I, SM Masud Rana, hereby submit this original work as part of the requirements for the degree of Master of Science in Environmental Engineering.

It is entitled:
Contaminant Spread Forecasting and Sampling Location Identification in a Water Distribution Network

Student’s name: SM Masud Rana

This work and its defense approved by:

Committee chair: Dominic Boccelli, Ph.D.
Committee member: Robert Janke, M.S.
Committee member: James Uber, Ph.D.
CONTAMINANT SPREAD FORECASTING AND SAMPLING LOCATION IDENTIFICATION IN A WATER DISTRIBUTION NETWORK

A thesis submitted to the Division of Research and Advanced Studies of the University of Cincinnati in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

in the Department of
Biomedical, Chemical and Environmental Engineering of the College of Engineering and Applied Science

2013

by

SM Masud Rana

B.Sc., Civil Engineering, 2011
Bangladesh University of Engineering and Technology (BUET)
Abstract

A safe drinking water distribution system is an indispensable requirement for a developed and healthy community as they rely principally on distributed water for their everyday needs. A contamination event in the distribution system can have severe impacts on the health of the unsuspecting consumers. The large spatial extent of a distribution system provides a big challenge to utility managers by making physical monitoring of water quality inefficient and expensive. Monitoring water quality with online sensors has provided a solution to this problem and is becoming a more common practice in the USA. However, monitoring for specific contaminants is not practical as such sensors will be useless for different contaminants and during normal operating conditions. Hence, monitoring non-specific water quality parameters (e.g., chlorine, pH, etc.) has gained more preference as these monitoring systems can play a dual role - provide non-specific water quality information during normal operations and act as sensors to detect contamination intrusion during possible contamination events.

A large amount of data can be obtained from online monitoring stations but the interpretation of these data presents a significant challenge. Event detection algorithms (EDA) have been developed as part of contamination warning systems (CWS) with significant success for distinguishing anomalous events from normal operating signals using non-specific water quality data. Additionally, source identification algorithms have been developed that can locate contamination sources after a CWS triggers an alarm. While these developments have been progressing, there remains a lack of tools and methods to interpret these results for the forecasting of contamination spread. An integrated framework is essential that would combine both source
identification and spread forecasting together to be useful for the utility managers to take precautionary and mitigating actions.

In this research, a spread forecasting algorithm has been developed that is based upon the results of a probabilistic contamination source identification (PCSI) algorithm, which estimates the probability that a node maybe contaminated. The forecasted spread of contamination probability to downstream nodes was estimated by assuming the impact of the upstream nodes to be proportional to their flow contributions to the downstream nodes. Additionally a confirmatory sampling location selection algorithm has also been developed that utilizes the forecasted spread to estimate sampling locations to improve the information about the overall water distribution network, which will lead to improved forecasting and source identification. Sampling location selection was based on the expected improvement of information as quantified by concepts of entropy from Information Theory. In Information Theory, lower entropy corresponds to higher certainty i.e., a greater amount of information. The developed sampling location selection algorithm determines the best sampling locations to be those that are expected to minimize the entropy of the entire distribution system when sampled at a particular time.

Two distribution systems with varying complexity were used to evaluate the performance of the developed algorithms. The first one was a small 97 node network, while the second one was a large 12,527 node network. For the first network, only one case with 5 sensors was analyzed while for the second network four cases with 5, 10, 20 and 50 sensors were analyzed.

The accuracy of the contamination spread algorithm was dependent on the amount of available past node-time (N-T) pair information. Spread forecast accuracy increased with increasing number of sensors due to the concomitant increase of past N-T pair information and generally decreased with increasing forecasting horizon. The estimated sampling locations were shown to maximize benefit in terms of the correct
identification of past N-T pairs by the PCSI algorithm. The locations estimated to
be the best sampling locations among all the potential sampling locations provided
the greatest increase in percent correct identification and greatest decrease in percent
incorrect identification and nodes with no information. Similarly, the spread forecast
was also improved by the estimated sampling locations, which increased the percent
of correct identification, and decreased the percent incorrect identification and nodes
with no forecast. Locations that provided more new information about unknown N-T
pairs were found to be better sampling locations than the locations that strengthened
existing information about known N-T pairs. Increasing the number of sensors had
diminishing effect on the impact of confirmatory sampling. Thus, the improvements
associated with the PCSI algorithm and the spread forecasting algorithm through
confirmatory samplings decreased with increasing number of sensors.
Acknowledgements

I would like to offer my special thanks to . . .

...my advisor, Dominic Boccelli, for providing me a wonderful opportunity to pursue my Master’s and for providing me with the encouragement to continue to my PhD. Your encouragement and sheer confidence in me has pushed me to believe that with proper diligence I can surmount great heights. I have learned valuable lessons both in and outside the classroom about how to approach a problem and conduct research all while have fun doing it. The several opportunities you provided to present my work, at meetings and conferences, helped me to mold my work into better shape every time. You have unwittingly provided me with an immense inspiration about life and have guided my future career in a completely different path than what I have imagined before. For the reasons mentioned here and those that remain unstated, I am extremely grateful for the time that I have shared with you so far.

...James Uber, for his guidance and the helpful discussions on approaching problems of this research. His unique ways of dealing with research and problems inspired me to look at complex problems from surprisingly simple perspectives.

...my friend Alissa O’Donnell, whose help and valuable criticism meant that my writings were much better shaped than it would have been without it. I appreciate her countless hours spent on my writings and her helping hand in other life aspects which made my Masters life a lot easier.
...all my friends who made my graduate life, far away from home, enjoyable and provided me with the support that I couldn’t get from from my family.

...Lipi Aktar, a very special friend, who mentally pushed and shoved me during my down times to keep me going after my dream. Her valuable criticisms and endless encouragement ultimately helped me reach this successful completion of my Master’s.

...Finally, my parents and family back at Bangladesh, whose hard work and sacrifice allowed me to come here to pursue my dream. Even from thousands of miles away, they constantly provided me with all kinds of love and support that they could. I am eternally grateful and indebted to them.

These few lines of writing cannot fully express my feelings and it is just a token of my gratitude. From the day I landed up to this day, the University of Cincinnati has provided me with tremendous hospitality which has opened endless opportunities for me. The University of Cincinnati has supported my education, fortified me with wisdom and has invigorated me to continue for the pursuit of knowledge for the rest of my life.
Contents

1 Introduction 1

2 Methodology 5
   2.1 Spread Forecasting ................................. 5
   2.2 Sampling Location Selection ...................... 9

3 Results 13
   3.1 Small Network: Net3 .............................. 13
   3.2 Large Network: Network2 .......................... 23

4 Discussion 45

5 Conclusion 49

6 Future Work 53

7 References 55
List of Figures

2.1 Illustrative spread forecast example. 7
2.2 Entropy of an event space with two possible outcomes. 10
3.1 Layout of the small test network, Net3. 14
3.2 Layout of the large test network, Network2. 15
3.3 Spread forecasting results for Net3. 17
3.4 Expected Network Entropy and Actual Network Entropy for sampling during the 7th hour in Net3. 19
3.5 Cumulative relative frequency of ANE for sampling at all the potential locations in Net3. 20
3.6 Spatial locations of the top estimated sampling locations for the 7th hour in Net3. 20
3.7 Changes in the PCSI characterization of N-T pairs due to sampling during hour 7 in Net3. 22
3.8 Spread forecasting results for Network2. 25
3.9 Cumulative Relative Frequency plots of ANE for Network2. 27
3.10 Spatial locations of the estimated and actual best sampling locations in Network2. 29
3.11 Spatial locations of the top 1% of the estimated sampling locations in Network2 for the 5 sensor case. 30
3.12 Spatial locations of the top 1% of the estimated sampling locations in Network2 for the 10 sensor case. ... 31
3.13 Spatial locations of the top 1% of the estimated sampling locations in Network2 for the 20 sensor case. ... 32
3.14 Spatial locations of the top 1% of the estimated sampling locations in Network2 for the 50 sensor case. ... 33
3.15 Changes in the PCSI characterization of N-T pairs due to sampling in Network2 with 5 sensors. ... 35
3.16 Changes in the PCSI characterization of N-T pairs due to sampling in Network2 with 10 sensors. ... 36
3.17 Changes in the PCSI characterization of N-T pairs due to sampling in Network2 with 20 sensors. ... 37
3.18 Changes in the PCSI characterization of N-T pairs due to sampling in Network2 with 50 sensors. ... 38
3.19 Changes in the forecasted spread due to sampling in Network2 with 5 sensors. ... 39
3.20 Changes in the forecasted spread due to sampling in Network2 with 10 sensors. ... 40
3.21 Changes in the forecasted spread due to sampling in Network2 with 20 sensors. ... 41
3.22 Changes in the forecasted spread due to sampling in Network2 with 50 sensors. ... 42
List of Tables

3.1 Time to first detection of contamination for different sensor cases for Network2. .................................................. 24

3.2 Top five estimated sampling locations and their ANE rankings for Network2, for the different false rates. .................. 44
**Notation**

ANE    actual network entropy  
$En$    entropy  
ENE    expected network entropy  
$\mathbb{E}[NE]$ expected network entropy  
FNR    false negative rate  
FPR    false positive rate  
i    output node  
$I_{ij}$    impact of the input node $j$ on the output node $i$ at time $t$  
j    input node  
$NE$    network entropy  
$NE_{T+\Delta t}^{+ve}$ network entropy for an assumed positive signal at $T + \Delta t$ hour  
$NE_{+ve}$ network entropy for an assumed positive signal  
$NE_{-ve}$ network entropy for an assumed negative signal  
$N$    total number of nodes  
N-T    node-time pair  
p    probability of a node-time pair to be contaminated  
$p_i^t$    probability of node $i$ to be contaminated at time $t$  
PCSI    probabilistic contamination source identification algorithm  
t    time  
T    current analysis time
Chapter 1

Introduction

Water distribution networks are one of the most expensive resources of a municipality and yet they are vulnerable to (un)intentional contamination events due to the ease of access by the public and their large spatial distribution, which makes physical monitoring inefficient and difficult. The events of September 11, 2001 raised a massive concern among utility managers and water distribution network specialists about the security of drinking water in distribution networks. Drinking water quality is threatened by intentional contamination intrusion and also unintentional intrusions during, for example, accidental low pressure episodes that can lead to major public health hazards. Online monitoring of water distribution networks is considered one of the most efficient ways to manage and monitor water quality and detect intrusion events (Janke et al., 2006; USEPA, 2005a,b). Monitoring for specific contaminants is not only expensive but also of very little use, if any, when different contaminants may be present or during normal operating conditions. Thus, there has been considerable efforts in developing methods to detect contamination events in water distribution networks using sensors associated with common water quality metrics (e.g., chlorine, pH, conductivity, etc.) (Allgeier et al., 2006; Byer and Carlson, 2005; McKenna et al., 2008; Yang and Boccelli, 2012; Yang et al., 2009). On the other hand, online moni-
toring of every node of a large water distribution network is not possible or practical leading to the development of various sensor placement algorithms to determine optimal sensor locations that can provide the greatest benefit to the public given a limited number of sensors (e.g., Berry et al. (2006); Isovitsch and VanBriesen (2008); Preis et al. (2011); Watson et al. (2010); Xu et al. (2010)).

Additional development of contaminant source identification and source characterization algorithms have been used to identify potential contaminant source locations utilizing the sensor information. De Sanctis et al. (2010) developed a contamination source identification (CSI) algorithm to identify source locations utilizing binary sensor data and the particle backtracking algorithm developed by Shang et al. (2002). This algorithm by De Sanctis et al. (2010) provided a much faster alternative to other source identification algorithms, especially those which are based on the optimization of simulated observations at sensors for numerous potential injection scenarios. While this approach provided a framework for fast source identification, the CSI did not consider the uncertainties that accompany the hydraulics and sensor information in a water distribution system. In addition to hydraulic uncertainty, sensors will also produce false positive and false negative readings that must be accounted for in source identification. Yang and Boccelli (2013) addressed this issue and extended the CSI algorithm by developing a probabilistic contamination source identification (PCSI) algorithm that was capable of incorporating sensor uncertainty for estimating the source probability.

The PCSI algorithm by Yang and Boccelli (2013) relies on the hydraulic connectivity of sensors with upstream nodes at previous times to perform the source identification and hence its performance is critically dependent on the number and placement of the sensors. As the number of sensors available within a network will almost always be inadequate, as they are limited by the budget of a municipality, the sparse data obtained from these sensors will result in false identification of nodes.
as either contaminated or not contaminated. There will also be many locations for which the sensors will not be able to provide any information due to lack of hydraulic connection(s) with those nodes. One approach to fill in these information gaps is to perform confirmatory sampling at different locations during a possible contamination event. The task of finding a location to sample to increase information about a possible contamination event is complicated since the location that would maximize any benefit is dependent on several factors, but most importantly will depend on the potential spread of the contaminant plume in the network during the time of sampling. Although there has been significant effort in characterizing and identifying contaminant source locations, (e.g., Guan et al. (2006); Kumar et al. (2012); Laird et al. (2005); Neupauer et al. (2010); Wagner and Neupauer (2012); Yang and Boccelli (2013)) little has been done to utilize these results for the prediction of contaminant spread for subsequent planning of mitigation actions (Haxton and Uber, 2010). This study aims to bridge the gap between event detection, source identification and response action. More specifically, an algorithm will be developed to predict the spread of contamination and, utilizing that spread information, confirmatory sampling locations will be identified to increase the overall information associated with the system.

The proposed research has two major objectives, the first objective is to develop a software framework to make probabilistic contamination spread predictions by translating the information associated with upstream source nodes to future downstream output nodes. The PCSI algorithm will be used to analyze real-time binary sensor information from online sensors to obtain upstream source node probabilities. The probabilities of these upstream source nodes will be translated downstream by assuming that their impacts are proportional to their flow contributions to the downstream nodes. The second objective is to develop a sampling location selection algorithm, utilizing the forecasted data and the concepts of entropy from Information Theory. The term “entropy” from Information Theory will be used to quantify the amount
of information available about the status of the nodes in the network. This concept will be utilized to search for a sampling location that will correspond to the greatest increase in the amount of information when sampling at a future time given the forecasted probability of a downstream location being contaminated at that time. The positive impact of this research will be a more comprehensive and integrated software framework associated with the detection, assessment, and mitigation of contamination events.
Chapter 2

Methodology

2.1 Spread Forecasting

At the onset of an event detected by a CWS, the PCSI algorithm developed by Yang and Boccelli (2013) will utilize the binary signals from existing online sensors to estimate the probability of possible contamination sources. In addition to identifying potential injection locations, the PCSI algorithm also characterizes the contamination status of nodes at different past times. To determine the contamination status, the algorithm relies on hydraulic connection of the sensors with the upstream node-time (N-T) pairs. Hence the PCSI cannot assign any probability to those upstream N-T pairs that are not hydraulically connected to the sensors. If a large number of N-T pairs are characterized as contaminated, estimating the spread of contaminant can become computationally intensive if injections are simulated at all those N-T pairs individually and superimposing the forward simulation results.

In the proposed method, the spread of the contamination at downstream nodes will be estimated by translating the flow weighted source probabilities of the upstream nodes obtained using the PCSI algorithm. In order to calculate the flow weighting factors, the backtracking algorithm EPANET-BTX (Shang and Uber, 2009) will be
used. Given a set of forecasted demands, EPANET-BTX can generate the impact coefficients of a particular upstream input N-T pair on a downstream output N-T pair. An impact coefficient is defined as the sensitivity of the output concentration to the strength of the input. Two types of contaminant source injections can be modelled in EPANET-BTX, namely the “Mass Booster” and the “Flow Paced Booster” (Rossman, 2000). When the Flow Paced Booster source is used for a conservative tracer (no reduction in chemical concentration with time) the impact coefficients obtained from EPANET-BTX are equivalent to flow weights; that is, the impact coefficients represent the fraction of water reaching a downstream N-T pair from an upstream N-T pair.

To forecast the contaminant spread at a future hour for an output node $i$ the following steps will be followed: First, all the upstream N-T pairs that are physically connected to the downstream output node at the forecasted time will be identified. Second, using EPANET-BTX the impact coefficients of the physically connected upstream N-T pairs on the output node will be estimated. Finally, using the upstream source N-T pair probabilities, obtained using the PCSI algorithm, the probability of the output N-T pair will be estimated by an impact coefficient weighted (or flow-weighted) average of the source probabilities, which can be mathematically constructed as:

\[
p_{i}^{T+\Delta t} = \frac{\sum_{j \in J} I_{ij}^T p_{j}^{T}}{\sum_{j \in J} I_{ij}^{T}}
\]  

(2.1)

where $p_{i}^{T+\Delta t}$ is the estimated probability of observing contamination at node $i$ at the forecasting hour $T + \Delta t$, $T$ being the time up to which the sensor data has been collected and $\Delta t$ is the period of time ahead up to which the forecast is desired; $j$ is an upstream node whose impact on the output node is being considered; $J$ is the set that contains all the nodes that are physically connected to the output node; $I_{ij}^{T}$ is the flow weighted impact of the upstream node $j$ at a past time $t$ on the downstream node $i$ at the forecasting hour, $T + \Delta t$; and $p_{j}^{T}$ is the probability of node $j$ being a
contaminant source at the past time $t$.

A simple example is illustrated in Figure 2.1 to demonstrate the working principle of the spread forecasting algorithm. In this example the node labelled “Target Node” in Figure 2.1a, is connected to four different nodes and the directions of water flow along different pipes are marked with arrows. The Target Node is receiving water from Node 1 and Node 2, while providing water to Node 3 and Node 4. Assuming a contamination event has been detected at hour 5 and the current analysis time is hour 6, a forecast is desired of the target node at hour 7. From the assumed hydraulics it can be estimated that Node 1 contributes 78% of the flow to the target node, and the time delay for the water to travel to the Target Node has been calculated to be 2 hours. Therefore, the contamination status of Node 1 at hour 5 is required, which is shown in Figure 2.1b. Similarly, if Node 2 contributes 22% of the flow with a time delay of 3 hours then the contamination status of Node 2 at hour 4 is required, which is shown in Figure 2.1c. The probabilities of these two upstream nodes (shown in Figure 2.1b and 2.1c) will be flow-weighted to obtain the probability of the Target Node at hour 7, which will be $(0.9 \times 0.78 + 0.8 \times 0.22) ÷ (0.78 + 0.22) = 0.88$.

While forecasting for a particular node, there might be upstream N-T pairs whose
probability of being a contaminant source will not be known due to the lack of hydraulic connection of those N-T pairs with the sensors. In such cases, those nodes will be replaced from the set \( J \) with the next upstream nodes that are physically connected to them and the algorithm will be repeated for those newly added nodes. In certain cases, for particular flow paths, replacing a N-T pair with unknown source probability with its upstream N-T pair(s) may result in one or more N-T pairs with unknown probability. In such cases, the algorithm will continue replacing the N-T pairs with their upstream connections, travelling backwards in time, until an upstream N-T pair with known source probability is reached or until a N-T pair is reached where further backtracking cannot be continued. In the case where the probability of only a part of the flow can be traced back to a nearest upstream N-T pair, the spread will be calculated by assuming that portion to represent the whole flow. On the other hand if no upstream N-T pair with known contamination status can be found, the spread for that particular N-T pair will not be possible and an “unknown” status will be assigned.

By only backtracking to the node with the next measurable information, the spread is estimated upon the best information available at the time of the forecasting. While increasing the number of sensors is one way to attain more information in estimating the N-T pair probabilities of being a source, a potentially more practical alternative is to perform confirmatory or grab sampling in the field. This additional data from a confirmatory sample should improve the amount of information about the potential spread of a contaminant within a water distribution system. However, different locations will provide different amounts of information when sampled at a particular time. Thus, it is necessary to identify potential sampling locations that will provide the highest amount of useful information.
2.2 Sampling Location Selection

At the onset of an event, grab samples can be taken from the network to supplement the existing sensor data to provide additional information about the event. These grab samples can lead to an increase in the amount of available information about the upstream N-T pair probabilities and can potentially improve the estimated contaminant spread. The degree of additional information will, however, depend on the location and time of sampling. In order to select the location that will maximize the information gain, an algorithm will be developed utilizing Shannon’s entropy from Information Theory (Reza, 1961). In Information Theory, entropy is used as a metric of information and to evaluate the amount of information that a finite discrete probability scheme contains. For example, if a particular event or outcome $E_k$ is defined over the entire probability space of $n$ events and has a probability of occurring equal to $p_k$, then the amount of self-information associated with that event is $-\log_2 p_k$. Entropy, $En$, is then the average self-information of all the events in the probability space and can be mathematically expressed as:

$$En = -\sum_{k=1}^{n} (p_k \log_2 p_k)$$  \hspace{1cm} (2.2)

For a binary event space, as in the case of a N-T pair probability, the value of $n$ will be 2. The base of log can be arbitrarily chosen and, for communication theory, a base of 2 is historically chosen for binary signals which is also chosen here (Reza, 1961). Conceptually, lower entropy is associated with higher amounts of information while the opposite is true for higher entropy. Figure 2.2 shows the plot of entropy of an event space with two possible outcomes. Entropy reaches a maximum value for the probability of 0.5, which represents the least amount of information about an event since both outcomes are assigned an equal probability of 0.5. On the other hand, when there is no uncertainty about the outcomes, that is the probability is equal to
one or zero, the entropy is a minimum and equals zero. The probability that a single N-T pair is a contaminant source forms a complete finite probability scheme, with two possible outcomes contaminated (positive) and not contaminated (negative). The metric “Network Entropy”, (NE), will be defined here as the sum of the individual entropies of all the N-T pairs where the window of time will range from the current analysis time to a certain period backwards expressed as,

\[
NE = - \sum_{i=1}^{N} \sum_{t=t_0}^{T} \left( p_i^t \log p_i^t + (1 - p_i^t) \log (1 - p_i^t) \right) 
\]  

(2.3)

where, \( N \) is the total number of nodes; \( T \) is the current analysis time; and \( t_0 \) is the starting point of the time window. Thus, NE quantifies the amount of information that is available about the network at a particular time. A change in NE in response to a change in sensor information will indicate the amount of increase or decrease of information about the network associated with that particular change in sensor information. The best location to sample at a future time will then be the location
that would minimize the NE when sampled at that time.

The amount of change in NE when sampled at a certain location will depend on whether contamination has reached that location or not. In this research, the metric “Expected Network Entropy” (ENE), which is the probabilistic average of NE for an assumed positive and negative signal, will be used to predict the change in entropy that would result when sampled at a particular location at a future time. ENE can be expressed mathematically as,

\[ \mathbb{E}[NE] = NE_{+ve}p + NE_{-ve}(1 - p) \]  \hspace{1cm} (2.4)

where \( \mathbb{E}[NE] \) is the expected network entropy; \( NE_{+ve} \) and \( NE_{-ve} \) are the network entropies for an assumed positive signal and negative signal respectively; and \( p \) and \( 1 - p \) are the estimated forecasted probability of observing a positive and negative signal respectively.

The optimal sampling location at a forecasting hour \( T + \Delta t \) will be determined using complete enumeration by implementing the following steps: First, the spread forecasting algorithm described in section 2.1 will be used to estimate the probability of a particular location to be contaminated at the forecasting hour. If no forecasting is possible for that node a probability of 0.5 will be assumed. Second, a positive and a negative signal will be assumed to have occurred at that location at \( T + \Delta t \) hour and, treating them as new sensor information, the PCSI results will be updated separately for the two signals. Third, the NE for the updated source probabilities for both the signals will be calculated using eqn. 2.3 with the ENE calculated using eqn. 2.4. Finally, the above steps will be repeated for all locations within the network (except for the locations of the existing sensors) to find the location that produces the minimum ENE. Mathematically the expected network entropy for the forecasted
sampling can be expressed as:

\[
E[NE]_i^{T + \Delta t} = NE_{+ve}^{T + \Delta t} p_i^{T + \Delta t} + NE_{-ve}^{T + \Delta t} (1 - p_i)^{T + \Delta t} \tag{2.5}
\]

where \(E[NE]_i^{T + \Delta t}\) is the ENE for the sampling location \(i\) at \(T + \Delta t\) hour; \(NE_{+ve}^{T + \Delta t}\) and \(NE_{-ve}^{T + \Delta t}\) are the network entropies for an assumed positive and negative signal at \(T + \Delta t\) hour respectively; and \(p_i^{T + \Delta t}\) and \((1 - p_i^{T + \Delta t})\) are the forecasted probability of observing a positive signal and negative signal respectively (estimated using eq. 2.1) at node \(i\) at \(T + \Delta t\) hour.
Chapter 3

Results

The proposed algorithms were applied to two hypothetical water distribution networks with increasing complexity and size. The first network was a small 97 node water distribution network, Net3 (Figure 3.1), which is distributed with EPANET (Rossman, 2000). The second network was a 12,523 node water distribution network, Network2 (Figure 3.2), which was used during the “The Battle of the Water Sensors Network” (BWSN) (Ostfeld et al., 2008). Sensor locations for both networks were obtained from Yang and Boccelli (2013), where the authors used TEVA-SPOT (Berry et al., 2010) to locate sensor locations for both the networks. The results of the application of the algorithms are presented in following sections.

3.1 Small Network: Net3

The Net3 network is shown in Figure 3.1 along with the injection and sensor locations. To simulate a contamination event an hour long injection of a conservative chemical was simulated at node 10 starting at hour 3. The injection was simulated as a Flow Paced Booster with a concentration of 100 mg/L. A false positive and false negative rate of 10% was assumed for all sensors and also for the confirmatory sampling results. The base demands and the demand patterns of the nodes were kept the same as
Figure 3.1: Layout of the small test network Net3. The injection location has been represented with a cross and the sensor locations with rectangles.
Figure 3.2: Layout of the large test network Network2 along with sensor locations for the (a) 5 sensor; (b) 10 sensor; (c) 20 sensor; and (d) 50 sensor cases. The injection location has been marked with a cross.
assigned to them in the original network input file. Contaminant concentrations at
the sensor locations, calculated using EPANET, were converted into binary signals
assuming a threshold concentration of $10^{-6}$ mg/L and fed into the PCSI algorithm.
First detection of contamination was made at 5.25 hours by the sensor at node 193.

To test the forecasting algorithm, sensor measurements were acquired until the 6th
hour at 15 minute intervals and then the forward simulation was stopped. The spread
forecasting algorithm was executed for the 6th through 10th hours with the spread es-
timations summarized in Figure 3.3. Nodes with probability higher than or equal to
0.50 were considered “contaminated” and nodes with probability less than 0.50 were
considered “not contaminated.” Three quantities - percent correct, percent incorrect
and percent unconnected (no information), were analyzed for all nodes at each hour
from the beginning of simulation to the end of the forecasting horizon. Here, the
percent correct/incorrect is defined as the percentage of nodes correctly/incorrectly
identified as contaminated or not contaminated. Percent unconnected, which is appli-
cable to the PCSI results only, is the percentage of nodes that were not hydraulically
connected to any of the sensors; hence the PCSI algorithm could not assign any prob-
ability those nodes. Similarly, percent “no forecast”, (applicable to the forecasted
results only) is the percentage of nodes for which no probability could be forecasted.
Both unconnected and no forecast nodes represent nodes for which no information is
available.

In Figure 3.3, plots from hour 0 (start of simulation) to hour 6 constitute the re-
results from the the PCSI algorithm. The percentage of correct identification decreased
with increasing percentage of “unconnected” nodes (nodes with no information). This
result was not unexpected, because, as the current simulation time (hour 6) was ap-
proached, increasing numbers of upstream nodes lost hydraulic connection with the
sensors and their status became unknown. The percentage of incorrect identification
also increased with decreasing hydraulic connections with the sensors. At the current
Figure 3.3: Spread forecasting summary for the Net3 network. PCSI results (solid lines) up to hour 6 and forecasting results (dashed lines) for hours 6 through 9, showing percent correct, incorrect and unconnected (PCSI)/no forecast (forecasting) nodes; sensor data was acquired up to hour 6 for this forecasting.

Simulation time, all upstream nodes lose hydraulic connections with the sensors, and the percent correct and incorrect identification became zero while the percent unconnected became one hundred percent. Forecasting results for the same quantities are shown, from hour 6 to hour 9, with dashed lines. The percentage of correct and incorrect identification for the forecasted probability decreased with the forecasting horizon, while the percentage of nodes with no forecasted probability increased. The forecasted probabilities do not rely on hydraulic connections with the sensors, but only with the hydraulic connections to the upstream nodes. With increasing forecast horizon, estimating the information about the nodes at the forecasted time required more upstream N-T pair information; the lack more information about previous N-T pairs resulted in an increase in the number of nodes with no possible forecasted probability.

The sampling location selection algorithm was employed to estimate the best
sampling location at hour 7 utilizing the sensor information observed up to hour 6. Figure 3.4a shows the expected change in entropy for all possible sampling locations relative to the baseline entropy (entropy without any sampling) if sampled during the 7th hour. Nodes 1 and 2 were expected to produce the two largest decreases in entropy. Nodes 40 and 179 along with nodes 237, 206, 208, 209, 211 and 213 (peaks inside the gray box in Figure 3.4a) were also expected to produce a significant decrease in entropy, although not as prominent as nodes 1 and 2.

To determine the location that would have actually produced the minimum entropy, forward simulation was continued until hour 7 and the sensor data was collected, assuming a grab sample was collected for each possible location at the 7th hour. The data was then used to update the PCSI results and update the NE calculation, as previously described, for each of the potential sampling locations, termed hereafter as “Actual Network Entropy” (ANE). Figure 3.4b illustrates how the ANE would have changed for all the possible sampling locations when sampled during the 7th hour. Nodes 1 and 2 were still shown to produce large decreases in entropy. However, nodes 40 and 179 were also found to produce quite large decreases in entropy, which were not expected from the ENE enumeration (Figure 3.4a). A plot of the cumulative relative frequency of the ANE, created using a Blom plotting position, is shown in Figure 3.5. The ANEs for the top three estimated sampling locations (i.e., the three lowest ENE nodes), are also shown in this figure. Although the first two estimated sampling locations were also evaluated to have the two lowest ANEs, the third location was not, however, the third location was within the bottom 5% of the ANE.

Figure 3.6 shows the spatial locations of the sampling nodes highlighted in Figure 3.4. Two distinct regions have been identified where the potential sampling locations appear to cluster. The first region consists of the tank at node 1 with nodes 40 and 179 connected immediately before the tank. The second region includes the other
Figure 3.4: (a) Changes in ENE for sampling at all the locations during the 7th hour in the Net3 network. The peaks inside the gray box corresponds to the cluster of nodes inside the lower gray box in Figure 3.6; and (b) changes in ANE for sampling at all the potential sampling locations during the 7th hour.

tank at node 2 along with nodes 237, 206, 208, 209, 211 and 213 just upstream of the tank.

Figure 3.7 demonstrates how sampling at different locations during the 7th hour will change the PCSI results. Figure 3.7a, shows the percentage of nodes correctly identified at each hour for all potential sampling locations. As before, correct identification means that the PCSI algorithm correctly assigned a probability higher or equal to 0.5 for a contaminated node and lower than 0.5 for an uncontaminated node. The baseline case is shown, which is the case without any sampling, along with plots for sampling at nodes 1 and 2, which were the best estimated sampling locations. The shaded region show the range of correct identification for sampling at all remaining locations. Similarly, Figure 3.7b shows the percentage of nodes incorrectly identified, and Figure 3.7c shows the percentage of unconnected nodes, for which no information was available. It was found that if node 1 was used as a sampling location at hour 7, the PCSI results would have had the greatest increase in correctly identified N-T pairs and greatest decrease in the percentage of N-T pairs with no information. However, the percent incorrect identification did not improve but was seen to decrease at hour
Figure 3.5: Cumulative relative frequency of ANE for sampling during the 7th hour at all the potential locations in the Net3 network; entropies of the top three estimated sampling locations are indicated.

Figure 3.6: Spatial locations of the top estimated sampling locations for the 7th hour in the Net3 network. Insets show blown up images of their locations. Darker shades indicate greater potential to decrease entropy. The tanks at node 1 and 2 were estimated to be the two best locations.
3 and increase at hour 5 with an overall increase in incorrect identifications. Node 2, which was expected to be the second best sampling location in terms of entropy, also would have produced a good increase in the percentage of correctly identified N-T pairs immediately following the performance of node 1. Similarly, sampling at node 2 would have produced a meaningful decrease in the percentage of N-T pairs with no information closely following the performance of node 1. On the other hand, node 2 performed better, in terms of percent incorrect identification, as it does not differ from the baseline case in contrast to the slight increase in the case of node 1. In terms of characterization of contamination status of N-T pairs by the PCSI algorithm, node 1 was shown to provide the greatest overall benefit followed by node 2 as was expected by the sampling location selection algorithm.

However, it is to be noted that in Figure 3.4b, nodes 40 and 179 were shown to produce large decreases of entropy, yet they did not have considerable impact on the percent correct, incorrect and unconnected N-T pairs. Further investigating into these locations showed that sampling at these two locations modified the contamination status of fewer N-T pairs but that the degree of modification was such as to result a large decrease of NE. On the other hand, sampling at nodes 1 and 2 modified a greater number of N-T pairs, relative to nodes 40 and 179, resulting in a similarly large decrease of NE. Another point to consider is the fact that the entropy function is parabolic (see Figure 2.2) and the rate of change of entropy increases as the probability moves away from 0.5 either towards zero or one. A simple example would be that a change of probability from 0.8 to 0.9 will decrease entropy by a greater amount than a change of probability from 0.6 to 0.7. Thus, if a sampling location modified the probability of a certain number of nodes that had probability closer to 0 or 1, the decrease of NE would be larger than if it had modified the probability of the same number of nodes, by the same amount with a probability closer to 0.5. Therefore, since nodes 40 and 179 mostly modified the probability of N-T pairs that already
Figure 3.7: Changes in: (a) % correct identification; (b) % incorrect identification; and (c) % unconnected, in the PCSI results for sampling during hour 7 at different locations in the Net3 network. The baseline case is shown along with plots for sampling at node 1 and 2. The range of change for sampling at all the remaining locations is shown in gray.
had some probability assigned to them, their contribution in decreasing entropy was more because of the non-linear nature of entropy rather than because of providing information about a greater number of N-T pairs.

3.2 Large Network: Network2

Node JUNCTION-5416 (see Figure 3.2) was selected to be the contamination injection location for this network. Injection of a conservative tracer was simulated at this node at hour 3 for a duration of one hour. The injection was simulated as a Flow Paced Booster with a concentration of 100 mg/L. The number of sensors were varied for this network and four cases, with 5, 10, 20 and 50 sensors, were analyzed. A false positive rate (FPR) and false negative rate (FNR) for the sensors were assumed to be the same for all the cases and two different false rates, 10% and 30%, were used to analyze the algorithms for each of the four sensor cases. Sensor data was collected at every 15 minute interval starting from the beginning of simulation at hour zero. The base demands and the demand patterns of the nodes were kept the same as were assigned to them in the original network file. Contaminant concentrations at the sensor locations, calculated using EPANET, were converted into binary signals using a simple threshold analysis and fed into the PCSI algorithm. A sensor detection threshold of $10^{-5}$ mg/L was used for an injection of 100 mg/L. After collecting sensor data up to a certain time, which is termed as the “current simulation time”, the contaminant spread and confirmatory sampling locations were estimated for subsequent hours. For the four different sensor cases, Table 3.1 shows time of first detection of the contamination event, and the time to which the simulation continued (with data being collected by the sensors) before the forecasting and confirmatory sampling location selection algorithms were applied.

The results of the PCSI algorithm up to the sensor data collection time, in Table
<table>
<thead>
<tr>
<th>Number of Sensors</th>
<th>First Detection</th>
<th>Node ID</th>
<th>Sensor Data Collection Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>9.5</td>
<td>JUNCTION-11304</td>
<td>11</td>
</tr>
<tr>
<td>10</td>
<td>9.25</td>
<td>JUNCTION-10614</td>
<td>11</td>
</tr>
<tr>
<td>20</td>
<td>7.50</td>
<td>JUNCTION-5365</td>
<td>9</td>
</tr>
<tr>
<td>50</td>
<td>7.50</td>
<td>JUNCTION-5365</td>
<td>9</td>
</tr>
</tbody>
</table>

Table 3.1: Time to first detection of contamination for different sensor cases for Network2 and node IDs of the sensors that detected the contamination. The sensor data collection hour is also shown up to which sensor data was collected for subsequent spread forecasting and sampling location estimation.

3.1, and the spread forecasting algorithm for up to four hours of forecasting horizon for the different sensor cases are presented in Figure 3.8. The sensor data collection times for these forecast estimations are given in Table 3.1. The solid lines in these figures represent the N-T pair characterization of previous hours, estimated by the PCSI algorithm, and the dashed line represents the N-T pair characterization of future hours estimated by the forecasting algorithm. For the four hours of forecast horizon, the percent correct identification increased for all sensor cases except for the 50 sensor case. On the other hand, the percent incorrect identification increased and percent of nodes with no forecasted probability decreased for all sensor cases.

Using these forecasted data, the sampling location selection algorithm was utilized to estimate potential sampling locations at one hour after the current simulation time (sensor data collection hour). Figure 3.9 shows the cumulative relative frequency plots of ANE for all potential sampling locations in the network. The change of ANE for the estimated best locations (i.e., the lowest ENE location) are also shown in the plots. As described before, ANE is the NE calculated by obtaining sensor data up to the sampling hour along with a single confirmatory sample from all possible sampling locations. The shapes of the ANE cumulative relative frequency plots for the four sensor cases were very similar to each other. The plots are almost horizontal at the low end of entropy indicating a large variation in entropy among the top potential
Figure 3.8: PCSI and spread forecasting results for Network2 including the percent correct, incorrect and unconnected/no forecast nodes for the: (a) 5 sensor case; (b) 10 sensor case; (c) 20 sensor case; and (d) 50 sensor case. The vertical lines inside the plots mark the current simulation time up to which sensor data has been collected.
locations. This may be a desirable trait as it would make the best locations more distinct in terms of entropy than the rest of the nodes. The plots then start to rise rapidly and become almost vertical at a particular entropy indicating that the majority of the nodes (about 90%) have entropy lower than this particular value. The range of ANE, which is the difference between maximum and minimum ANE, generally decreased for increasing number of sensors, which is not unexpected, as with increasing number of sensors the benefits from a confirmatory sampling is expected to diminish. Also, the absolute magnitude of ANEs decrease with the increasing number of sensors due to the increase in the amount of information from the sensors. It was observed from these plots, that the estimated best sampling location, which is the lowest ENE location, do not always coincide with the lowest ANE location. The spatial details of the best ANE and best ENE locations are discussed in the next paragraph.

Figure 3.10 shows the locations of the estimated (using ENE) and actual (using ANE) best sampling locations in the network. For the five sensor case, the estimated best sampling location ranked 10\textsuperscript{th} when evaluated in terms of ANE. The actual best location (i.e., the lowest ANE location) was estimated as the 7\textsuperscript{th} best location to sample, but, physically, they were not very far apart in the network (Figure 3.10a). When ranked similarly, the estimated best sampling locations for the 10 and 20 sensor cases were also ranked the best in terms of ANE (Figure 3.10b and 3.10c). For the 50 sensor case, the estimated best location ranked 4\textsuperscript{th} in terms of ANE and the best ANE location was estimated to be the second best sampling location by the algorithm. Unlike the 5 sensor case, the best ENE location and the best ANE location were in different regions of the network. It is evident from these results that sampling at two different regions may provide very similar amount of information, albeit for a completely different set of N-T pairs. As in the case of the small network Net3, here too the top estimated sampling locations clustered together. Figures 3.11 through
Figure 3.9: Cumulative relative frequency plots of ANE for the: (a) 5 sensor case; (b) 10 sensor case; (c) 20 sensor case; and (d) 50 sensor case, for sampling at all locations in Network2. ANE of the estimated best locations are marked with squares.
3.14 show the top 1% (125 nodes) of the estimated potential sampling locations for the different sensor cases. For the five sensor case, shown in Figure 3.11, the top 1% locations are grouped in two clusters, both of them in the top right portion of the network. Two major clusters were also seen for the 10, 20 and 50 sensor cases. A small cluster at the lower left portion of the network showed up in the latter three sensor cases, which also contained the best location (in terms of ANE) - JUNCTION-3357, for the 50 sensor case. Clustering of the estimated best locations is advantageous in a sense that even if the actual best sampling location was not selected, comparable amounts of useful information would still be obtained when sampled from nearby regions.

The impacts of confirmatory samplings on previous N-T pair characterization by the PCSI algorithm is illustrated in Figures 3.15 through 3.18. Trends that were observed for the four sensor cases were very similar to each other and also very similar to the small network case. For the PCSI results, as the current simulation time was approached the percent correct and incorrect identification decreased as more and more N-T pairs lost hydraulic connections with the sensors. For the 5 sensor case, the impact of performing a confirmatory sampling at the estimated best sampling location (i.e. the lowest ENE location) and the actual best sampling location (i.e. the lowest ANE location) is illustrated in Figure 3.15a along with the case of no sampling (baseline case). The shaded region in this figure represents the range of percent correct identification for all possible sampling locations. The plots for the estimated and actual best sampling locations closely follow each other and are on the top of the shaded region indicating that they are indeed among the best places to sample. A similar plot for percent incorrect identification is illustrated in Figure 3.15b where the plots for the best locations lie at the bottom of the shaded region indicating minimum percent incorrect identification would result if sampled at those locations. Figure 3.15c shows the percent unconnected nodes, i.e., N-T pairs that were not hy-
Figure 3.10: Spatial locations of the estimated and actual best sampling locations for the: (a) 5 sensor case; (b) 10 sensor case; (c) 20 sensor case; and (d) 50 sensor case, in Network2.
Figure 3.11: Spatial locations of the top 1% of the estimated sampling locations for the 5 sensor case, in Network2.
Figure 3.12: Spatial locations of the top 1% of the estimated sampling locations for the 10 sensor case, in Network2.
Figure 3.13: Spatial locations of the top 1% of the estimated sampling locations for the 20 sensor case, in Network2.
Figure 3.14: Spatial locations of the top 1% of the estimated sampling locations for the 50 sensor case, in Network2.
draulically connected to the sensors. For the aforementioned best sampling locations (the best ENE and ANE locations), the decrease in the percent of unconnected nodes were among the highest relative to the other locations. Improvements in the percent correct, incorrect and percent unconnected showed similar trends for the remaining sensor cases, but the degree of improvement decreased with increasing number of sensors. It is evident from these figures that, as the sensor number increases, the impacts of confirmatory samplings on the percent correct, incorrect and unconnected nodes decrease. This is because, with increasing density of sensors, the number of N-T pairs with existing information will increase and consequently the number of new N-T pair information that a particular grab sample could provide would decrease.

To analyze the impacts of confirmatory sampling on spread forecasting, sensor data was collected up to the sampling hour with additional data from the sampling location included during that hour. Then the contamination spread was estimated for the sampling hour and three hours into the future. Figure 3.19 shows the impacts of confirmatory sampling on the forecasted N-T pair probability estimation for the 5 sensor case. In this case, sensor data was collected up to hour 12 and, during the 12th hour, an additional confirmatory sample was included for PCSI analysis and subsequent spread calculations. Figure 3.19a shows the impacts of sampling on percent correct identification of the forecasted N-T pair probabilities at the estimated best sampling location - JUNCTION-448, and at the actual best sampling location (according to ANE) - JUNCTION-10740. The spread forecast was actually improved by a greater amount by the estimated best sampling location than the actual best sampling location in this case. Similarly, Figure 3.19b and 3.19c shows the percent incorrect identification and percent of N-T pairs with no possible forecast, respectively. Confirmatory samples were shown to decrease both the percent incorrect identification and percent of nodes with no possible forecast. For all the three quantities, the estimated best sampling location outperformed the sampling location with the
Figure 3.15: Changes in: (a) % correct identification; (b) % incorrect identification; and (c) % unconnected, in the PCSI results for sampling during hour 12 at all the potential locations in Network2, for the 5 sensor case. The baseline case is shown along with plots for sampling at the estimated and actual best locations. The range of change for sampling at all the locations is shown in gray.
Figure 3.16: Changes in: (a) % correct identification; (b) % incorrect identification; and (c) % unconnected, in the PCSI results for sampling during hour 12 at all the potential locations in Network2, for the 10 sensor case. The baseline case is shown along with plots for sampling at the estimated and actual best locations. The range of change for sampling at all the locations is shown in gray.
Figure 3.17: Changes in: (a) % correct identification; (b) % incorrect identification; and (c) % unconnected, in the PCSI results for sampling during hour 10 at all the potential locations in Network2, for the 20 sensor case. The baseline case is shown along with plots for sampling at the estimated and actual best locations. The range of change for sampling at all the locations is shown in gray.
Figure 3.18: Changes in: (a) % correct identification; (b) % incorrect identification; and (c) % unconnected, in the PCSI results for sampling during hour 10 at all the potential locations in Network2, for the 50 sensor case. The baseline case is shown along with plots for sampling at the estimated and actual best locations. The range of change for sampling at all the locations is shown in gray.
Figure 3.19: Impacts of sampling during the 12\textsuperscript{th} hour at the estimated and actual best sampling locations on: (a) \% correct identification; (b) \% incorrect identification; and (c) \% no forecast, of the forecasted spread, for the 5 sensor case for Network2.

lowest ANE. Similar plots for the 10, 20 and 50 sensor cases are shown in Figures 3.20 through 3.22, respectively. In all the cases, confirmatory sampling improved the forecast, but as in the characterization of past N-T pairs by the PCSI algorithm, here too, the diminishing impact of confirmatory sampling could be observed with increasing number of sensors.

\textit{Impact of false rates of the sensors.} The results discussed above considered a false positive and false negative rate of 10\% for both the sensor and sampling data. All these results for Network2 were reanalyzed by changing the false rates from 10\% to 30\%. The impact on the PCSI characterization was very little - percent correct and incorrect identification changed by less than one percent, as was the impact on
Figure 3.20: Impacts of sampling during the 12\textsuperscript{th} hour at the estimated and actual best sampling locations on: (a) \% correct identification; (b) \% incorrect identification; and (c) \% no forecast, of the forecasted spread, for the 10 sensor case for Network2.
Figure 3.21: Impacts of sampling during the 10th hour at the estimated and actual best sampling locations on: (a) % correct identification; (b) % incorrect identification; and (c) % no forecast, of the forecasted spread, for the 20 sensor case for Network2.
Figure 3.22: Impacts of sampling during the 10th hour at the estimated and actual best sampling locations on: (a) % correct identification; (b) % incorrect identification; and (c) % no forecast, of the forecasted spread, for the 50 sensor case for Network2.
spread forecast estimation. Results from the sampling location selection algorithm were also very similar, although, some reordering among the top estimated potential sampling locations was observed. On the other hand, the actual network entropy (ANE) changed significantly with the change of false rates and the estimated sampling locations were not ranked as high as for the lower false rate cases. Table 3.2 shows the ranks of the top five estimated sampling locations for the two false rate scenarios. For the 5 sensor case, four of the top five sampling locations for the false rate of 10% were still selected using the false rate of 30%, but their ANE rankings were much lower than the lower false rate case. Similar results for the 10, 20 and 50 sensor cases were also observed, although the effect of higher false rate was found to be compensated by the increasing number of sensors and the ANE ranking of the estimated locations improved. Even though the estimated locations were not ranked as high as before, they were still found to outperform the locations that had high ANE rankings, as they did better in terms of N-T pair characterization (percent correct, incorrect and unconnected nodes). Visual inspection of the spatial distribution of the top 1% of the estimated locations did not show any significant differences.
<table>
<thead>
<tr>
<th>Estimated Best Sampling Locations</th>
<th>ANE Rank</th>
<th>Estimated Best Sampling Locations</th>
<th>ANE Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>5 Sensors</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>JUNCTION-448</td>
<td>10</td>
<td>JUNCTION-448</td>
<td>129</td>
</tr>
<tr>
<td>JUNCTION-4032</td>
<td>67</td>
<td>JUNCTION-11177</td>
<td>110</td>
</tr>
<tr>
<td>JUNCTION-1153</td>
<td>20</td>
<td>JUNCTION-4032</td>
<td>430</td>
</tr>
<tr>
<td>JUNCTION-11177</td>
<td>5</td>
<td>JUNCTION-4075</td>
<td>102</td>
</tr>
<tr>
<td>JUNCTION-12289</td>
<td>87</td>
<td>JUNCTION-1153</td>
<td>154</td>
</tr>
<tr>
<td><strong>10 Sensors</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>JUNCTION-448</td>
<td>1</td>
<td>JUNCTION-448</td>
<td>36</td>
</tr>
<tr>
<td>JUNCTION-11177</td>
<td>2</td>
<td>JUNCTION-11177</td>
<td>47</td>
</tr>
<tr>
<td>JUNCTION-11191</td>
<td>3</td>
<td>JUNCTION-11191</td>
<td>26</td>
</tr>
<tr>
<td>JUNCTION-11126</td>
<td>4</td>
<td>JUNCTION-11126</td>
<td>31</td>
</tr>
<tr>
<td>JUNCTION-11193</td>
<td>9</td>
<td>JUNCTION-11174</td>
<td>77</td>
</tr>
<tr>
<td><strong>20 Sensors</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>JUNCTION-11191</td>
<td>1</td>
<td>JUNCTION-11191</td>
<td>17</td>
</tr>
<tr>
<td>JUNCTION-11126</td>
<td>2</td>
<td>JUNCTION-11126</td>
<td>18</td>
</tr>
<tr>
<td>JUNCTION-11192</td>
<td>4</td>
<td>JUNCTION-11192</td>
<td>21</td>
</tr>
<tr>
<td>JUNCTION-11190</td>
<td>3</td>
<td>JUNCTION-11190</td>
<td>29</td>
</tr>
<tr>
<td>JUNCTION-4032</td>
<td>5</td>
<td>JUNCTION-4032</td>
<td>10</td>
</tr>
<tr>
<td><strong>50 Sensors</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>JUNCTION-11177</td>
<td>4</td>
<td>JUNCTION-11177</td>
<td>1</td>
</tr>
<tr>
<td>JUNCTION-3357</td>
<td>1</td>
<td>JUNCTION-11174</td>
<td>6</td>
</tr>
<tr>
<td>JUNCTION-3356</td>
<td>3</td>
<td>JUNCTION-11176</td>
<td>7</td>
</tr>
<tr>
<td>JUNCTION-3358</td>
<td>2</td>
<td>JUNCTION-11193</td>
<td>9</td>
</tr>
<tr>
<td>JUNCTION-11174</td>
<td>8</td>
<td>JUNCTION-4075</td>
<td>11</td>
</tr>
</tbody>
</table>

Table 3.2: Top five estimated sampling locations and their ANE rankings for the different sensor cases and different false rates, for Network2.
Chapter 4

Discussion

The primary objectives of this research were to forecast the contaminant spread in a water distribution network after the detection of a contamination event and identify potential confirmatory sampling locations to improve information about the system.

The lack of an integrated framework to translate information on network contamination to estimate the spread of contamination was the principle driving force of this research in order to identify potential confirmatory sampling locations.

*The spread forecasting algorithm.* A method was presented to make spread forecasts by utilizing the contamination status of N-T pairs provided by the source identification algorithm PCSI, developed by Yang and Boccelli (2013). To forecast the spread, the probabilities of N-T pairs have been treated as a conservative chemical with complete mixing at nodes. The reason for adopting this approach, as opposed to treating any node receiving water from a contaminated upstream node to be contaminated, was to account for the fact that some of the nodes may receive a very small quantity of contaminated water from its upstream sources and may not be detectable if sampling was performed at those locations. Thus, the method adopted here takes into account the dilution effect by flow weighting the probabilities of upstream N-T pairs. The spread forecasting results obtained by this method worked well, and the
The accuracy of the algorithm was dependent on the amount of past N-T pair information that was available and generally improved with increasing number of sensors.

The sampling location identification algorithm. A sampling location identification method was presented that utilized the forecasted spread to identify potential sampling locations during an event to improve the information associated with the overall distribution network. In this method, the concept of entropy from Information Theory was utilized to quantify information about the network. Expected change in entropy, i.e., change in information, was estimated by assuming a positive and a negative signal from a potential sampling location and then making a probabilistic average of their impacts by weighing the corresponding impacts with the probability of observing a positive and negative signal at that location. By enumerating the expected change in entropy for all locations, the best sampling locations were identified to be the ones producing the greatest decrease in entropy. This method worked well as sampling locations were selected that provided results that proved to be more beneficial than others. The benefits expected to be realized from a confirmatory sampling were evaluated using two different criteria. The first criterion was the actual network entropy (ANE), calculated using simulated sensor data up to the sampling hour to quantify the change in the amount of information about the network before and after sampling. The second criterion was the amount of change in the percent correct and incorrect identified nodes and the unconnected N-T pairs, as characterized by the PCSI algorithm and the spread forecasting algorithm before and after sampling. It was shown that the estimated potential sampling locations were among the top 1% of the lowest ANE producing nodes, and produced the greatest increases in percent correct identification and greatest decreases in percent unconnected nodes. However, the percent of incorrect identifications did not change significantly from the baseline cases but generally were among the lowest of all the potential sampling locations.

Information about unknown N-T pairs versus N-T pairs with existing information.
The PCSI algorithm only assigns probability to previous N-T pairs that are hydraulically connected to the sensors. Hence the PCSI results, in most cases, will have N-T pairs with unknown contamination status. A confirmatory sample may increase information about the network by providing new information about unknown N-T pairs or by providing additional information about known N-T pairs, and in most cases it will be a combination of both. The best sampling locations selected by the algorithm presented here predominantly increased information (or decreased entropy) by providing new information about unknown N-T pairs. Hence, it can be contended that the developed sampling location selection algorithm looks for a location that creates more new information rather than modifies existing information. Entropy, however, will decrease for both cases. Because of the parabolic nature of entropy, as discussed in Section 3.1, the change in entropy for obtaining new information about an unknown N-T pair may generally be lower than the change of entropy for a similar change in the probability of a N-T pair with existing information. This is because when new information is obtained about an unknown N-T pair, the probability of that N-T pair moves away from 0.5 to a certain value. Since the rate of change of entropy tends to zero as probability tends to 0.5, a change from 0.5 will always be lower than a similar change from any other value. Hence a sampling location that predominantly provides new information must be able to provide information about a larger number of unknown N-T pairs to outcompete (in terms of entropy) a sampling location that predominantly modifies existing information about N-T pairs. This is quite beneficial for the purpose at hand as knowledge about unknown N-T pairs can potentially lead to a greater improvement in the estimation of contaminant spread.

*False rate of sensors.* When the false rate of sensors increase, more uncertainty is introduced into their contamination status information. Characterization of a system using data from such sensors becomes more uncertain and sensor signals or confirmatory samples induce less change in information, hence there is less change in the
associated entropy than would occur for a lower false rate. Thus, a higher false rate caused an overall decrease of information about the distribution system. Also, due to the parabolic nature of entropy, the change of entropy for providing new information about an unknown N-T pair was more affected than the change of entropy for modifying the information of a N-T pair with existing information. For this reason, sampling locations that predominantly provided new information about N-T pairs were shown to have lower ANE ranks for the higher false rate case.

Hydraulic uncertainty. The presented spread forecasting and sampling location selection algorithms assumed known hydraulics with fixed demands and demand patterns, but actual hydraulics will rarely be the same as the assumed hydraulics. Also, at the onset of a contamination event, consumers may be made aware of a possible contamination, which will result in dynamic changes in water usage that could significantly affect both spread and optimal sampling location identification (Shafiee and Zechman, 2012; Yang and Boccelli, 2009). A short term demand forecasting algorithm, capable of forecasting demand during an emergency contamination event, can be coupled with the spread forecasting and sampling location identification algorithm, which can potentially improve the estimated spread and sampling location identification.

Portability of the developed algorithms. Both the algorithms presented here were developed to work in an integrated manner with the PCSI algorithm and relies intimately on PCSI results for N-T pair characterization. However, these algorithms are not limited to the PCSI algorithm and should work equally well with any other algorithm that can characterize N-T pair contamination status utilizing sensor information with similar or better accuracy than the PCSI algorithm.
Chapter 5

Conclusion

This research developed two algorithms to estimate the spread of contamination after
the detection of an event by a CWS, and to identify confirmatory sampling loca-
tions using the estimated contamination spread. The first algorithm estimated the
spread of contamination by flow weighting the estimated upstream source proba-
bilities. The source probabilities were obtained using the PCSI algorithm and the
flow weights were obtained using the backtracking algorithm, EPANET-BTX. The
second algorithm identified confirmatory sampling locations, utilizing the estimated
contamination spread, to improve the information about the overall system and the
contamination event.

The backtracking algorithm, EPANET-BTX, was preferred to a more intuitive
forward simulation approach because of the computational efficiency. Simulating mul-
tiple injections separately and superimposing individual contaminant concentrations
at different nodes can become computationally burdensome if the potential number
of sources are very large. The spread forecasting algorithm described here required
about 35 seconds to estimate the contaminant spread one hour ahead into the fu-
ture for the large network, Network2, with 10 sensors on a laptop running on a 2.1
GHz Intel Core i3 processor. Computation time generally increased with increasing
number of sensors and increased forecast horizon.

Confirmatory sample location selection was based on the degree of increased information, quantified using entropy, that a potential sampling location was expected to contribute. Conceptually, lower entropy corresponds to greater certainty about a stochastic event space. This concept was used to find the best sampling location by complete enumeration that would minimize the overall entropy of the entire system.

The developed algorithms were applied to two hypothetical water distribution networks of which the first network, Net3, was a small 97 node network and the second network, Network2, was a large 12,527 node network. For the small network, only one case with 5 sensors was analyzed and, for the larger network, four cases with 5, 10, 20 and 50 sensors were analyzed. Sensor locations for both these networks were obtained using TEVA-SPOT, as reported in Yang and Boccelli (2013). A false rate of 10% was used for all sensor cases and confirmatory sampling results. A single injection was simulated for both networks starting at hour 3 of the simulation with a duration of one hour.

For both the networks, the spread forecasting algorithm generated satisfactory results with the prediction accuracy dependent on the accuracy of the past N-T pair characterization accuracy, which increased with the number of sensors. Sampling locations were also correctly identified when evaluated in terms of changes in the characterization of N-T pairs by the PCSI algorithm and the spread forecasting algorithm, and were also among the top ANE locations. The estimated sampling locations identified for both the networks were found to cluster together. For the smaller network, the tanks were observed to be good places to sample as the two best locations were both tanks. Spatial locations of the top 1% of the estimated sampling locations, for the larger network, revealed two major clusters. Also, these clusters were found to be in very similar regions, especially for the 5 and 10 sensor cases, which were almost identical. With increasing number of sensors, the clusters seemed to become more
scattered as other regions emerged. The greatest amount of scattering was observed for the 50 sensor case, although the entropies remained similar.

The impact of a confirmatory sampling was evaluated in terms of ANE and improvement in N-T pair characterization by the PCSI and spread forecasting algorithms. For the smaller network the two estimated best locations were also the two lowest ANE nodes. Confirmatory samples from these locations, when fed into the PCSI algorithm, resulted in the greatest improvements in N-T pair characterizations. However, the confirmatory sampling for the smaller network had little impact on the forecasted spread. The most probable reason for this was that the number of sensors for such a small network had a diminishing effect on the outcome of sampling. Similar results were also observed for the larger network case as the number of sensors increased. The estimated sampling locations resulted in the greatest improvements in PCSI characterization and also improved the forecasting results for all the sensor cases. The impact of confirmatory sampling was found to diminish with increasing number of sensors, both for the PCSI results and spread forecasting results. As explained earlier, this is due to the fact that with increasing number of sensors there will be fewer number of unknown N-T pairs and greater number of N-T pairs with existing information about which grab samplings will be able to provide information.

The two algorithms presented in this paper make extensive use of the PCSI algorithm, which accounts for the problem of hydraulic uncertainty, to some extent, by including false positive and false negative rates of the sensor in past N-T pair probability estimates. Here, the performance of the algorithms was also evaluated by changing the false rates of the sensors. Two different false rates, 10% and 30%, were used and the general performance was similar for both false rates. The accuracy of the forecasted spread remained very similar and the sampling locations did not significantly change for the two false rates. However, the ANE rankings of the estimated sampling locations for the higher false rate dropped, particularly for the cases with
smaller number of sensors. With increasing number of sensors, the ANE rankings of the estimated sampling locations became more similar to the rankings for the lower false rate case.

The reason for this drop in ANE ranking is due to the parabolic nature of entropy. A higher false rate, as a whole, decreased the impact of information gain by confirmatory sampling. However, the impact of information gain about unknown N-T pairs was reduced more than the impact information gain about N-T pairs with existing information. This resulted in a lower ANE ranking for the nodes which predominantly provided new information about unknown N-T pairs. Notwithstanding the lower ANE rankings, the estimated sampling locations for the higher false rate cases were still the best locations to sample as they resulted to the greatest increases in N-T pair characterization by the PCSI algorithm and by the spread forecasting algorithm.
Chapter 6

Recommendations and Future Work

For the two networks analyzed, the developed algorithms successfully made forecasts and the best sampling locations were also correctly identified. However, as discussed in Chapter 4, improving information regarding contaminant spread can be achieved both by gaining new information about unknown N-T pairs and/or modifying information about N-T pairs with existing information. Although the presented sampling location selection algorithm tends to select sampling locations that predominantly provide new information about unknown N-T pairs, the opposite may also be desired. If so, the algorithms could be modified to give more importance to a certain type of information improvement.

*Spread Forecast.* The relatively simple approach adopted for the estimation of spread worked well enough to be able to estimate the best sampling locations. However, a more rigorous approach for the spread estimation could result in improved sample location identification and is worth exploring.

*Hydraulic Uncertainty.* The developed algorithms, the PCSI algorithm and the backtracking algorithm, all assume known hydraulics. Before practical implementa-
tion of the spread forecasting and sampling location identification algorithms, it is very important to test the performance of these algorithms under changing or unknown hydraulics to better approximate actual conditions. Such dynamic hydraulics can be simulated with a short-term demand forecasting algorithm.

Impacts of multiple samples. The presented sampling location selection algorithm was developed to maximize the benefit from a single sample obtained from a particular location. In reality, it might be more convenient, or practical, to obtain multiple samples from one location than from different locations. It is quite possible that the best location for obtaining a single sample be quite different from the best location to obtain a series of samples. Hence the algorithms could be modified to estimate the expected change in entropy (information) for a series of samples from a particular location.

A long term goal of this work is to develop a tool to aid the utility managers and decision makers in developing response action and formulating mitigation action plans during a contamination event. Interpretation of the data produced by this and other tools can be greatly expedited if they can be stored in a database coupled with a visualization software like GIS. Also, in the future, algorithms could be developed to provide decision makers with several automated response action plans, describing their pros and cons, so that decision makers would only need to choose one (or several) from the alternatives provided. Such an algorithm, if implemented, would minimize the cost of damage, allocate necessary resources to the most needed areas and, most importantly, save the amount of time required to make a decision.
7 References


