurement information can result in a large number of likely source locations. To further refine the results of source identification, water utilities can send out manual sampling teams to obtain additional measurements. Since the utilities are constrained by the number of sampling teams and the time it takes to mobilize them, intelligent selection of sampling locations is important. Wong et al. (2010) propose an optimization formulation that selects grab sampling locations to maximize the pair-wise distinguishability between candidate incidents. In this work, an analogous greedy algorithm is proposed to iteratively select grab sampling locations that provide maximum pair-wise distinguishability.

The greedy algorithm can be explained with the help of a simple example network shown in Figure 6.1.

The first part of the algorithm, which involves data generation, is identical to the optimization-based method proposed by Wong et al. (2010). It involves simulating the candidate incidents to generate an *Impact matrix*, which is then used to create sets of pairwise incidents that are distinguished by each sample location. For example, if we consider candidate incidents at all six nodes in Figure 6.1, then the Impact

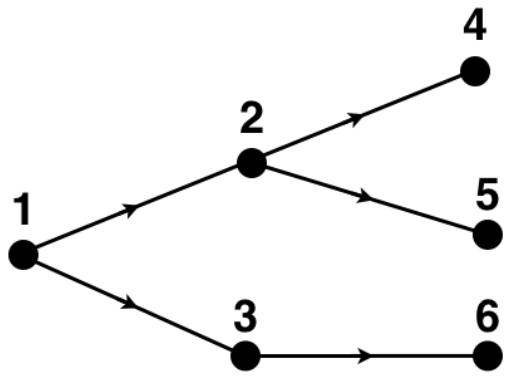


Figure 6.1.: Illustrative six node example network. Arrows represent flow direction.

matrix can be generated as shown in Table 6.1. The rows in the Impact matrix represent incident locations and the columns represent candidate sampling locations. Each incident is simulated up to an estimated future *sample time* when the manual samples will be drawn. If a simulated incident results in a positive measurement at a candidate sampling node (base on a concentration threshold), then the corresponding value in the impact matrix is 1, otherwise it is 0.

Using the Impact matrix (Table 6.1), the set of distinguishable pairs for each sampling location can be identified as shown in Table 6.2. The basic idea here is that a sampling location will be able to distinguish between two incidents if one incident will lead to a positive measurement while the other will result in a negative.

The greedy sampling algorithm proceeds by selecting the sampling location that will distinguish the highest number of incident pairs. Ties are broken arbitrarily. For instance, in our example node 4 can distinguish 9 pairs of incidents and is selected as the first sampling location. Next, all the pairs of events distinguished by node 4 are removed from Table 6.2 to obtain Table 6.3. Again, the next sampling team can select node 3, 4, or 5 because they will all distinguish the highest number of remaining incident pairs. The algorithm continues to select sampling location until

Table 6.1: Impact matrix for injection at all nodes in Figure 6.1

Incident Locations	Sampling Locations 1	2	3	4	5	6
1	1	1	1	1	1	1
2	0	1	0	1	1	0
3	0	0	1	0	0	1
4	0	0	0	1	0	0
5	0	0	0	0	1	0
6	0	0	0	0	0	1

Table 6.2: Sets of distinguishable incident pairs based on Impact matrix in Table 6.1.

	Sampling Locations 1	2	3	4	5	6
	1-2	1-3	1-2	1-3	1-3	1-2
	1-3	1-4	1-4	1-5	1-4	1-4
	1-4	1-5	1-5	1-6	1-6	1-5
	1-5	1-6	1-6	2-3	2-3	2-3
	1-6	2-3	2-3	2-5	2-4	2-6
		2-4	3-4	2-6	2-6	3-4
		2-5	3-5	3-4	3-5	3-5
		2-6	3-6	4-5	4-5	4-6
				4-6	5-6	5-6
# of pairs	5	8	8	9	9	9

Table 6.3: Sets of distinguishable incident pairs after node 4 has been picked as the first sampling location.

	Sampling Locations				
	1	2	3	5	6
	1-2	1-4	1-2	1-4	1-2
	1-4	2-4	1-4	2-4	1-4
			3-5	3-5	3-5
			3-6	5-6	3-6
# of pairs	2	2	4	4	4

all the sampling teams have been deployed. Note that before the algorithm begins, pairs distinguished by fixed water quality sensors are removed from the analysis to avoid double counting.

6.3 Source identification case study on large-scale network

The Bayesian-probability based source identification method and the greedy sampling algorithm have been incorporated in US EPA's Water Security Toolkit (WST). Along with results of the case study performed in this section, we provide the details of the parameters necessary to reproduce this case study with WST.

The ability of the proposed Bayesian-probability based source identification method and the greedy sampling algorithm to quickly narrow down the contamination source is demonstrated with a simulated case study on a large-scale water network model. A mass injection is simulated at Junction-6632 of the BWSN (Battle of the Water Sensor Networks) Network 2 (12,523 Nodes) (Ostfeld et al., 2008) shown in Figure 6.2. The injection takes place at 8:00 AM in the morning, which is 8 hours from the simulation start time, and gets detected at 9:45 AM when a fixed water quality sensor

goes off at Junction-12325. An additional positive measurement is obtained at 10:00 AM before the source identification procedure is started.

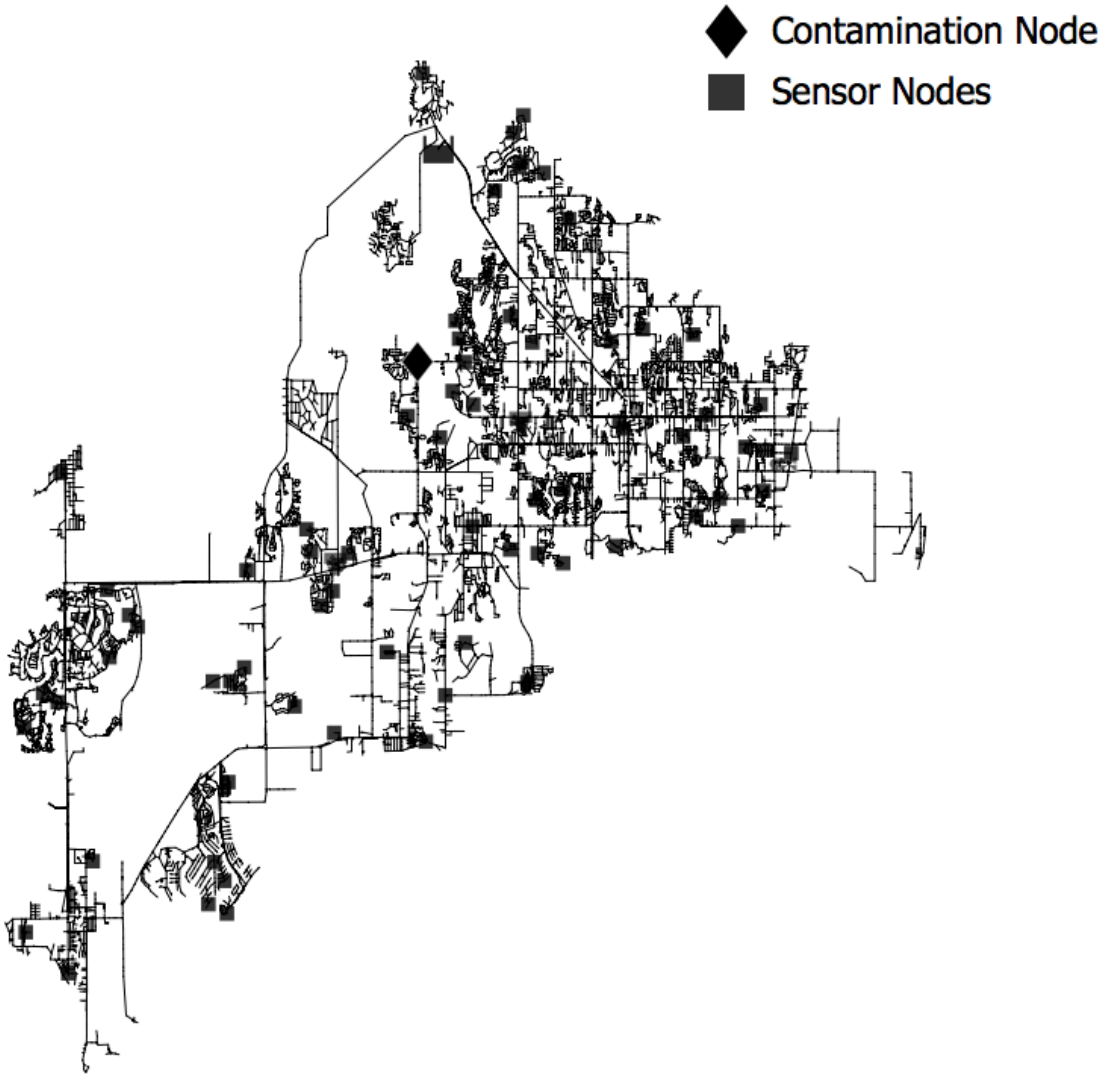


Figure 6.2.: BWSN Network 2 diagram with contamination location and sensor locations.

For this case study, the EDS is composed of 130 fixed water quality sensor that were placed using the *sp* module of WST. This module provides an optimization-

Table 6.4: Parameters used for placing fixed water quality sensors in BWSN Network 2.

Specification Parameter	Value(s)
Injection Scenario Locations	Non-zero demand nodes
Injection Scenario Start Time	12 AM
Injection Scenario Duration	24 Hours
Injection Scenario Strength	10 grams/minute
Hydraulic Time Step	1 Hour
Water Quality Time Step	15 min
Objective	Total Population Exposed

based sensor placement techniques that can be used to minimize impact over a set of contamination scenarios. The population exposed impact metric was used for the optimal sensor placement. Details of all the parameters used for the sensor placement are provided in Table 6.4. The measurement frequency of these sensors is assumed to be fixed (15 minutes) and for simplicity, the detection threshold is set to 0 mg/L.

At 10:00 AM source identification is performed using the Bayesian probability-based method, which results in 72 potential source locations. The parameters used for the Bayesian probability-based source identification method and the greedy grab sampling algorithm are provided in Table 6.5.

For this case study, we assume that it takes 1 hour for three teams to gather and analyze manual samples. Therefore, using the greedy algorithm, the sampling locations are determined for a sample time 1 hour into the future. At 11:00 AM, new measurements are available from three sampling locations along with additional measurements from fixed sensors. This new information is used to again perform source identification resulting in 30 candidate source locations. The source identification and sampling cycle is repeated every hour until 1:00 PM, when the possible source locations have been narrowed down to 4 nodes that include the true source location - Junction-6632.

Table 6.5: Parameters used for the Bayesian probability-based source identification method and the greedy grab sampling algorithm.

Specification Parameter	Value(s)
Time Horizon	24 Hours
Measurement Failure Probability (p_f)	0.1
Measurement Detection Threshold	0 mg/L
Feasible Source Nodes	All nodes
Cumulative Probability Cutoff	0.95
Manual Sampling Time Delay	1 Hour
Feasible Sampling Locations	Non-zero demand nodes

Figure 6.3 provides the number of candidate source locations, and the computation times for the source identification and grab sample selection calculations during each cycle. All computations were done in serial on a machine with 24 Intel(R) Xeon(R) processors (E5-2697 v2 @ 2.70GHz). The computational speed is made possible by selection of appropriate data structures to avoid cache misses. As we can observe from Figure 6.3, only 3 sampling cycles were necessary to sufficiently narrow down the candidate source locations. As more measurement data was gathered, the computational time of the Bayesian probability-based method went up and as the number of candidate injections went down, the computational time of the greedy grab sampling algorithm also went down. However, the overall time for both these calculations between each cycle, which needs to be performed as soon as new measurement data is available to quickly deploy the sampling teams, was less than 3 minutes. These results imply that the proposed methodology for source identification and grab sampling location selection are a viable approach for real-time response to a contamination incident in large-scale networks.

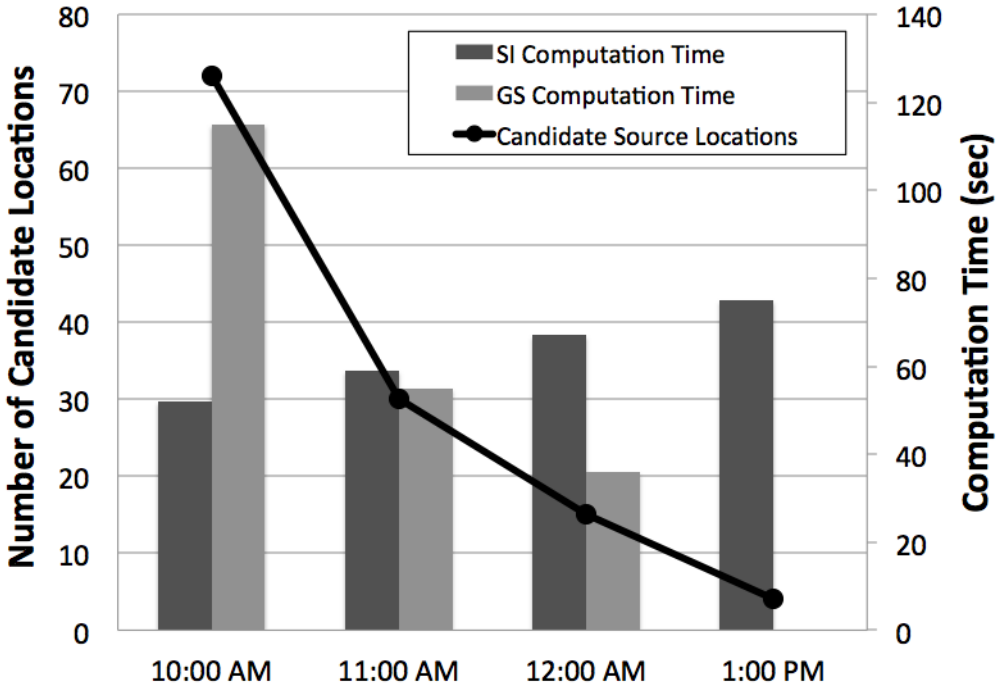


Figure 6.3.: Performance of the Bayesian probability-based method and the greedy grab sampling algorithm. Left axis indicates number of candidate source locations. Right axis indicates overall computation time for the source identification (SI) and grab sampling (GS) calculations.

7. TESTING METHODOLOGY FOR CONTAMINATION SOURCE IDENTIFICATION METHODS [1]

In Chapter 5 we discussed a significant body of research describing different approaches for the source identification problem. In order to contrast and compare the variety of source identification methods available in the literature, each having there own advantages and disadvantages, in this chapter we propose a testing methodology for source identification methods. Using this testing methodology in the next chapter, the performance of three different type of source identification methods is evaluated on a wide range of realistic scenarios (e.g., measurement and modeling errors).

7.1 Performance Metrics

In this chapter, the following assumptions are made in regards to the source identification problem. It is assumed that an EDS provides discrete yes/no measurements that indicate the presence or absence of contamination in the water. All sensors in the network are assumed to provide measurements at the same constant frequency. Note that although this assumption is used here, it is easy to relax this assumption for any of the tests. For example, the source identification methods studied in this work can make use of the information provided from manual grab sample measurements, however, the test scenarios discussed in this chapter do not consider these kind of measurements. It is assumed that contaminant ingress can take place at any node (junctions, tanks, and reservoirs) in the entire network, and both single and multiple injections can occur. Additionally, it is assumed that source identification methods

[1]Part of this section is reprinted with permission from "Testing Contamination Source Identification Methods for Water Distribution Networks" by Seth, A., Klise, K.A., Siirola, J.D., Haxton, T., and Laird, C.D., 2015. to appear in Journal of Water Resources Planning and Management, Copyright 2015 by American Society of Civil Engineers.

output a list of possible injection nodes along with their corresponding *measure of likeliness*. The likeliness measure is dependent on the source identification method being used. For example, for a probability-based method that reports a probability of a node being the true injection node, this probability value is used as the measure of likeliness. The measure of likeliness used for each source identification methods studied in this chapter is described later in the section describing these methods.

To measure the performance of a source identification method, two important criteria must be considered. First and foremost, a method should be able to correctly identify the true injection location(s) as the most likely location(s). Secondly, a method should be able to distinguish the true injection location(s) from the rest of the candidate nodes as effectively as possible. To this end, Yang and Boccelli (2014) introduce two performance metrics - *Accuracy* and *Specificity* - and this work uses modified versions of these metrics as shown in Equations 7.1 and 7.2.

$$Accuracy\,(\%) = \frac{Likeliness\ measure\ of\ the\ true\ injection\ node}{Highest\ likeliness\ measure\ over\ all\ candidate\ nodes} \times 100 \tag{7.1}$$

$$Specificity\,(\%) = \frac{Number\ of\ nodes\ with\ lower\ likeliness\ than\ true\ injection\ node}{Total\ number\ of\ candidate\ nodes} \times 100 \tag{7.2}$$

Here, a 100% accuracy indicates that the true injection node had the highest likeliness value, while a high value of specificity indicates a high rank of the true injection node among all the candidate nodes. For example, if the true injection node is given a likeliness value of 5, and the remaining nodes are 2, then the accuracy is 100%, and the specificity is as close to 100% as possible. However, it is possible to have a high accuracy with low specificity. For example, if a method returns a likeliness value of 2 for all nodes, the accuracy value is still 100%, but the specificity value is zero. Although Equations 7.1 and 7.2 are defined for scenarios with only a single true injection node, in the case of multiple injection nodes, multiple values of both metrics are calculated with respect to each true injection node.

Table 7.1: Standard specifications used for generating test sets.

Specification Parameter	Value(s)
Network	EPANET example Net3
Injection Nodes	111, 151, 189, 183, 229
Sensor Locations	149, 117, 167, 213, 253
Injection Start Time	24 hours into simulation
Injection Length	10 hours
Measurement Start Time	10 hours before detection
Measurement End Time	2 hours after detection
Sensor Frequency	2 measurements per hour
Time Horizon	24 hours

7.2 Factors effecting source identification and testing methodology

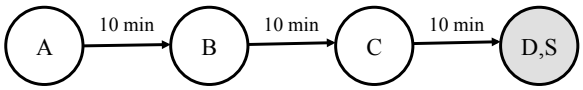
In this section, some of the challenges inherent for source identification methods in a real-time response system are introduced. The following subsections describe a category of test cases designed to demonstrate the effectiveness or identify limitations of different source identification methods. Most of the test cases presented in this chapter are created using Net3 (an example network from EPANET). Table 7.1 provides the standard specifications used in generating these test cases. Some test cases require varying these standard specifications, while most of the test cases require additional specifications discussed in the corresponding subsections.

7.2.1 Preliminary Tests

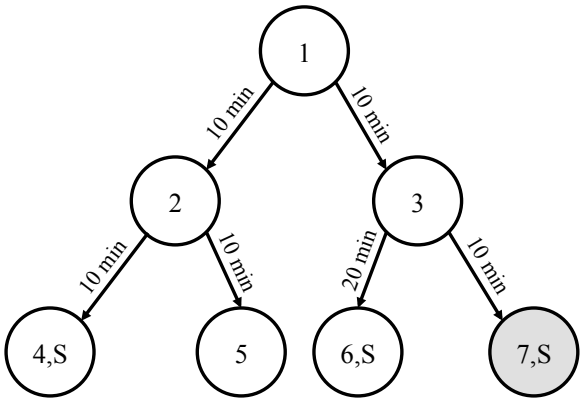
The testing framework provides a couple of preliminary tests using small simple networks. For these tests, the analytical solution to source identification problem is known, and therefore, these tests can be used to validate the behavior a source identification method during or after its development process.

The first test in this set is created using the simple four node linear network shown in Figure 7.1a. The demand patterns in this network are calibrated to ensure that

the time delay between each node is equal to 10 minutes. A continuous injection is simulated starting at node A at time zero, which is detected at sensor node D at 40 minutes. For this test, with a single sensor, all nodes are equally likely candidates and indistinguishable. For this test, the accuracy should be 100%, and the specificity should be 0%. The second test in this set comprises a seven node binary tree network shown in Figure 7.1b. A contaminant is injected at node 3 at time step 3 and detected at node 7 at time step 4. Here, the solution to the source inversion problem should indicate node 3 and 7 as the most likely source nodes. For this test, the accuracy and specificity should be 100% and 71% respectively.



(a) A four node network linear network. A contaminant injection is simulated at node A and is detected at the sensor placed at node D.



(b) A seven node tree network. A contaminant injection is simulated starting at node 3 and is detected at the sensor at node 7.

Figure 7.1.: Simple network structures used to create basic source identification tests with know analytical solutions.

7.2.2 Measurement Error

Various researchers have demonstrated the impact of measurement error on their source identification method's performance. Water quality sensors can encounter measurement error in the form of both false positives and false negatives. In general, false positives can lead an source identification calculation to identify an unnecessarily large number of candidate nodes, while false negatives could decrease the likelihood of identifying the true contamination node(s).

In order to test whether an source identification method is reliable in the likely occurrence of both false positive and false negative measurements, a set of tests with a range of false positive and false negative rates is used. These tests are generated by simulating contamination scenarios in EPANET Net3 at five different nodes located in different parts of the water distribution network. The specifications provided in Table 7.1, with modified measurement start and end times, are used in designing this test set. For each scenario, the binary measurements from five sensors (optimally placed using the *sp* module in WST) are collected over a 12-hour span starting 8 hours before the detection time and ending 4 hours after the detection time. Next, measurement error is artificially introduced by randomly selecting which measurements are in error based on the specified false positive rate (FPR) and false negative rate (FNR). All permutations of FPR and FNR values in the set - [0, 0.1, 0.2, 0.3, 0.4] are chosen in designing these tests. Finally, each combination of FPR and FNR is sampled 50 times to obtain statistics. To summarize, each test is identified by an injection node, an FPR value, an FNR value, and a sample number, giving a total of 6005 tests (FNR=FPR=0 does not require multiple samples).

7.2.3 Modeling Error

Due to the limited availability of data for model tuning and a lack of real-time demand information, error can be expected between a network model and the true flow fields in the distribution system. Therefore, it is important to include test cases

that are designed to assess the performance of a source identification method in the presence of network modeling errors.

Demand variability is a naturally stochastic phenomenon that is challenging to estimate especially with the lack of real-time data. Typically, source identification methods incorporate a hydraulic model that uses estimated demand patterns to model the flow rates and directions inside each pipe in the water distribution network. Therefore, errors in demand estimates are propagated to errors in modeled flow rates and directions. Furthermore, errors in flow direction can have a drastic impact on the results by potentially switching the set of upstream nodes which could contain the true injection node(s). Inaccurate flow rates can also impact source identification results by shifting the estimated time profile of the contaminant at a sensor node. To evaluate a method in the presence of modeling error, a set of test cases are generated with different levels of demand variability between the model used to simulate the contamination incident and the model used to perform source identification. This produces error between the model used by the source identification method and the measurement data. Details of the specifications used in designing this test set are provided in Table 7.1. In this set of tests, a base case model is used to generate the measurement data. To form the base case, the demands of all nodes of Net3 are reduced by 20%. This is done to avoid infeasibly high demand values when random error is added to the system. Base cases are formed from each of the five contamination scenarios described earlier, and the simulations are used to obtain the measurement data. Next, random demand error over a range of error percentages - [1, 2, 4, 8, 10, 20] - is added to this base case to form "erroneous" models given to the source identification methods. This is achieved by generating model input files (EPANET INP format) containing error in the base demand values at each node. Finally, for every error range value, multiple samples (50) are taken to generate a number of different input files. Hence, each test in this set consists of measurements obtained for a particular injection scenario, a model input file with a particular error range, and a sample number, giving a total of 1500 test cases.

7.2.4 Injection Characteristics

Because of the limited number of sensor measurements, the source identification problem is inherently non-unique (many possible locations and/or injection profiles). Most source identification methods acknowledge these limitations and some methods impose additional constraints on the characteristics of a possible solution. Two such characteristics are the number of simultaneous injection locations considered during a contamination scenario, and the length of the contaminant injection at a node.

While an algorithm that assumes a single injection location might be able to more accurately identify the source in a single source test, it is important to know how it performs if multiple injections occur. The accuracy of methods that do not make this assumption should be evaluated. Therefore, this study includes test cases where the number of injection nodes in a scenario is varied from a single injection node to three injection nodes in Net3. Details of the specifications used in designing this test set are provided in Table 7.1. Three injection scenarios are generated - single injection at node 151, two simultaneous injections at nodes 111 and 151, and three simultaneous injections at nodes 111, 151, and 189. In the case of multiple injections, the time horizon is chosen in reference to the longest detection time of any of the individual injections.

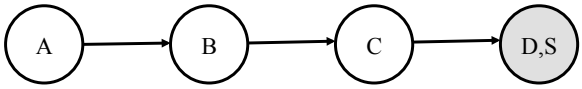
This study also incorporates tests to investigate the capabilities of a source identification method in identifying contamination scenarios of varying injection lengths. In the presence of measurement and modeling errors, longer injections are generally more distinguishable compared with shorter injections, which could be completely missed by periodic sampling of the sensors or produce a short pulse of positive measurements that can be more difficult to assess. The details of the first set of tests are provided in Table 7.1 with the exception of the injection length, which is varied over a range. For each injection node, this range contains the following injection lengths (in minutes) - [60, 120, 240, 480, 720]. In order to capture the difference in measurement profiles produced by the injections of different lengths, the measurement end time

is increased to include 12 hours following the initial detection. Another set of tests that contain the same simulated injections (nodes and injection lengths) but with added measurement error (FPR=10%, FNR=10%) is also included in this test set. The test cases with measurement error are run 50 times with different random seeds to obtain performance statistics. Each case in both injection length test sets (with and without error) is identified by an injection length, an injection node, and (in case of measurement error) a sample number. Therefore, this category contains a total of 1275 tests.

7.2.5 Time Horizon

In order to keep the size of the source identification problem reasonable, the calculations are typically performed by limiting the window of time under consideration. This is generally called the analysis time horizon or the time horizon. While reducing the time horizon can save computational expense, if it is too short, it falsely limits the space of potential injection locations. The impact of the time horizon on source identification calculations is explained using a simple example shown in Figure 7.2. The figure shows a four node linear network with a sensor at the terminal node D and a table listing connections between injection node-time pairs and the sensor detection times. For example, the top left entry in the table indicates that an injection at node A at time 0 hours will be witnessed at the sensor node D at time 3 hours, and so forth. To illustrate the significance of analysis time horizon, two cases should be considered. In the first case, an injection takes place at node D at time 4 hours and is detected by the sensor at node D at the same time. If the time horizon is 1 hour and the table in Figure 7.2 is used to investigate a detection at node D at 4 hours, the possible injections can be determined to be C at 3 hours and D at 4 hours. Using the same injection and detection scenario, the second case uses a time horizon of 3 hours and determines the candidate injections as node D at 4 hours, node C at 3 hours, node B at 2 hours and node A at 1 hours. Notice that in the second case, the

larger time horizon has lead to a larger number of candidate nodes, thereby affecting the specificity of the source identification calculations.



A,0 →D,3	A,1→D,4	A,2→D,5	A,3→D,6
B,0→D,2	B,1→D,3	B,2→D,4	B,3→D,5
C,0→D,1	C,1→D,2	C,2→D,3	C,3→D,4
D,0→D,0	D,1→D,1	D,2→D,2	D,3→D,3

Figure 7.2.: Four node linear network with a sensor at Node D. Time delay between each node is assumed to be 1 hour. Table shows the node-time relationships between all nodes and the sensor node.

Picking the right time horizon is not straight forward, ideally it should be at least as big as the longest flow path to any sensor in the network. However, for realistic large-scale networks this can limit the efficiency of real-time source identification calculations. Therefore, good algorithms need to be aware of the limitations imposed by a selected time horizon and indicate these limitations in the results they produce. This study includes tests that vary the time horizon used by a method. Details of this test set are provided in Table 7.1, with the exception of the time horizon that is varied over the set - [1, 2, 4, 8, 16, 24] (hours). Each case in this test set consists of an injection node and a time horizon used to perform source identification for that injection, giving a total number of 30 tests.

7.2.6 Network Size

The size and complexity of the water distribution network can impact the performance of a source identification method. One major factor is the non-uniqueness of the solution, which, for a fixed number of sensors, increases significantly with the network size. Not only can the quality of the solution to the source identification problem be negatively impacted as the network size increases, but the computational

effort in terms of solution time and memory requirement can also increase substantially. A set of tests are provided that contains water distribution networks of three different sizes - EPANET Net3 (97 Nodes), Network2 (3,358 Nodes) (Watson et al., 2009) and BWSN (Battle of the Water Sensor Networks) Network 2 (12,523 Nodes) (Ostfeld et al., 2008).

Figures 7.3, 7.4, and 7.5 provide a graphical representation on EPANET Net3, Network2, and BWSN Network 2 respectively. All three figures show the fixed sensors placed using U.S. EPA's Water Security Toolkit and the injection locations used to obtain average performance statistics when performing source identification.

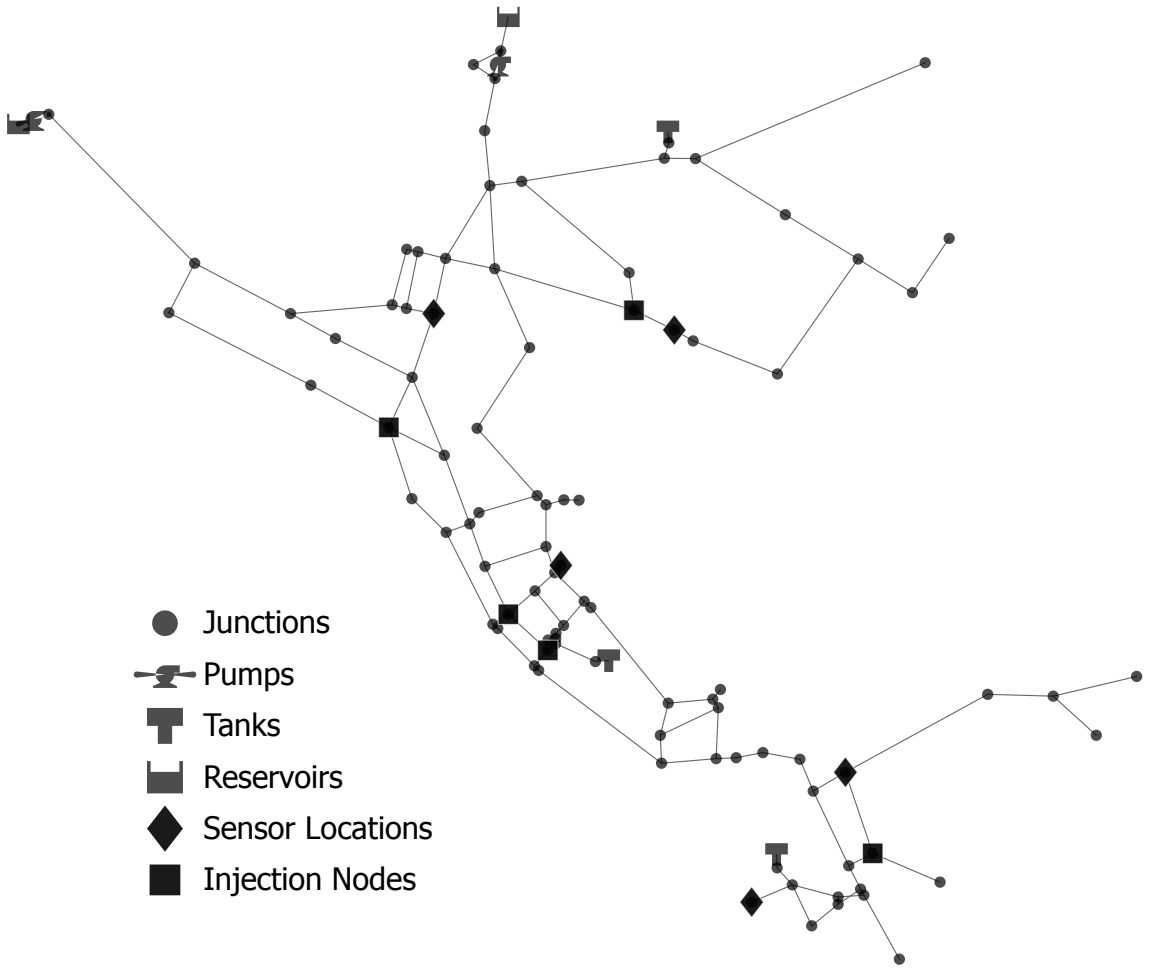


Figure 7.3.: EPANET Net3 with fixed sensor locations and injection locations.

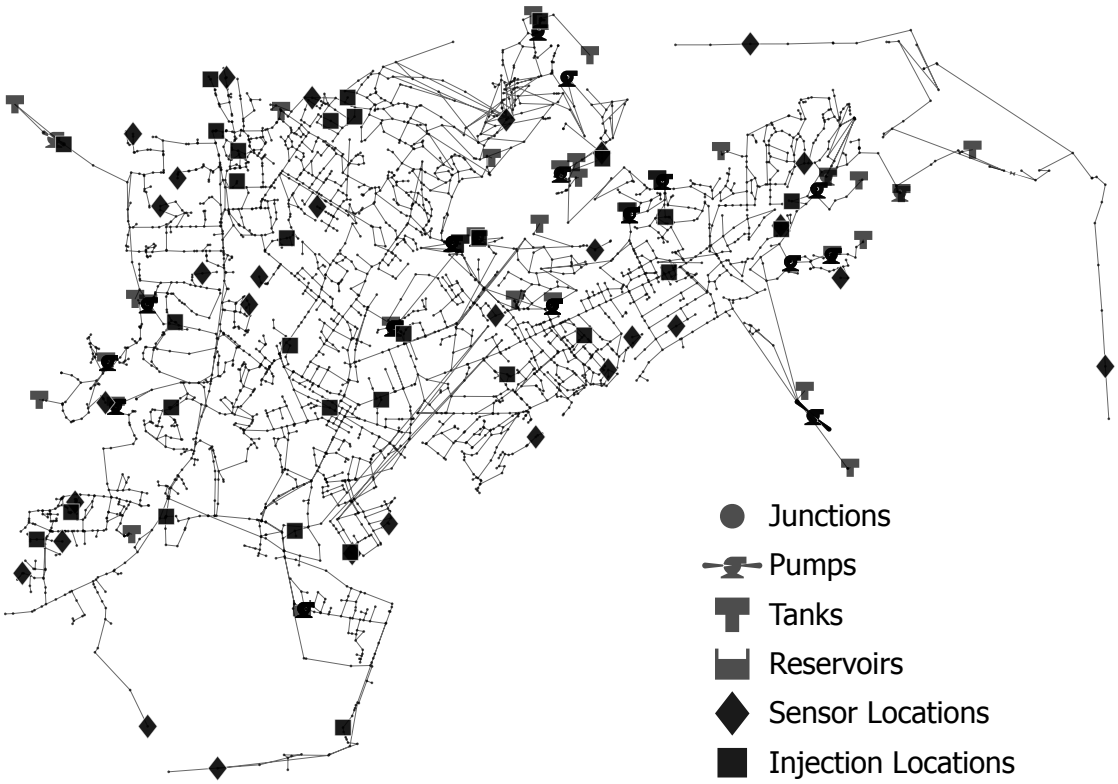


Figure 7.4.: Network2 with fixed sensor locations and injection locations.

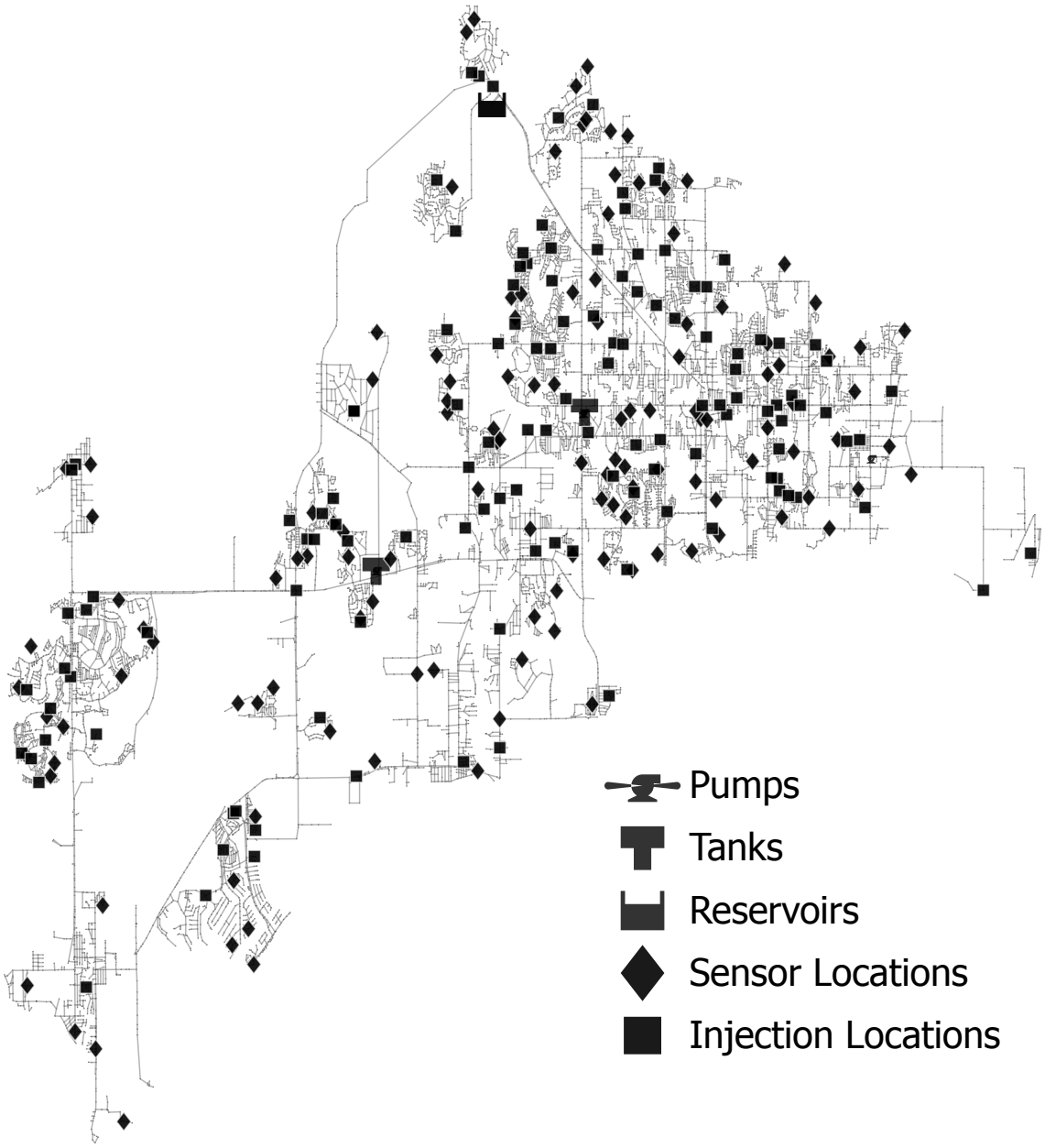


Figure 7.5.: BWSN Network 2 with fixed sensor locations and injection locations.

To obtain average performance statistics for each of these networks, injection scenarios are simulated by injecting a contaminant at several nodes selected from different parts of that network. Note that for each of these scenarios, all the nodes

Table 7.2: Specifications used for generating the network size test set.

Specification Parameter	Value(s)
Number of Nodes	97, 3,358, 12,523
Number of Injection Scenarios	5, 30, 30
Number of Sensors	5, 30, 130
Injection Start Time	12 hours into simulation
Injection Length	10 hours
Measurement Start Time	10 hours before detection
Measurement End Time	2 hours after detection
Sensor Frequency	2 measurements per hour
Time Horizon	12 hours
Total Number of Test Cases	65

in a network are considered as candidate injection nodes for the source identification methods. The number of injection scenarios selected for each network is provided in Table 7.2 along with other specifications used in designing this test set.

7.2.7 Sensor Placement

The cost of buying, operating, and maintaining water quality sensors limits the number of sensors that can be installed in a water distribution network. Therefore, researchers have proposed optimal placement of fixed sensors to minimize the impact on the population or the network infrastructure due to a contamination incident (Berry et al., 2005a; Ostfeld and Salomons, 2004b; EPA, 2010a). Although optimal sensor placement is a separate problem to source identification, the location of these sensors can have a major impact on the performance of source identification methods. Only a few papers have investigated sensor placement for better source identification. Tryby et al. (2010) provide a sensor placement technique that is designed to reduce the ill-posedness of the source identification problem. An approach of dynamically selecting manual grab sample locations to improve distinguishability between can-

didate source nodes has also been proposed in Mann et al. (2012). However, other objectives typical for sensor placement might not be appropriate for source identification. For example, a typical maximum coverage objective would place a sensor that detects multiple scenarios, which would reduce a source identification methods ability to distinguish between those scenarios.

The sensor density in a network is defined as the percentage of nodes where a water quality sensor is located. This corresponds to the amount of information available for source identification. A good source identification method should be able to accurately identify candidate contamination nodes with limited information. To investigate the impact of sensor density and layout, two different sets of tests are provided. The first test set varies the density of optimally placed sensors, while the second test set varies the density of sensors that are randomly placed at nodes around the network. Two networks of different sizes are used for this test set: EPANET Net3 (97 Nodes) and Network2 (3,358 Nodes). For Net3, apart from the sensor nodes, the rest of the specifications used in designing the optimal and random sensor placement test sets are provided in Table 7.1. The tests for Network2 use the same set of parameters with different injection nodes. The list of sensor placements are selected by varying the sensor density over the set [2%, 4%, 6%, 10%, 20%] for Net3 and [0.2%, 0.4%, 0.6%, 0.8%, 1%] for Network2.

For the test set with optimally placed sensors, sensor placement is performed using WST with the objective set to minimize population exposed. To summarize, each case in both test sets (optimal and random) is identified by a network, a sensor density and an injection node, for a total of 100 test cases.

8. COMPARATIVE STUDY AND SENSITIVITY ANALYSIS OF SOURCE IDENTIFICATION METHODS [1]

The testing methodology described in the previous chapter (Chapter 7) is used to compare the performance of three different source identification methods. Additionally, pairwise sensitivity analysis is performed for some of the tests cases to analyze the effect of two factors (described in Section 7.2) at a time.

8.1 Overview of methods studied

The Bayesian probability-based method proposed in Chapter 6 is the first method studied. The following subsections describes the Contaminant Status Algorithm (De Sanctis et al., 2009) and an optimization-based method (Mann et al., 2012). These two methods are briefly overviewed and readers are referred to their respective publications for more details. The underlying assumptions of each method is highlighted to help explain the performance results presented later.

8.1.1 Contaminant Status Algorithm

The Contaminant Status Algorithm (CSA), proposed by De Sanctis et al. (2009), performs source identification by assigning status to each candidate node-time pair as either being safe (not an injection candidate), unsafe (possible injection candidate), or unknown. Since the performance metrics calculation requires the results of a source identification method to be in the form of a list of candidate nodes with their corre-

[1]Part of this section is reprinted with permission from "Testing Contamination Source Identification Methods for Water Distribution Networks" by Seth, A., Klise, K.A., Siirola, J.D., Haxton, T., and Laird, C.D., 2015. to appear in Journal of Water Resources Planning and Management, Copyright 2015 by American Society of Civil Engineers.

sponding measure of likeliness, the CSA was modified to assign a likeliness measure of 1 to a node if it is contained in the list of unsafe node-time pairs, while all other nodes are assigned a likeliness measure of 0. Essentially, CSA operates by iterating over all measurements and pruning out (marking as safe) upstream node-time pairs that are hydraulically connected to negative measurements. Consequently, CSA allows for multiple simultaneous injections, however, it assumes perfect measurements when marking candidate injections as safe.

8.1.2 Optimization-Based Method

The third method used for this comparative study is the MILP based technique from Mann et al. (2012). This method incorporates the linear input-output water quality model Laird et al. (2005); Mann et al. (2012a) directly into an optimization formulation that seeks to find an injection source profile that minimizes the mismatch between yes/no measurements and those in the model. This formulation assumes that a sensor yields a positive measurement if the contaminant concentration is above a specified detection threshold concentration and a negative measurement otherwise. Therefore, if a sensor yields a positive measurement, any corresponding calculated concentration from the water quality model above the threshold is in agreement with this measurement data. Hence, while constructing an objective for estimation, only calculated concentrations below this threshold are penalized. Likewise, if a sensor yields a negative measurement, only the corresponding calculated concentration above the threshold is penalized. To identify a number of possible source candidates, the method repeatedly solves the MILP problem, each time adding integer cuts to remove previously found solutions until the objective value at the solution has deteriorated significantly. Note that for each candidate solution, the corresponding inverse of the objective value is used to represent the measure of likeliness of all nodes identified in that solution. Also note that this method allows for multiple simultaneous injections.

8.2 Performance results and sensitivity analysis of three methods

In this section, the performance of the three source identification methods is compared on the test cases proposed in Chapter 7. Each of the following subsections provides the performance plots for the three methods using a particular test set.

8.2.1 Preliminary Tests

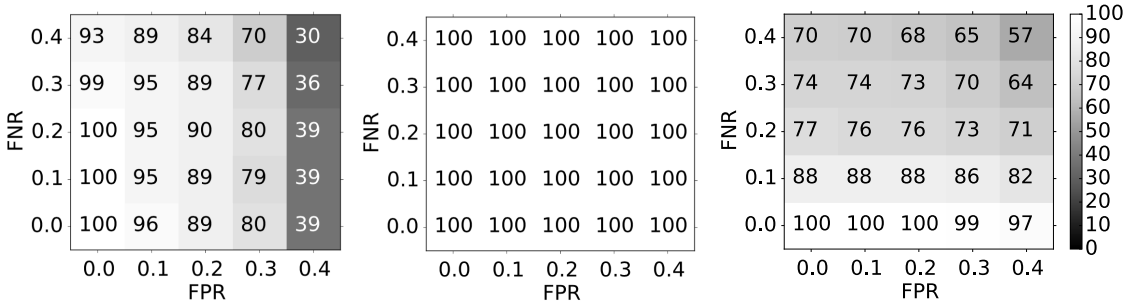
All the three source identification methods produced the expected results for the two preliminary tests.

8.2.2 Measurement Error

An increase in the false positive and false negative rate is expected to degrade the performance of all three source identification methods. Figures 8.1 and 8.2 highlight how the three source identification methods behave differently in the presence of false positives versus false negatives.

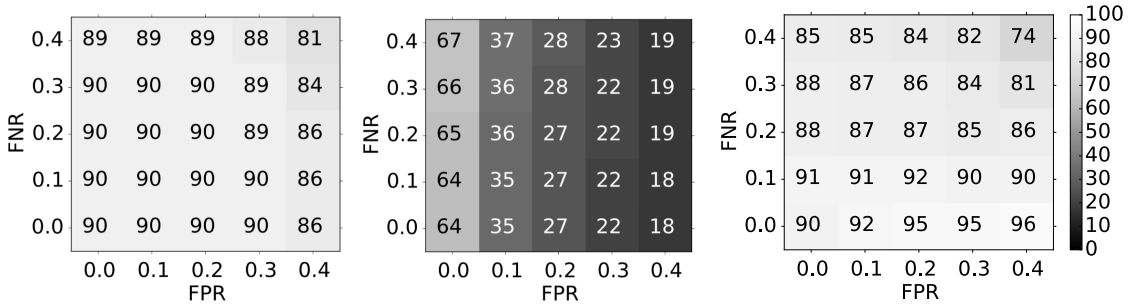
Figures 8.1a and 8.2a show that the Bayesian probability-based method performs worse in the presence of false positive measurements as opposed to false negative measurements (both in terms of accuracy and specificity). This behavior can be attributed to the fact that the probability-based method starts off by selecting a list of initial candidate injections that contains upstream node-time pairs that are hydraulically connected to positive sensor measurements. These candidate injections are then simulated to obtain measurement profiles, which are then matched against the true measurements. Therefore, a higher number of false positives leads to more candidate injections and also increases the possibility of finding injections that match the measurement data better than the true injection, hence degrading both accuracy and specificity.

In Figure 8.2b, the CSA shows a dramatic decrease in specificity with increased false positive rate. This is because the CSA selects, as injection location candidates, all node-time pairs that are hydraulically connected to any positive measurements.



(a) Probability-based (b) CSA (c) Optimization-based

Figure 8.1.: Mean accuracy of the three source identification methods as a function of FPR and FNR calculated over 5 injection nodes and 50 samples.



(a) Probability-based (b) CSA (c) Optimization-based

Figure 8.2.: Mean specificity of the three source identification methods as a function of FPR and FNR calculated over 5 injection nodes and 50 samples.

Therefore, an increase in the number of positive measurements leads to an increase in the size of the candidate set. Figure 8.1b also shows that the CSA maintains 100% accuracy across all levels of negative and positive measurement error. The CSA removes node-time pairs from the candidate set (marks node-time as safe) only when a negative measurement confirms the absence of contamination. Within the framework of this study, the CSA was modified to aggregate over time, including a candidate node if it appears in at least one node-time pair. Therefore, while the CSA

will remove a node-time pair in the face of a false negative measurement, as long as the true node is hydraulically connected to at least one positive measurement, the accuracy will be 100%.

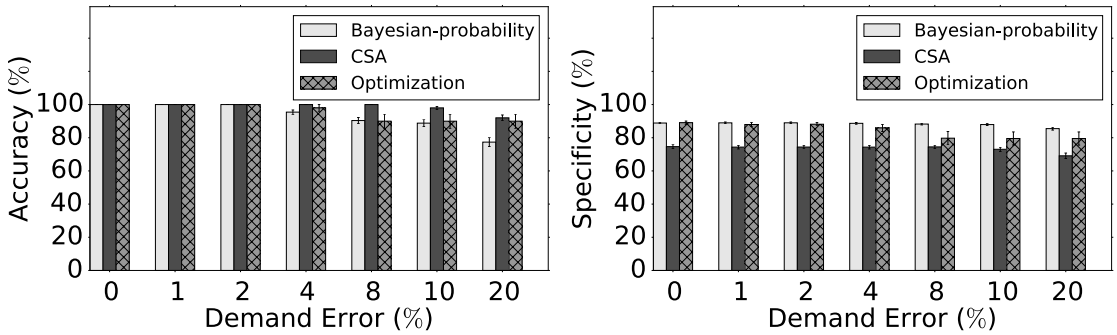
The optimization-based method is generally more balanced in the trade-off between accuracy and specificity even in the presence of large amounts of measurement error. A trend to notice in Figure 8.1c is that an increase in the FNR has a higher impact on the accuracy of this method compared with a similar increase in the FPR. This is likely due to the fact that a typical test scenario has a fewer number of positive measurements (often from a single location of initial detection) compared with the number of negative measurements taken from all sensors over the complete 12 hour time horizon.

8.2.3 Modeling Error

Figure 8.3 shows a decrease in performance for all three methods as the amount of demand error is increased. For instance, the mean specificity of the optimization-based method at 0% demand error is 90%. This means that considering the average performance over the 5 simulated scenarios, 10 nodes will have to be investigated before finding the true injection node. However, when the demand error is increased to 10%, the specificity reduces to 80%, which means that 20 nodes will have to be investigated. Therefore, for the optimization-based method the percentage of candidate nodes to be investigated doubles with only a 10% increase in the demand error. Similar conclusions can be made about the other two source identification methods. This highlights the need for accurate demand estimation when performing source identification.

8.2.4 Injection Characteristics

In the first set of tests, multiple simultaneous injections are simulated and the performance of the three methods is measured based on their ability to identify each



(a) Accuracy (b) Specificity

Figure 8.3.: Mean accuracy and specificity of the three source identification methods as a function of demand error calculated over 5 injection nodes and 50 samples. The error bars represent ± standard deviation of the mean.

individual injection node. Tests are run with a single injection, two simultaneous injections, and three simultaneous injections.

The Bayesian probability-based method assumes only a single injection node when performing source identification. This method uses a binomial distribution to calculate the prior source probabilities. For a large number of measurements, the binomial distribution has a sharp peak, which means that small changes in the number of matching measurements can lead to big differences in the probability values. This results in a big drop in probability values over the ranked list of candidate nodes. This means that even though the true injection node(s) can be high in rank (high specificity), it can still have low accuracy. Hence, while the test results show that this method has 100% accuracy for a single injection location, for two simultaneous injections the method can only identify one true injection node with 100% accuracy, while the other injection node has 1% accuracy and 80% specificity. For three simultaneous injections, the accuracy for all injection nodes drop to 1% while the specificities are 90%, 70%, and 60%. In contrast, the CSA allows for multiple injections and has 100% accuracy going from one to three simultaneous injections. However, the specificity deteriorates quickly from being 90% for one injection, to 75% for two injections, to

30% for three injections. The optimization-based method performs reasonably on these tests even though the maximum number of injections were set to one in the optimization formulation. The accuracy value(s) for one injection is 100%, for two simultaneous injections is 100% and 60%, and for three simultaneous injections is 100%, 70%, and 40%. The specificity values for the optimization-based method are very similar to the Bayesian probability-based method. These go from 90% for a single injection, to 90% and 60% for two simultaneous injections, to 95%, 80%, and 78% for three simultaneous injections.

Tables 8.1, 8.2, and 8.3 show results from a pairwise sensitivity study analyzing the effect of multiple simultaneous injections ranging from 1 to 3 injection nodes along with two levels of hydraulic uncertainty (demand error). As expected, low demand error (5%) did not have much impact on the accuracy and specificity of the three SI methods, while high demand error (20%) had more impact in reducing the performance of all three methods especially in the case of three simultaneous injections.

Table 8.1: Performance of the Bayesian-probability based method in the presence of multiple simultaneous injections and low (5%) and high (20%) demand error.

Demand	1 Injection		2 Injection		3 Injection	
Error (%)	Acc. (%)	Spe. (%)	Acc. (%)	Spe. (%)	Acc. (%)	Spe. (%)
0	100	90	100, 1	95, 80	1, 1, 1	90, 70, 60
5	100	88	67, 1	93, 79	1, 1, 1	86, 70, 60
20	100	87	66, 1	93, 77	1, 1, 1	86, 66, 40

Table 8.2: Performance of the CSA in the presence of multiple simultaneous injections and low (5%) and high (20%) demand error.

Demand	1 Injection		2 Injection		3 Injection	
Error (%)	Acc. (%)	Spe. (%)	Acc. (%)	Spe. (%)	Acc. (%)	Spe. (%)
0	100	90	100, 100	75, 75	100, 100, 100	30, 30, 30
5	100	87	100, 100	72, 72	100, 100, 100	22, 22, 22
20	100	86	100, 100	66, 66	100, 100, 100	18, 18, 18

Table 8.3: Performance of the optimization based method in the presence of multiple simultaneous injections and low (5%) and high (20%) demand error.

Demand	1 Injection		2 Injection		3 Injection	
Error (%)	Acc. (%)	Spe. (%)	Acc. (%)	Spe. (%)	Acc. (%)	Spe. (%)
0	100	90	100, 60	95, 80	100, 70, 40	95, 80, 78
5	100	88	100, 60	95, 79	100, 69, 38	95, 78, 76
20	100	87	100, 60	95, 77	97, 69, 20	95, 76, 42

Source identification techniques often make assumptions about the length of candidate injections. The probability-based method assumes continuous injections and therefore performs poorly for tests that involve short injection durations. For injection lengths below 4 hours, both mean accuracy and specificity values are under 50%. As expected, both metrics have high mean values (∼100%) when the injection length is over 8 hours.

On the other hand, CSA does not make any assumption regarding the injection length and is therefore more capable of identifying short injections. The method shows 100% mean accuracy and close to 60% mean specificity for all injection lengths. The

optimization-based method also shows 100% mean accuracy for all injection lengths, while the mean specificity goes from 60% for 1 hour injections to 80% for 12 hour injections. The optimization-based method does assume continuous injections, but it is still able to accurately identify short injections since an injection at the true node matches the measurement better than any other injection node, even if it is a poor match (many negative measurements do not match). However, short injections are more difficult to identify in the presence of measurement error. Therefore, as expected, adding 10% measurement error (FNR=10%, FPR=10%) leads to an extra 20% reduction in mean accuracy for both optimization-based method and Bayesian probability-based method on injection lengths less than 4 hours, while the CSA still shows 100% mean accuracy on these tests. Adding measurement error results in similar trends in mean specificity for all three methods over all injection lengths.

8.2.5 Time Horizon

As expected, the performance of all three methods improves as the time horizon is increased from 1 to 24 hours. For some cases, the true incident time is outside the time horizon. In those cases, the source identification method can identify the true source node, but with an incorrect incident time. Since the node-time pairs were aggregated to only identify nodes in the metrics, in some cases, the correct node will be identified. Nevertheless, on average, small time horizons are expected to result in poor performance. The mean accuracy of Bayesian probability-based method ranges from 1% for 1 hour horizon to 100% for 24 hour horizon, while the mean specificity ranges from 35% to 90%. The mean accuracy for both CSA and optimization-based method ranges from 20% to 100%, while the mean specificity for CSA has a slightly lower range (15% to 80%) as compared to the optimization-based method (40% to 100%).

Tables 8.4, 8.5, and 8.6 show results from a pairwise sensitivity study analyzing the effect of time horizon ranging from 1 to 24 hours and two levels of hydraulic uncertainty (demand error). As expected, low demand error (5%) did not have much

impact on the accuracy and specificity of the three SI methods, while high demand error (20%) had more impact in reducing the performance of all three methods. The trends are consistent with the Time Horizon results reported above.

Table 8.4: Impact of time horizon along with low (5%) and high (20%) demand error on the performance of the Bayesian-probability based method.

Time	0% Demand Error		5% Demand Error		20% Demand Error	
Horizon (%)	Acc. (%)	Spe. (%)	Acc. (%)	Spe. (%)	Acc. (%)	Spe. (%)
1	1	35	1	35	1	35
2	1	52	1	52	1	52
4	44	72	43	72	43	67
8	100	90	94	88	88	87
16	100	90	94	88	88	87
24	100	90	94	88	88	87

Table 8.5: Impact of time horizon along with low (5%) and high (20%) demand error on the performance of the CSA.

Time	0% Demand Error		5% Demand Error		20% Demand Error	
Horizon (%)	Acc. (%)	Spe. (%)	Acc. (%)	Spe. (%)	Acc. (%)	Spe. (%)
1	20	15	20	15	20	15
2	40	30	40	34	40	30
4	80	64	80	64	80	63
8	100	75	100	75	80	64
16	100	78	100	78	80	66
24	100	80	100	80	80	71

Table 8.6: Impact of time horizon along with low (5%) and high (20%) demand error on the performance of the optimization based method.

Time	0% Demand Error		5% Demand Error		20% Demand Error	
Horizon (%)	Acc. (%)	Spe. (%)	Acc. (%)	Spe. (%)	Acc. (%)	Spe. (%)
1	20	40	20	40	20	40
2	28	54	28	54	28	54
4	72	72	70	71	68	66
8	100	90	98	87	92	82
16	100	90	98	87	92	82
24	100	90	98	87	92	82

8.2.6 Network Size

Figure 8.4 shows an increase in the specificity values as the size of the network increases, however this is difficult to compare since the size of the networks differ. Therefore, the figure also includes the numeric values above each specificity bar to indicate the mean number of nodes that need to be investigated before the true injection node is identified. For instance, using CSA on the BWSN2 network on average involves investigating 214 nodes before the true node is identified. On the other hand, for the same network the Bayesian probability-based method requires only 45 nodes to be investigated. This is primarily because the CSA allows for the possibility of multiple injections and also because it produces a relatively large list of all equally likely candidate incidents. All methods showed 100% accuracy on all tests.

8.2.7 Sensor Placement

As expected, for both optimal and random sensor placement, the specificity of all three source identification methods (as shown in Figure 8.5 and Figure 8.6) improves

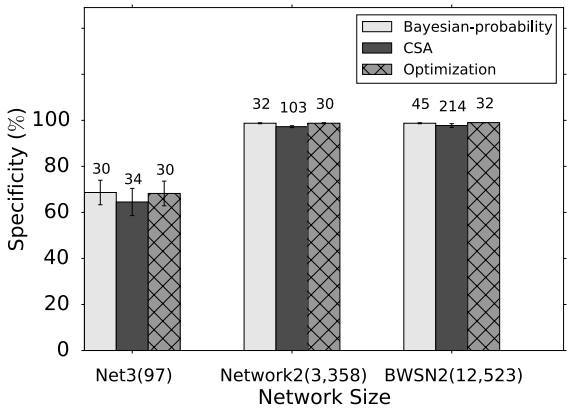


Figure 8.4.: The effect of network size on the performance of all three source identification methods. Each bar represents a mean specificity over the number of injection locations provided in Table 7.2. Error bars represent ± standard deviation of the mean. The number above each bar represents the absolute specificity value (i.e., the number of nodes with higher or equal likeliness to the true injection node).

with higher sensor density due to the increase in the amount of measurement information available from a larger number of locations around the network. All methods showed 100% accuracy on all test cases. It is interesting to see that the optimal sensor placement does not perform as well as the random sensor placement. This is due to the fact that optimal placement of sensors is typically done based on an objective (e.g., minimize population impact, maximum coverage) that is not designed for source identification. Typical optimal sensor placement results in sensors being placed at locations that detect larger number of scenarios, which can have a negative impact on a source identification method's ability to distinguish between possible injection scenarios. A pairwise sensitivity study that simultaneously considers hydraulic uncertainty is provided in the supplemental data (Figures S4 and S5).

Figures 8.7 and 8.8 show results from a pairwise sensitivity study considering sensor placement and hydraulic uncertainty. In general, low demand error (5%) did not have much impact on the performance of the three SI methods, while high demand

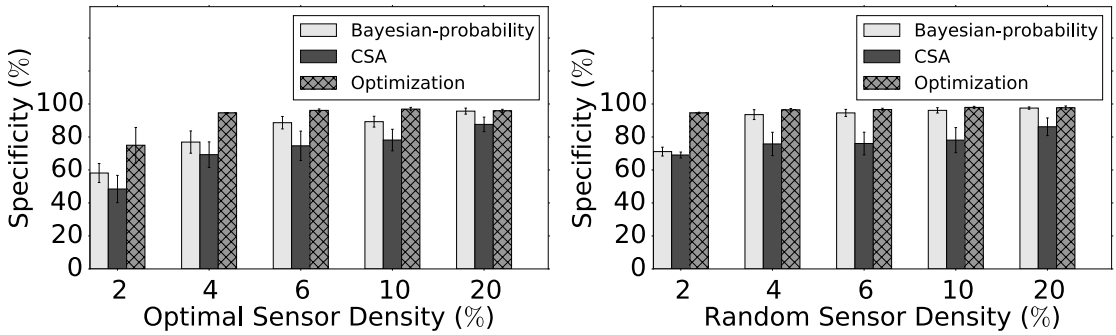


Figure 8.5.: The effect of sensor density and sensor placement on the specificity of all three source identification methods using Net3 (97 nodes). Each bar represents the mean specificity over 5 different injection locations. Error bars represent ± standard deviation of the mean.

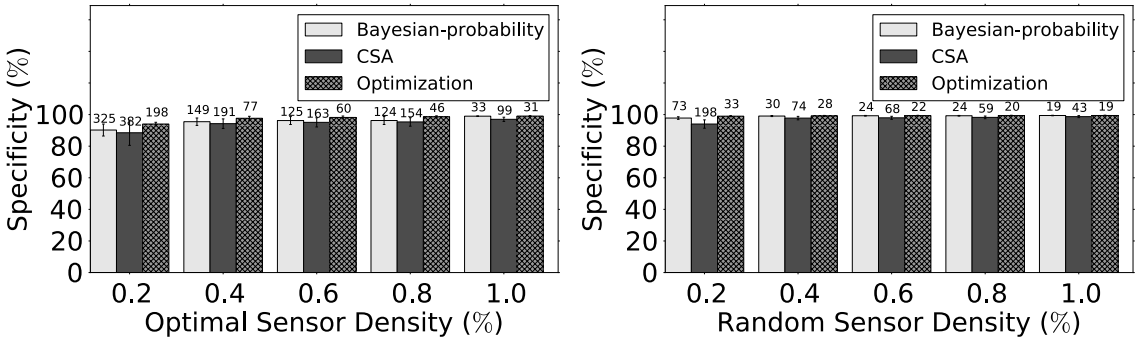
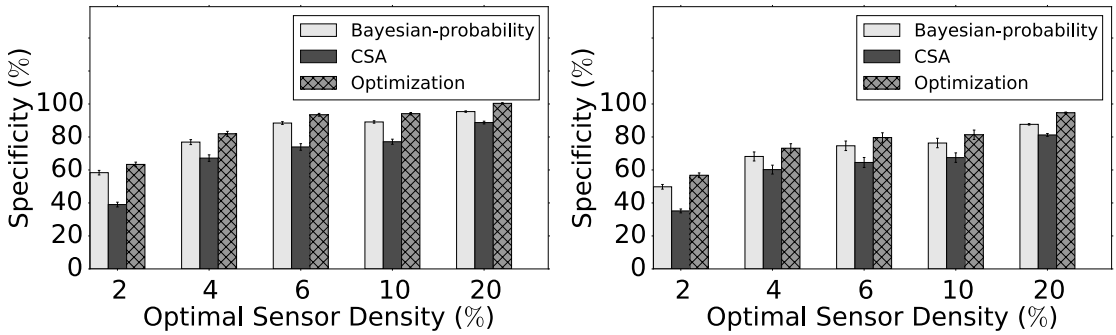


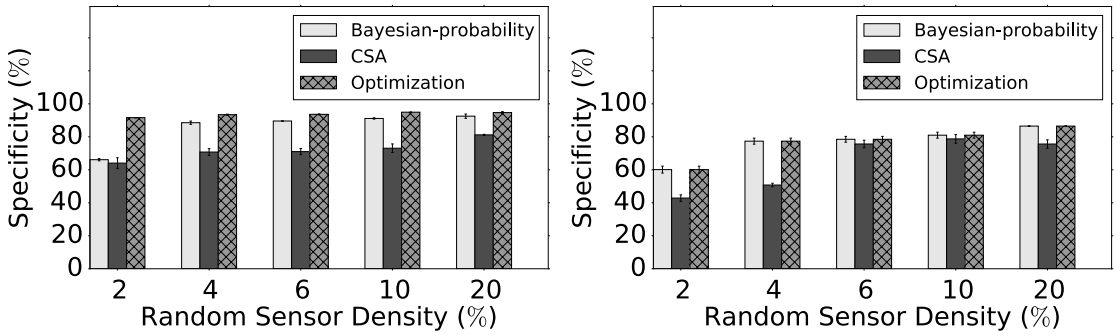
Figure 8.6.: The effect of sensor density and sensor placement on the specificity of all three source identification methods using Network2 (3,358 nodes). Each bar represents the mean specificity over 5 different injection locations. Error bars represent ± standard deviation of the mean. The number above each bar represents the absolute specificity value (i.e., the number of nodes with higher or equal likeliness to the true injection node).

error (20%) had more impact in reducing the performance of all three methods. The trends are consistent with the Sensor Placement results reported above.



(a) 5% Demand error (b) 20% Demand error

Figure 8.7.: The effect of sensor density of optimally placed sensors and modeling error on the specificity of all three SI methods using Net3 (97 nodes). Each bar represents the mean specificity over 5 different injection locations and 20 random samples of demand error. Error bars represent ± standard deviation of the mean.



(a) 5% Demand error (b) 20% Demand error

Figure 8.8.: The effect of sensor density of randomly placed sensors and modeling error on the specificity of all three SI methods using Net3 (97 nodes). Each bar represents the mean specificity over 5 different injection locations and 20 random samples of demand error. Error bars represent ± standard deviation of the mean.

8.3 Conclusions form comparative study

In general, the test cases presented in this work were effective at illustrating key differences in the methods and the following basic conclusions can be drawn. Note

that these results are not necessarily indicative of the performance of all methods of a particular class (i.e., all Bayesian probability-based methods, all optimization-based methods, or all CSA type methods).

- The Bayesian probability-based method assumes only single candidate injections and therefore performs poorly (at least in terms of accuracy) in the presence of multiple simultaneous injections. This method does not explicitly consider hydraulic connections between the sensor and candidate nodes (unconnected nodes-time pairs can match negative measurements). Furthermore, this method has poor accuracy in the presence of a large amount of false positive measurements. However, with reasonably good information (low measurement error, low demand error) this method shows higher accuracy and specificity in identifying single injections compared with the other two methods.

- The Contaminant Status Algorithm has higher accuracy than the other two methods, but typically shows lower specificity since it provides an exhaustive list of hydraulically connected node-time pairs with no negative measurement to mark them as safe. Unlike the other two methods, CSA does not make any assumptions about the length of candidate injections and therefore shows better performance in identifying short injection lengths. The specificity of this algorithm becomes worse as the number of positive measurements are increased, since more candidate injections are hydraulically connected to these measurements. Nevertheless, the fact that this method has good accuracy in the presence of large amount of measurement and modeling error can be used to shortlist the candidate set for further source identification calculations. More recent work by De Sanctis et al. (2008) extends this method to a Bayesian probabilistic approach.

- The optimization-based method shows good performance in most test cases, especially in the presence of large amount of measurement error. However, this method has tuning parameters (e.g., detection threshold) that could af-

fect performance in a real system, is more difficult to implement, and can be computationally intensive.

9. SUMMARY, CONCLUSIONS, AND FUTURE WORK [1]

Water distribution networks are vulnerable to inadvertent or intentional intrusion of chemical or biological species that can cause significant harm to a city population and the network infrastructure. Efficient design of detection, planning, and response systems can aid in minimizing the negative consequences of such incidents and speed up the mitigation process. In this thesis, we address two critical aspects of the response planning: (1) early mitigation of a potential contamination incident by injecting additional disinfectant into the network, and (2) identification of the source of the contamination stop contamination and begin cleanup operations. Systems modeling and optimization techniques provide a great tool for addressing these problems. Additionally, these techniques can be used to rigorously account for various uncertainties that need to be considered when planning for potential contamination incidents in the future (e.g., location of contaminant injection, time of contaminant injection, etc.). However, there are significant challenges associated with using systems modeling and optimization techniques for large and complex water distribution networks. As these networks become larger, the size of the models describing the flow of chemical or biological species also grows considerably. Multiplied by the fact that

[1]Part of this section is reprinted with permission from "Testing Contamination Source Identification Methods for Water Distribution Networks" by Seth, A., Klise, K.A., Siirola, J.D., Haxton, T., and Laird, C.D., 2015. to appear in Journal of Water Resources Planning and Management, Copyright 2015 by American Society of Civil Engineers.
Part of this section is reprinted from "Efficient Reduction of Optimal Disinfectant Booster Station Placement Formulations for Security of Large-Scale Water Distribution Networks" by Seth, A., Hackebeil, G.A., Klise, K.A., Haxton, T., Murray, R., and Laird, C.D., 2015. Submitted to Computational Optimization and Applications.
Part of this section is reprinted from "Evaluation of Chlorine Booster Station Placement for Water Security" by Seth, A., Hackebeil, G.A., Haxton, T., Murray, R., Laird, C.D., and Klise, K.A., 2015. Submitted to Journal of Water Resources Planning and Management, American Society of Civil Engineers.

large networks have a high number of potential contamination locations, considering this uncertainty in a optimization framework results in tremendously large and often intractable problems. In this thesis, we propose efficient optimization and modeling techniques that can tackle the two problems mentioned above for large-scale water distribution network.

In the first part of this thesis, we propose two optimization methods for the placement of disinfectant booster stations that can inject additional chlorine (but within safe limits) into the network as an early response to a potential contamination incident. When planning for potential contamination incidents, we do not have a priori knowledge of the contaminant species. Therefore, reasonable assumptions need to be made in order to model the contaminant-chlorine reaction. The two proposed methods provide two different ways of modeling the unknown reaction.

In Chapter 3, the first method for the optimal placement of disinfectant booster stations is proposed, which we call the "Neutralization method." The following contributions are made in this chapter:

- A model for the chlorine-contaminant reaction is proposed that makes two assumptions: (1) the reaction rate is assumed to be fast, and (2) the chlorine is assumed to be in stoichiometric excess, i.e., chlorine completely neutralizes a contaminant and remains in excess as it flows through the network. Additionally, we use a linear water quality model to describe the flow of the contaminant and chlorine in a network (Mann et al., 2012a). These assumptions allow us to decouple the linear chlorine and contaminant simulations and use superposition to calculate the chlorine and contaminant concentrations in the network for all the possible injection combinations (e.g., multiple booster injections, booster and contaminant injections). This modeling technique also lets us pre-simulate the booster and contaminant injections and use the resulting data in an optimization formulation, removing the need to embed the water quality model into the formulation.

- Two stochastic Mixed-Integer Linear Programming (MILP) formulations are proposed for placement of booster stations. These scenario-based formulations account for uncertainty in both the location and the time of a contamination incident. The objective in the first formulation is to minimize the expected mass consumed by the public in the form of demand from junctions (*MC formulation*). The objective of the second formulation is to minimize the expected number of people that ingest the contaminant above a dose threshold (*PD formulation*).

- Considering large-scale networks with potential contamination scenarios at every node and at every hour over a 24 hour demand cycle, the extensive form of the stochastic program is intractably large. However, a tremendous amount of structure in the problem is induced by the contamination scenario-based formulation and the network model itself. We propose three reduction techniques that dramatically decrease the size of the formulations. In the case studies considered in this chapter, the problem sizes were reduced as much as five orders of magnitude. With the proposed reductions the solution is possible considering realistic network models with more than 3,000 nodes.

- We analyze the effectiveness of booster stations in reducing the expected impact of contamination incidents. Case studies performed on three different networks highlight the significant benefits of using booster disinfection as an early response strategy.

In Chapter 4, the second method for the optimal placement of disinfectant booster stations is proposed, which we call the "Limiting reagent method." The following contributions are made in this chapter:

- We propose a stochastic MILP formulation that lets the user provide a stoichiometric ratio (Mass Chlorine/Mass Contaminant) as a parameter to approximate different kinds of contaminant-chlorine reactions. This is in contrast to the

Neutralization method, which assumes that chlorine remains in stoichiometric excess as it neutralizes the contaminant through the network. Therefore, the Limiting reagent method requires us to embed the linear water quality models describing the chlorine and contaminant flow into the optimization formulation.

- We provide a comparison of the Neutralization and Limiting reagent methods and show how these two methods can result in significantly different booster station placements on two different network models. In general, the Neutralization method gave an optimal placement that was closer to the upstream and downstream edges of the network, whereas the Limiting Reagent method resulted in booster stations being placed more centrally in the network as the stoichiometric ratio was increased.

- The effect of contaminant toxicity, sensor placement, and stoichiometric ratio was analyzed on the performance the two methods in terms of reduction in expected population dosed. Furthermore, each optimal booster station placement obtained using different levels of contaminant toxicity and stoichiometric ratios was evaluated over the same range of toxicities and stoichiometric ratios. The results show that under the assumption that the probability of contaminant toxicities and stoichiometric ratios are uniformly distributed, the optimal booster station placement obtained assuming the worst case scenario of high contaminant toxicity and high chlorine to contaminant stoichiometric ratio, resulted in the lowest overall expected population dosed.

In conclusion, the Neutralization method makes two simplifying assumptions about the contaminant-chlorine reaction that enable us to solve the optimal booster placement problem for large-scale water networks. The Limiting reagent method is more realistic and lets us model the contaminant-chlorine reaction with respect to a stoichiometric ratio, however, it can only tackle moderately size networks with limited number of scenarios. As a policy maker, one would need to quantify the probability of a range of possible contaminant species to make a more informed decision. Under

the assumption that the probability of different contaminants (with different reaction stoichiometric ratios) is uniformly distributed, our results indicate that the booster placement done using the Limiting reagent method will give the best overall performance. However, we know that for certain type of contaminants (e.g. E. Coli) the stoichiometric ratio for reaction to chlorine is very small and for those contaminants using the Neutralization method would make more sense. Keeping this in mind, the following future research directions are proposed:

- It should be straightforward to modify the scenario-based optimization formulation for the Limiting reagent method to account for the stochasticity in the stoichiometric ratio. Additionally, the uncertainty in the contaminant toxicity (dose threshold) can also be included in the formulation. As previously mentioned, quantifying the probability of possible contaminant species can also be explore in the future.

- Tools like EPANET-MSX (Shang et al., 2011) that enable modeling complex reactions between multiple chemical and biological species can also be used in the future to evaluate the effectiveness of the booster chlorination in the presence of more complex reaction kinetics.

- Scenario reduction schemes can be explored in the future to make the Limiting reagent method more tractable for larger networks.

The optimal booster station placement problem is solved at the planning stage before a contamination incident has taken place. Once a contamination has been confirmed, identifying its source location as quickly as possible can help in stopping further contamination. Additionally, response and cleanup operations can greatly benefit from an accurate understanding of the contaminant plume, which in turn requires knowledge of the contamination source location.

In the second part of this thesis, we address the problem of source identification that needs to be solved in real-time as a contamination incident unfolds. For this problem, the following contributions are made:

- In Chapter 6, we propose a Bayesian probability based method that uses sensor data to assign a source probability to all upstream nodes from the sensor locations that flag a contamination incident. This method benefits from a fast water quality simulation framework, Merlion (Mann et al., 2012a), to efficiently simulate a large number of possible contamination scenario. A simulation case study performed on a large-scale (above 12,000 nodes) network with more than 100 sensor locations, highlights the computational speed and accuracy of the proposed method, which performs the source identification calculations within seconds.

- Due to a limited number of measurements obtained from fixed water quality sensors, the source identification problem is ill-posed, which can result in many possible source locations. Wong et al. (2010) proposed an optimization formulation for selection of manual grab sampling locations to get additional measurements that can help distinguish between potential source locations. As a corollary to the optimization method proposed by Wong et al. (2010), in Chapter 6 we propose a greedy algorithm that is shown to be computationally efficient for large-scale networks with similar effectiveness.

- Due to the wide range of source identification methods proposed in the literature, there is a need for a testing framework to contrast and compare different methods. In Chapter 7 a systematic testing methodology for contamination source identification methods is proposed. This methodology includes performance metrics and a set of test cases designed to analyze a variety of factors that can influence the performance of source identification methods (e.g., measurement error, modeling error, sensor placement, network size, etc.).

- In Chapter 8, the proposed testing methodology is used to perform a comparative study of the Bayesian probability-based method and two source identification methods from the literature. The study highlights the strengths and weaknesses of each method.

Decision makers must be aware of the underlying assumptions that a source identification method makes. The testing methodology proposed in Chapter 7 is designed to identify common issues that may arise due to these assumptions. To further extend this testing methodology the following future work is proposed:

- Random noise on individual measurement points was added to show that the performance of source identification methods can start to degrade at high false negative and false positive rates. In the future, it will be interesting to study the impact of more systematic sensor failures.

- While higher sensor density generally leads to better performance of source identification methods, typical criteria used for optimal sensor placement are not ideal for source identification. Identifying sensor placements to improve source identification performance is an interesting topic for future study. Additionally, multi-objective approaches that consider both minimizing the impact of contamination incidents and improving source identification should be explored in the future.

- The source identification methods studied in this work do not make use of the specific identity of the contaminant (e.g., to model decay/reactions). In general, the source identification methods should be effective immediately, before the contaminant may be identified through laboratory analysis. If the specific compound is known, then the water quality models could be modified to include kinetic models. This is a reasonable direction for future work.

- With the simultaneous development of real-time data collection systems, real-time modeling tools, and real-time source identification tools, there is a need to study and optimize their interactions, which opens up new challenges associated with monitoring and protecting drinking water networks.

LIST OF REFERENCES

LIST OF REFERENCES

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

Berry, J., W. E. Hart, C. A. Phillips, J. G. Uber, and J.-P. Watson (2006). Sensor placement in municipal water networks with temporal integer programming models. *Journal of Water Resources Planning and Management 132*(4), 218–224.

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

Boccelli, D., M. Tryby, J. Uber, and L. Rossman (1998). Optimal location of booster disinfection stations for residual maintenance. In *Water Resources and the Urban Environment*, pp. 266–271. ASCE.

Boccelli, D., M. Tryby, J. Uber, L. Rossman, M. Zierlof, and M. Polycarpuo (1998). Optimal scheduling of booster disinfection in water distribution systems. *Journal of Water Resources Planning and Management 124*(2), 99–111.

Boccelli, D. L., M. E. Tryby, J. G. Uber, L. A. Rossman, M. L. Zierolf, and M. M. Polycarpou (1998). Optimal scheduling of booster disinfection in water distribution systems. *Journal of Water Resources Planning and Management 124*(2), 99–111.

Brumbelow, K., J. Torres, S. Guikema, E. Bristow, and L. Kanta (2007). Virtual cities for water distribution and infrastructure system research. In *World Environmental and Water Resources Congress*, pp. 15–19.

Constans, S., B. Brémond, and P. Morel (2003). Simulation and control of chlorine levels in water distribution networks. *Journal of water resources planning and management*.

Cristo, C. and A. Leopardi (2008). Pollution source identification of accidental contamination in water distribution networks. *Journal of Water Resources Planning and Management 134*(2), 197–202.

Cunha, M. d. C. and J. Sousa (1999). Water distribution network design optimization: simulated annealing approach. *Journal of Water Resources Planning and Management 125*(4), 215–221.

Davis, M. J., R. Janke, and M. L. Magnuson (2014). A framework for estimating the adverse health effects of contamination events in water distribution systems and its application. *Risk Analysis 34*(3), 498–513.

De Sanctis, A., F. Shang, and J. Uber (2009). Real-time identification of possible contamination sources using network backtracking methods. *Journal of Water Resources Planning and Management 136*(4), 444–453.

De Sanctis, A. E., D. L. Boccelli, F. Shang, and J. G. Uber (2008, May). Probabilistic approach to characterize contamination sources with imperfect sensors. In *World Environmental and Water Resources Congress 2008*, Reston, VA, pp. 1–10. American Society of Civil Engineers.

Eliades, D. and M. Polycarpou (2011). Water contamination impact evaluation and source-area isolation using decision trees. *Journal of Water Resources Planning and Management 138*(5), 562–570.

Ellison, D. (2003). *Investigation of pipe cleaning methods*. American Water Works Association.

EPA, U. S. (2010a). Water quality event detection systems for drinking water contamination warning systems: Development, testing, and application of CANARY. Technical Report EPA/600/R-10/036, U.S. Environmental Protection Agency, Office of Research and Development, Washington, DC.

EPA, U. S. (2014). Water security toolkit user manual: Version 1.1. Technical Report EPA/600/R-13/353, U.S. Environmental Protection Agency, Office of Research and Development, Washington, DC.

EPA, U.S. (2010b). *Technology and cost document for the proposed revised total coliform rule*. Office of Water.

Eusuff, M. M. and K. E. Lansey (2003). Optimization of water distribution network design using the shuffled frog leaping algorithm. *Journal of Water Resources Planning and Management 129*(3), 210–225.

Geem, Z. W. (2009). Particle-swarm harmony search for water network design. *Engineering Optimization 41*(4), 297–311.

Guan, J., M. Aral, M. Maslia, and W. Grayman (2006). Identification of contaminant sources in water distribution systems using simulation-optimization method: Case study. *Journal of Water Resources Planning and Management 132*(4), 252–262.

Hackebeil, G., A. Mann, W. Hart, K. Klise, and C. Laird (2012). A stochastic programming formulation for chlorine booster station placement to protect large-scale water distribution systems. *Computer Aided Chemical Engineering 31*, 1462–1466.

Hochbaum, D. S. (1996). *Approximation algorithms for NP-hard problems*. PWS Publishing Company.

Huang, J. and E. McBean (2009). Data mining to identify contaminant event locations in water distribution systems. *Journal of Water Resources Planning and Management 135*(6), 466–474.

Isovitsch, S. L. and J. M. VanBriesen (2007). Integrating scada and gis to understand the effectiveness of on-line chlorine boosters used in response to contamination incidents within a water distribution network. *Proceedings of the Water Environment Federation 2007*(1), 424–433.

Jowitt, P. W. and G. Germanopoulos (1992). Optimal pump scheduling in water-supply networks. *Journal of Water Resources Planning and Management 118*(4), 406–422.

Kang, D. and K. Lansey (2010). Real-time optimal valve operation and booster disinfection for water quality in water distribution systems. *Journal of Water Resources Planning and Management 136*(4), 463–473.

Laird, C., L. T. Biegler, B. G. v. B. Waanders, and R. A. Bartlett (2005). Contamination source determination for water networks. *Journal of Water Resources Planning and Management 131*(2), 125–134.

Laird, C. D., L. T. Biegler, and B. G. van Bloemen Waanders (2006). Mixed-integer approach for obtaining unique solutions in source inversion of water networks. *Journal of Water Resources Planning and Management 132*(4), 242–251.

Laird, C. D., L. T. Biegler, B. G. van Bloemen Waanders, and R. A. Bartlett (2005). Contamination source determination for water networks. *Journal of Water Resources Planning and Management*.

Lansey, K., F. Pasha, S. Pool, W. Elshorbagy, and J. Uber (2007). Locating satellite booster disinfectant stations. *Journal of Water Resources Planning and Management 133*(4), 372–376.

Liu, L., S. R. Ranjithan, and G. Mahinthakumar (2011). Contamination source identification in water distribution systems using an adaptive dynamic optimization procedure. *Journal of Water Resources Planning and Management 137*(2), 183–192.

Mackle, G., D. Savic, G. A. Walters, et al. (1995). Application of genetic algorithms to pump scheduling for water supply. In *Genetic Algorithms in Engineering Systems: Innovations and Applications, 1995. GALESIA. First International Conference on (Conf. Publ. No. 414)*, pp. 400–405. IET.

Mann, A. V., G. a. Hackebeil, and C. D. Laird (2012a, August). Explicit water quality model generation and rapid multi-scenario simulation. *Journal of Water Resources Planning and Management 140*(5), 666–677.

Mann, A. V., G. A. Hackebeil, and C. D. Laird (2012b). Explicit water quality model generation and rapid multiscenario simulation. *Journal of Water Resources Planning and Management 140*(5), 666–677.

Mann, A. V., S. A. McKenna, W. E. Hart, and C. D. Laird (2010). Real-time inversion and response planning in large-scale networks. *Computer Aided Chemical Engineering 28*, 1027–1032.

Mann, A. V., S. a. McKenna, W. E. Hart, and C. D. Laird (2012, February). Real-time inversion in large-scale water networks using discrete measurements. *Computers & Chemical Engineering 37*, 143–151.

Munavalli, G. and M. M. Kumar (2003). Optimal scheduling of multiple chlorine sources in water distribution systems. *Journal of water resources planning and management 129*(6), 493–504.

Murray, R., T. Haxton, R. Janke, W. E. Hart, J. Berry, and C. Phillips (2010a). Sensor network design for drinking water contamination warning systems: A compendium of research results and case studies using the teva-spot software. Technical Report EPA/600/R-09/141, Cincinnati, OH Office of Research and Development, National Homeland Security Research Center, Water Infrastructure Protection.

Murray, R., T. M. Haxton, R. J. Janke, W. E. Hart, J. Berry, and C. A. Phillips (2010b). Sensor network design for drinking water contamination warning systems: A compendium of research results and case studies using the TEVA-SPOT-Report. US Environmental Protection Agency, Washington, DC. Technical report, EPA/600/R-09/141.

OECD (2012). *OECD Environmental Outlook to 2050: The Consequences of Inaction*. OECD Publishing, Paris.

Oliker, N. and A. Ostfeld (2014). Minimum volume ellipsoid classification model for contamination event detection in water distribution systems. *Environmental Modelling & Software 57*, 1–12.

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

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

Ostfeld, A. and E. Salomons (2006). Conjunctive optimal scheduling of pumping and booster chlorine injections in water distribution systems. *Engineering Optimization 38*(3), 337–352.

Ostfeld, A., J. Uber, E. Salomons, J. W. Berry, W. E. Hart, C. A. Phillips, J.-P. Watson, G. Dorini, P. Jonkergouw, Z. Kapelan, F. di Pierro, S.-T. Khu, D. Savic, D. Eliades, M. Polycarpou, S. R. Ghimire, B. D. Barkdoll, R. Gueli, J. J. Huang, E. A. McBean, W. James, A. Krause, J. Leskovec, S. Isovitsch, J. Xu, C. Guestrin, J. VanBriesen, M. Small, P. Fischbeck, A. Preis, M. Propato, O. Piller, G. B. Trachtman, Z. Y. Wu, and T. Walski (2008). The battle of the water sensor networks (BWSN): A design challenge for engineers and algorithms. *Journal of Water Resources Planning and Management 134*(6), 556–568.

Ozdemir, O. and M. Ucaner (2005). Success of booster chlorination for water supply networks with genetic algorithms. *Journal of Hydraulic Research 43*(3), 267–275.

Parks, S. and J. VanBriesen (2009). Booster disinfection for response to contamination in a drinking water distribution system. *Journal of Water Resources Planning and Management 135*(6), 502–511.

Perelman, L. and A. Ostfeld (2012). Bayesian networks for source intrusion detection. *Journal of Water Resources Planning and Management 139*(4), 426–432.

Prasad, T. D., G. Walters, and D. Savic (2004). Booster disinfection of water supply networks: Multiobjective approach. *Journal of Water Resources Planning and Management 130*(5), 367–376.

Preis, A. and A. Ostfeld (2006). Contamination source identification in water systems: A hybrid model treeslinear programming scheme. *Journal of Water Resources Planning and Management 132*(4), 263–273.

Preis, A. and A. Ostfeld (2007). A contamination source identification model for water distribution system security. *Engineering Optimization 39*(8), 37–41.

Preis, A. and A. Ostfeld (2008, March). Genetic algorithm for contaminant source characterization using imperfect sensors. *Civil Engineering and Enviornmental Systems 25*(1), 37–41.

Propato, M., F. Sarrazy, and M. Tryby (2009). Linear algebra and minimum relative entropy to investigate contamination events in drinking water systems. *Journal of Water Resources Planning and Management 136*(4), 483–492.

Propato, M. and J. Uber (2004a). Booster system design using mixed-integer quadratic programming. *Journal of Water Resources Planning and Management 130*(4), 348–352.

Propato, M. and J. Uber (2004b). Linear least-squares formulation for operation of booster disinfection systems. *Journal of Water Resources Planning and Management 130*(1), 53–62.

Rossman, L. (2000). EPANET 2:Users Manual. Cincinnati, OH: U.S. Environmental Protection Agency.

Rossman, L. A. and P. F. Boulos (1996). Numerical methods for modeling water quality in distribution systems: A comparison. *Journal of Water Resources planning and management 122*(2), 137–146.

Shang, F., J. Uber, and L. Rossman (2011). Epanet mutli-species extension user's manual. Technical Report EPA/600/S-07/021, USEPA.

Shang, F., J. G. Uber, and M. M. Polycarpou (2002a). Particle backtracking algorithm for water distribution system analysis. *Journal of Environmental Engineering 128*(5), 441–450.

Shang, F., J. G. Uber, and M. M. Polycarpou (2002b). Particle backtracking algorithm for water distribution system analysis. *Journal of Environmental Engineering 128*(5), 441–450.

Shen, H. and E. McBean (2011). False negative/positive issues in contaminant source identification for water-distribution systems. *Journal of Water Resources Planning and Management 138*(3), 230–236.

Tryby, M., D. Boccelli, J. Uber, and L. Rossman (2002). Facility location model for booster disinfection of water supply networks. *Journal of Water Resources Planning and Management 128*(5), 322–333.

Tryby, M., M. Propato, and S. Ranjithan (2010). Monitoring design for source identification in water distribution systems. *Journal of Water Resources Planning and Management 136*(6), 637–646.

Uber, J. G., D. Boccelli, R. Summers, and M. Tryby (2003). *Maintaining Distribution System Residuals Through Booster Chlorination*. Awwa Research Foundation.

Uber, J. G., M. M. Polycarpou, and P. Subramaniam (1998). Optimal decoupling of booster disinfection systems in water distribution networks. In *Water Resources and the Urban Environment*, pp. 297–302. ASCE.

Van Zyl, J. E., D. A. Savic, and G. A. Walters (2004). Operational optimization of water distribution systems using a hybrid genetic algorithm. *Journal of water resources planning and management 130*(2), 160–170.

Vasan, A. and S. P. Simonovic (2010). Optimization of water distribution network design using differential evolution. *Journal of Water Resources Planning and Management*.

Wagner, D. E. and R. M. Neupauer (2013). Probabilistic contaminant source identification in water distribution systems with incomplete mixing at pipe junctions. *World Environmental and Water Resources Congress 2013*, 930–935.

Wagner, D. E., R. M. Neupauer, and C. Cichowitz (2015). Adjoint-based probabilistic source characterization in water-distribution systems with transient flows and imperfect sensors. *Journal of Water Resources Planning and Management*, 04015003.

Wang, H. and K. W. Harrison (2012). Improving efficiency of the bayesian approach to water distribution contaminant source characterization with support vector regression. *Journal of Water Resources Planning and Management 140*(1), 3–11.

Watson, J.-P., R. Murray, and W. Hart (2009). Formulation and optimization of robust sensor placement problems for drinking water contamination warning systems. *Journal of Infrastructure Systems 15*(4), 330–339.

Wong, A., J. Young, W. E. Hart, S. A. McKenna, and C. D. Laird (2010). Optimal determination of grad sample locations and source inversion in large-scale water distribution systems. *Water Distribution System Analysis 2010*, 412–425.

Wong, A. V., S. A. McKenna, W. E. Hart, and C. D. Laird (2010). Real-time inversion and response planning in large-scale networks. *Computer Aided Chemical Engineering 28*, 1027–1032.

Yang, X. and D. L. Boccelli (2014). Bayesian approach for real-time probabilistic contamination source identification. *Journal of Water Resources Planning and Management 140*(8), 04014019.

Yu, G., R. Powell, and M. Sterling (1994). Optimized pump scheduling in water distribution systems. *Journal of optimization theory and applications 83*(3), 463–488.

Zecchin, A. C., H. R. Maier, A. R. Simpson, M. Leonard, and J. B. Nixon (2007). Ant colony optimization applied to water distribution system design: comparative study of five algorithms. *Journal of Water Resources Planning and Management 133*(1), 87–92.

Zechman, E. M. and S. R. Ranjithan (2009). Evolutionary computation-based methods for characterizing contaminant sources in a water distribution system. *Journal of Water Resources Planning and Management 135*(5), 334–343.

Zhao, H., D. Hou, P. Huang, and G. Zhang (2014). Water quality event detection in drinking water network. *Water, Air, & Soil Pollution 225*(11), 1–15.

VITA

VITA

Arpan Seth was born in Lucknow, Uttar Pradesh, India. After finishing his schooling at Seth M.R. Jaipuria School, Lucknow, he received his Bachelor of Science in Chemical Engineering from Louisiana State University, Baton Rouge. In August 2010, he started graduate school at Texas A&M University, College Station. Arpan joined the Carl Laird Research Group in January 2011 where he gained knowledge and experience in numerical modeling and optimization. He began conducting his research on real-time response techniques for water distributions networks. In January 2014, Arpan transferred to Purdue University, West Lafayette, to continue his research with his advisor Dr. Carl Laird.

During graduate school, Arpan also did several research internships in the industry - two at ExxonMobil Corporation and one at Rockwell Automation. These internships helped him gain valuable insight about the state of the oil and gas industry and the role modeling and optimization can play in their future.