

**HETEROGENEOUS DATA AND PROBABILISTIC SYSTEM
MODEL ANALYSES FOR ENHANCED SITUATIONAL
AWARENESS AND RESILIENCE OF CRITICAL
INFRASTRUCTURE SYSTEMS**

A Dissertation
Presented to
The Academic Faculty

by

Cynthia Ting Lee

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in Civil Engineering

Georgia Institute of Technology
May, 2020

COPYRIGHT © 2020 BY CYNTHIA TING LEE

**HETEROGENEOUS DATA AND PROBABILISTIC SYSTEM
MODEL ANALYSES FOR ENHANCED SITUATIONAL
AWARENESS AND RESILIENCE OF CRITICAL
INFRASTRUCTURE SYSTEMS**

Approved by:

Dr. Iris Tien, Advisor
School of Civil and Environmental
Engineering
Georgia Institute of Technology

Dr. Laurence Jacobs
School of Civil and Environmental
Engineering
Georgia Institute of Technology

Dr. John Taylor
School of Civil and Environmental
Engineering
Georgia Institute of Technology

Dr. Erik A. Johnson
Viterbi School of Engineering
University of Southern California

Dr. Seymour Goodman
The Sam Nunn School of International
Affairs
Georgia Institute of Technology

Date Approved: March 4, 2020

ACKNOWLEDGEMENTS

Support for this work by the National Science Foundation through Grant No. CNS-1541074 is acknowledged. Support by the United States Department of Homeland Security under Grant Award Number, 2015-ST-061-CIRC01 is also acknowledged. The views and conclusions contained in this document are those of the author and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the U.S Department of Homeland Security.

I would like to express my gratitude to my advisor, Dr. Iris Tien, for her guidance and support throughout my graduate studies. Your encouragement and investment have been invaluable for my development as a researcher, and I have learned so much by your example. I would also like to thank the members of my committee for their input in this dissertation: Drs. Seymour Goodman, John Taylor, Erik Johnson, and Laurence Jacobs.

I am also grateful to the mentors that have given me new perspectives on the role of teachers in higher education: Drs. Chloé Arson, David Lawrence, Tammy McCoy, and Kate Williams. Thank you all for your time in meeting with me and your feedback on my teaching experiences.

Thank you to Dr. Larry Jacobs for your consistent and persistent affirmations when I lacked confidence in my work or abilities, and to Dr. David Torello for helping me navigate my own maze of thoughts. Thank you to my fellow research groupmates, past and present, Dr. Chloe Applegate, Dr. Ajay Saini, Yanjie Tong, Albert Zhang, Jorge-Mario Lopez, and Christopher Dixon. You have all contributed to this dissertation through your

advice, insights, and technical expertise, and I feel fortunate to have been able to work with each of you. To Meron Wolde-Tensae, Scotty Smith, Rebecca Nylen, Maya Goldman, Sam Corrado, and Erol Unal – you have all kept me afloat from our first year at Georgia Tech. You have seen me through academic and personal struggles and have been steadfast in your love and support through it all.

This dissertation would not have been possible without the guidance, input, and friendship of so many others. Whether through catch-up calls and meals, reunions abroad, or late-night conversations about confidence and purpose (and all other existential crises), I am so fortunate to be able to count on friends and colleagues from different chapters of my life, in Atlanta, Boston, the Bay Area, Medford, New York, and now so many other places scattered around the globe. You have all sustained me and uplifted me at my lowest, celebrated my achievements with me, and shared in times of joy and laughter. I am forever grateful.

Finally, I dedicate this dissertation to my family, who have been my inspiration and motivation for everything I do. To my parents and sisters – all of my accomplishments are the product of your unconditional love and support. You have taught me what is important in life. To my grandmother, Popo, you are one of the strongest people I know, and I hope to one day have a fraction of the courage that you have. And to my grandfather, Gonggong, I love and miss you, and I promise that I will always keep climbing higher.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS	iii
LIST OF TABLES	viii
LIST OF FIGURES	ix
LIST OF SYMBOLS AND ABBREVIATIONS	xii
SUMMARY	xiv
CHAPTER 1. INTRODUCTION	1
1.1 Motivation	1
1.2 Dissertation Outline	7
CHAPTER 2. BACKGROUND	9
2.1 Introduction	9
2.2 Data Availability and Heterogeneity for Monitoring Critical Infrastructures	9
2.3 Social Media as a Data Source	11
2.4 Data Integration for Heterogeneous Sources	16
2.5 Critical Infrastructure System Parameters and Analyses	17
2.6 A Bayesian Network Model for Critical Infrastructure Systems	20
CHAPTER 3. Framework for Processing and Classifying Social Media Data For Critical Infrastructure Monitoring	23
3.1 Introduction	23
3.2 Objectives and Data Collection	23
3.3 Proposed Framework	24
3.4 Step-by-Step Summary	27
CHAPTER 4. Framework for Data Integration of Heterogeneous Sources	30

4.1	Introduction	30
4.2	Variables and Definitions	30
4.3	Proposed Framework	31
4.4	Step-by-Step Summary	34
4.5	Sequential Updating	34
CHAPTER 5.	Application Examples of Proposed Frameworks	36
5.1	Introduction	36
5.2	Application of Proposed Text-Based Classification Framework for Building Supervised Learning Classifiers to Identify Infrastructure Damage	37
5.2.1	Twitter Data Collection and Database	38
5.2.2	Training and Testing Set Development	39
5.2.3	Classification and Evaluation	48
5.3	Application of Social Media Classifiers for Identifying Critical Infrastructure Damages During Tropical Storm Irma in Atlanta, Georgia	61
5.4	Application of Proposed Data Integration Framework for Flood Event Estimation in Louisiana, United States	64
5.4.1	Prior Probabilities and Data Source Identification	65
5.4.2	Data Collection and Indications of Flood Events	67
5.4.3	Likelihoods of Data from Each Source	74
5.4.4	Probabilities of Data, Integration of Data Likelihoods, and Final Updating	80
5.4.5	Results	81
5.5	Application of Proposed Data Integration Framework for Estimating Transportation Damage During Tropical Storm Irma in Atlanta, Georgia	90
CHAPTER 6.	Probabilistic System Model Analyses	96
6.1	Introduction	96

6.2	Atlanta Water Network, Varied Parameters, and Network Model	96
6.3	Component Conditional Probabilities of Failure	100
6.4	Redundant Power Supply Components	109
6.5	Additional Link Analysis	112
6.6	Vulnerability Assessment Across Multiple Networks	118
CHAPTER 7.	Conclusions	129
7.1	Introduction	129
7.2	Social Media Data Processing and Classification Framework and Applications	129
7.3	Heterogeneous Data Integration Framework and Applications	132
7.4	Probabilistic System Model Analyses	135
7.5	Future Work	138
7.5.1	Additional Machine Learning Applications	138
7.5.2	Data Integration within Sensor Sources	139
7.5.3	Impacts of Network Topology on the Resilience of Different Networks	139
REFERENCES		141

LIST OF TABLES

Table 1. Example tweets that are relevant to transportation and power system damages given definitions	41
Table 2. Example of tweets that are irrelevant to transportation and power system damages given definitions	41
Table 3. Example keywords/terms for searching database	43
Table 4. List of filter terms for training and testing set development	47
Table 5. Final training and testing set sizes	48
Table 6. Results of models in classifying training data	49
Table 7. Results of classification evaluated using 10-fold cross-validation	50
Table 8. Results of models in classifying testing data	52
Table 9. Examples of relevant and irrelevant tweets for damages in transportation systems	55
Table 10. Examples of relevant and irrelevant tweets for damages in power systems	56
Table 11. Confusion matrix for support vector machine classifier tested for tweet relevance to flood events	72
Table 12. Example of tweet classified as relevant and containing location indicators for Louisiana	74
Table 13. Estimated disruptions of service for Supply 4 failure scenario	118
Table 14. Descriptions of multiple small-size water networks for vulnerability comparisons	120

LIST OF FIGURES

Figure 1. Flowchart for development of text-based classifiers for CIS monitoring	29
Figure 2. Effect of training size on recall for relevance	59
Figure 3. Effect of training size on precision for relevance	60
Figure 4. Example flood risk map with 100-year floodplain shaded for Ascension Parish, Louisiana (LSUAgCenter and LADOTD 2017)	66
Figure 5. Example gage height data for a stream gage in Louisiana (Data from USGS 2017)	68
Figure 6. Stream gages in Louisiana indicating a flood, August 2016	70
Figure 7. Decaying probability function with distance for stream gage sensor likelihood	75
Figure 8. Resulting updated probability distributions for August 10, 11, 12, and 13	82
Figure 9. Updating and sequential updating results with sensor observations from August 11 to August 11-12 and August 12 to August 12-13	83
Figure 10. Resulting updating probability distributions for August 12, 2016, using gage height and tweet relevance data sources alone compared to estimation from integrating both data sources	84
Figure 11. Louisiana Disaster Declaration for flooding event in August 2016, FEMA-4277-DR (Reprinted from FEMA 2016)	86
Figure 12. Parishes with updated probabilities of flood risk (dark blue) from integrated data sources on the dates shown	86
Figure 13. Stream gage and Twitter data availability by parish from August 10-13, 2016	88
Figure 14. Precipitation in Louisiana from August 11-14, 2016 (left) and inundation map for Louisiana (right) (Reprinted from Watson et al. 2017)	89
Figure 15. Map of sequential updating from August 12-13, 2016, with regions of highest precipitation (left) and confirmed inundation (right) from Figure 14 highlighted	89

Figure 16. Roadways needed to access a water supply (Supply 4) for repair in Atlanta, Georgia	92
Figure 17. Example gage height data for a stream gage in Atlanta, Georgia (Data from USGS 2017)	94
Figure 18. USGS stream gage and transportation component shown	94
Figure 19. Example tweet related to transportation damage in Atlanta, Georgia during Tropical Storm Irma	95
Figure 20. Schematic of Atlanta's water distribution system comprising supply and distribution components	99
Figure 21. Representative Bayesian network model of interdependent CIS	99
Figure 22. Relative change in marginal probabilities of failure versus number of supplies, varying conditional probabilities of failure of power supplies	102
Figure 23. Relative change in marginal probabilities of failure versus shortest distance to supply for components with one reachable supply	103
Figure 24. Relative change in marginal probabilities of failure versus number of MLSs for components with one supply	103
Figure 25. Relative change in marginal probabilities of failure versus minimum number of links from components to supply	104
Figure 26. Exponential trendlines for each conditional failure probability for power supplies, for relative changes in component outcomes as a function of shortest distance to a supply	107
Figure 27. α -coefficient values of exponential fit functions versus varied network parameter	108
Figure 28. Relative changes in marginal probabilities of failures after adding redundancies at each water supply versus (a) number of supplies, (b) shortest physical distance to a supply, (c) number of MLSs, and (d) minimum number of links to any supply	111
Figure 29. Failure scenario where failure of Supply 4 leads to failure of two distribution nodes	113

Figure 30. (a)-(d) from top left, right, bottom left, bottom right – New links added to the network and inference results for Supply 4 failure	114
Figure 31. U.S. Census blocks associated with Supply 4 and two distribution components that fail in Supply 4 failure scenario	116
Figure 32. Graph metrics versus network size	122
Figure 33. Average failure probability of network components versus network density	123
Figure 34. Average failure probability of network components versus average node degree	124
Figure 35. Average failure probability of network components versus average number of reachable supplies per component	125
Figure 36. Average failure probability of network components versus fraction of one-supply components in each network	127
Figure 37. Average failure probability of network components versus average supply-to-distribution component path lengths in each network	128

LIST OF SYMBOLS AND ABBREVIATIONS

θ	Probability of event occurrence
λ	True class of a tweet for flood event prediction
AL	Average path length
BN	Bayesian network
CIS	Critical infrastructure systems
CPT	Conditional probability table
D	Graph diameter
d_{ij}	Length of an MLS path between components i and j
FEMA	Federal Emergency Management Agency
FN	False negatives
FP	False positives
m	Number of edges (Chapter 6)
MLS	Minimum link set
n	Number of nodes (Chapter 6)
P	Precision
pf_{old}	Original component marginal probability of failure

pf_{new}	Component marginal probability of failure after changing network parameters
q	Graph density
R	Recall or true positive rate
RC	Relative change in marginal probabilities of component failure
s	Sensor source (Chapters 4 and 5); number of supplies (Chapter 6)
TP	True positives
USGS	United States Geological Survey
y	Stream gage height
z	Tweet class predicted by a classification model

SUMMARY

The protection and resilience of critical infrastructure systems (CIS) are essential for public safety in daily operations and times of crisis and for community preparedness to hazard events. Situational awareness of CIS is key for prioritization and decision-making regarding repairs, maintenance, and emergency response. Increasing situational awareness and resilience of CIS includes both comprehensive monitoring of CIS and their surroundings, as well as evaluating CIS behaviors in changing conditions and with different system configurations. This is a significant challenge given that CIS are distributed over large geographic areas, often lack continuously available data for monitoring system states, and have complex interdependencies between system components and across different systems. This work addresses these factors to enhance CIS resilience.

First, two frameworks for increasing the monitoring capabilities of CIS and their surrounding communities are presented. The proposed frameworks are (1) a process for analyzing and classifying social media big data for monitoring CIS and hazard events that impact CIS and (2) a framework for integrating heterogeneous data sources, including social media data, using Bayesian inference to update prior probabilities of event occurrence as new information is collected. Applications of both frameworks are presented, including building and evaluating text-based machine learning classifiers for identifying CIS damages and integrating stream gage data from the United States Geological Survey and Twitter data to estimate flood risks in Louisiana and CIS component damage (i.e., roadway flooding) in Atlanta, Georgia.

Finally, probabilistic analyses of CIS vulnerabilities with varying system parameters and topologies are presented. In a single water network, the impact of varying parameters on component performance throughout the system is evaluated. In multiple, small-size water networks, the impacts of system topology and supply-to-distribution component connectivity are assessed to identify characteristics of more resilient networks. Each of these analyses is conducted with a Bayesian network model that represents interdependent CIS.

This body of work contributes insights and methods for monitoring CIS and their surroundings and for assessing CIS performance. Specifically, integrating disparate data sources, including nontraditional sources that account for public perceptions, increases situational awareness of CIS, especially during or after failure events, and evaluating the sensitivity of CIS outcomes to changes in the network facilitates decisions for infrastructure planning and investments and for emergency response.

CHAPTER 1. INTRODUCTION

1.1 Motivation

Critical infrastructure systems (CIS) are the lifelines of modern society, providing services and commodities such as transportation, communication, water, and power to communities for daily functions. The resilience of these systems is necessary to ensure public health and safety, societal efficiency, and economic growth. The United States federal government recognizes 16 critical infrastructure sectors, including transportation, water and wastewater, energy, emergency services, and communications and underscores the need to take a holistic approach in securing these complex systems (White House 2013, PPD-21). This Presidential Directive calls for the strengthening of CIS resilience, so that CIS are “able to withstand and recover rapidly from all hazards,” which will “require integration with the national preparedness system across prevention, protection, mitigation, response, and recovery.” This is a monumental task further compounded by aging and evolving systems and dynamically changing environments. Civil infrastructures in particular are deteriorating from lack of maintenance and financial investment (ASCE 2016). Many components are at or past the end of their design lives, yet they are expected to accommodate rapidly growing populations, uncertain effects from climate change, and increasing interdependencies across systems with continually advancing technologies (Ouyang 2014; Saini and Tien 2018).

Moreover, CIS face risks from a range of hazards that may occur in quick succession, impeding system and community recovery and resilience efforts. Hurricane Maria in 2017 wreaked havoc on CIS across Puerto Rico, causing widespread power

outages and damaged roadways, and a series of earthquake events in January 2020 created still more damage for communities that had not fully recovered from the hurricane event (Brakkton 2020). Hurricanes Harvey and Irma in 2017 also caused CIS damages and flooding across large portions of the southeastern United States (FEMA 2018). In the Tohoku 2011 earthquake and subsequent tsunami, buildings collapsed, portions of roadway and railway networks were destroyed, power and water supplies experienced long-term interruptions, and interdependencies between systems caused a nuclear disaster (Krishnamurthy et al. 2016; Lekkas et al. 2012). Meteorological hazards such as hurricanes have been increasing in intensity and frequency in recent years and have severe impacts on CIS and their surrounding communities (Coumou and Rahmstorf 2012), and these changing conditions must be considered in conjunction with other natural and manmade hazards that impact CIS.

Terrorist attacks and manmade hazards can have similar effects on CIS as from natural hazards. For instance, in 2017, an accidental fire caused an otherwise structurally sound section of Interstate 85 in Atlanta, Georgia to collapse. Approximately 220,000 vehicles a day depend on the collapsed section that was closed for six weeks (Twumasi-Boakye and Sobanjo 2018). Overall, damaging events affecting CIS can vary greatly in terms of predictability, intensity, and resulting consequences, and all potential impacts must be considered in CIS protection and resilience.

Optimal situational awareness of CIS is key for prioritization and decision-making regarding CIS repairs, maintenance, and emergency management, and enhancing situational awareness of CIS increases their resilience (i.e., their ability to provide continuous services and adapt in the wake of long-term stressors and short-term shocks).

A community's situational awareness is its ability to collect and synthesize available information to gain a holistic understanding of its surroundings (Ireson 2009; Vieweg et al. 2008). For improving the resilience of interdependent CIS, this means understanding the current state of system components and their interactions within and across systems during normal operations and any hazard events. Obtaining situational awareness includes comprehensive monitoring of systems and environmental conditions, as well as understanding system interdependencies and behaviors in a variety of performance scenarios.

There are many challenges and limitations in enhancing the situational awareness and ultimately, resilience of these complex and continually evolving systems. The work in this dissertation focuses on addressing two specific challenges: (1) the lack of data availability for monitoring CIS in real-time and (2) the need to evaluate the impacts of network parameters and connectivity on overall vulnerability, a key facet of CIS resilience.

Full, continuous monitoring of system states and interactions is often not available because CIS are highly distributed over large geographic areas. While physical sensors for monitoring the natural and built environments have increased capabilities as their technologies advance, they lack widespread deployment in all communities and where deployed, are often homogeneous and measure only a single condition or parameter (Rawat et al. 2014). This lack of information and data integration across varying conditions hinders comprehensive, real-time analyses of CIS and their interdependencies (Rinaldi 2004; Eusgeld et al. 2011).

At the same time, the rise of the web and social media has unlocked opportunities for gathering large amounts of information, observations, and opinions from the crowd in real time. Recent research has investigated the use of these nontraditional data sources to detect disaster events and CIS damages, e.g., Tien et al. (2016), Musaev et al. (2015), Vieweg et al. (2010), and Yin et al. (2012). Big data from social media is inherently noisy and unreliable, so discerning useful information from it can be difficult, but it has potential to supplement information from traditional data sources. To use heterogeneous data sources for monitoring CIS and their surroundings, the unique challenges associated with data analysis for each source must be considered.

This dissertation proposes two frameworks for using heterogeneous data sources, including nontraditional sources such as social media data, for monitoring CIS and their surroundings. The first is a framework for processing and classifying social media big data to identify CIS damages in real-time. As an application, the framework is used to build and evaluate three well-known machine learning classifiers for identifying transportation and power system damages in text-based Twitter data. The three classifiers compared are support vector machines, naïve Bayes models, and decision trees. In addition to evaluating the classifiers on manually labeled test data, the highest performing classifier is evaluated on a set of unlabeled data from a specific event (i.e., Tropical Storm Irma in Atlanta) to simulate a real-time detection scenario.

The second framework is for probabilistically integrating data from multiple and varied sensor types to estimate disaster and failure events. Such integration can improve situational awareness by providing more comprehensive and up-to-date evaluations of a community's surroundings than an individual sensor type alone. The framework uses

Bayesian updating to infer updated probabilities of event occurrence from assumed or computed prior probabilities. The result is a posterior event probability given collected data. Bayesian updating is a method for data fusion across multiple sensors, especially in the field of robotics (Durrant-Whyte and Henderson 2008). The novelty of the proposed approach is in its considerations and updating of prior understandings of risk, and in integrating likelihoods of observing datapoints within individual sensor networks first before integrating that information across unrelated sensor types to monitor a community condition or event. This approach enables the framework to account for uncertainties in system states and be applicable across events. It captures information specific to each data type, calculating specific sensor source likelihoods to account for data heterogeneity, before combining information from all sources. The framework supports real-time and dynamic updating of estimated risks of an event as different data is observed and more data becomes available. To explicate the framework's use, it is applied to estimate flood risks in Louisiana during August 2016, and results are assessed compared to the true flood events that occurred during the time period of analysis. The framework is also applied to estimate the risk of transportation damage (i.e., blocked roadways) during Tropical Storm Irma in Atlanta, Georgia. This application is provided to illustrate the framework's potential for updating inputs for CIS models.

In conjunction with increasing monitoring capabilities, it is important to understand the impacts of CIS network parameters and topology on system vulnerabilities and behaviors. Network parameters and topology change when components are damaged in hazard events or when actions are taken to strengthen and improve CIS in attempts to increase resilience. Quantifying the impacts of different network characteristics on CIS

behaviors enables the evaluation and prioritization of various actions to repair, retrofit, or build out new parts of an infrastructure system. CIS characteristics include component-specific attributes such as capacity and asset robustness and system-level attributes such as network topology and connectivity. Given the complexities of CIS, including their interdependencies and the range of potential changes to network parameters that can impact system behavior, there is the need to determine the sensitivity of system performance to those different parameters. That is, what changes to network parameters have the greatest impact on system behavior and overall resilience?

This portion of the dissertation investigates the effects varying of three network parameters on CIS vulnerability and resilience: component probabilities of failure, supply redundancies, and system connectivity. The parameters are varied in a single system, and each of these correspond with a decision a stakeholder might make regarding infrastructure investments to increase resilience, i.e., retrofitting components to better withstand a hazard, adding supply redundancies to the system, and building out new parts of the system. The effects of weakened system components, i.e., damaged components with increased probabilities of failure, are also evaluated. Outcomes are evaluated by comparing component state functionality under these varying conditions across component characteristics and populations affected in failure scenarios. Moreover, since CIS can have a wide variety of topologies that affect the paths through which supply components can reach non-supply components in a network, a series of analyses is conducted across multiple, small-size water networks. The overall vulnerabilities of the systems are compared according to their supply-to-distribution component connectivity and other graph theory metrics. To obtain the probabilistic outcomes of system performance under

varying parameters and across multiple system layouts, a Bayesian network (BN) model is used.

1.2 Dissertation Outline

The remainder of this dissertation is organized as follows:

Chapter 2 presents a background of the current body of work for areas related to the use of nontraditional data sources for CIS monitoring, data integration methodologies, and analyses of CIS in changing conditions. Specifically, related work on the use of social media big data for monitoring critical infrastructures and hazard events in conjunction with traditional data sources is presented, along with related work in evaluating CIS performance and resilience. This chapter also describes in detail the Bayesian network model used in this dissertation for probabilistic analyses of CIS performance.

Chapter 3 describes a new framework for processing and classifying social media big data for use in monitoring CIS components, from initial download to the development of supervised-learning models to identify specific, real-time CIS damages. The proposed framework focuses on text-based social media data.

Chapter 4 describes a new framework for probabilistically integrating heterogeneous data sources to monitor a single condition or event. Heterogeneous data sources include both physical and social sensor data.

Chapter 5 presents application examples for both of the proposed frameworks. **First**, three text-based machine learning classifiers are built and evaluated for identifying CIS damages in two systems through the proposed framework for processing social media

data. Specifically, the classifiers are built to identify damages in transportation and power systems. A detailed comparison and analysis of the classifiers' performance and generalizability is presented, including an evaluation of one of the classifiers on unlabeled data to simulate a real-time detection scenario. **Second**, the proposed data integration framework is used to estimate flood risk in Louisiana in August 2016 using two disparate data sources, including a social media sensor and a physical sensor. **Third**, the proposed data integration framework is used to estimate the risk of damage to a transportation system component (i.e., a blocked roadway) during Tropical Storm Irma in Atlanta, Georgia.

Chapter 6 analyzes the impacts of the parameters and topology of a single CIS with interdependencies across other system components and the impacts of topology on overall network vulnerability in multiple, small-size water networks. The Bayesian network model described in Chapter 2 is used for each of these analyses.

Chapter 7 discusses opportunities for extending and continuing the work in this dissertation and summarizes the major findings and contributions of the work and results presented.

CHAPTER 2. BACKGROUND

2.1 Introduction

This chapter presents a background of related work in monitoring CIS components and their surroundings and evaluating CIS network parameters and their impact on the overall behaviors of CIS. Areas where previous work can be built upon and new work is needed for increasing situational awareness and resilience of interdependent CIS is discussed. For monitoring CIS and their surroundings, different data sources, including the emergence of social media as a nontraditional data source, and current data integration methods are discussed. For evaluating CIS network parameters, current analyses of CIS performance with uncertain parameters, various topologies, and actions implemented to increase resilience are presented and discussed.

2.2 Data Availability and Heterogeneity for Monitoring Critical Infrastructures

Increasing the resilience of CIS is a global challenge (Amin 2002), with infrastructure performance recognized as a critical contributor to overall community resilience (Johansen et al. 2017). Monitoring CIS components in day-to-day operations and in times of crisis is important to obtain situational awareness for the sustainability, protection, and response of CIS and the communities they serve in today's rapidly changing world. Elements of CIS and their surroundings – both of the natural and built environments – can be difficult to estimate given uncertainties in hazards, responses, and impacts, and with continuous evolution in time.

A variety of tools are available, depending on the type of infrastructure, to monitor CIS and detect damages. Within these tools are many different physical sensors available

for monitoring conditions in the built and natural environments. For example, strain gages and accelerometers provide data to detect anomalies in structural health monitoring applications (Lynch and Loh 2006), and air quality and meteorological sensors collect data on atmospheric conditions. For power systems, smart grids detect and isolate power outages (U.S. Department of Energy 2019). These physical sensors can provide real-time data for system analyses. They do not, however, identify sources of damage, assess damages to other components and systems caused by the same source, or measure the impact of infrastructure damages on local communities.

Smart city initiatives are pushing forward the need for open data availability and more connected wireless sensor networks, e.g., Chicago's Array of Things (Mone 2015), to increase resilience, but new information and communication technologies are slow to implement uniformly across entire communities (Neirotti et al. 2014; Alawadhi et al. 2012). Furthermore, most interconnected sensor networks still lack heterogeneity to collect multiple types of data and require single or similar sensor types (Miorandi et al. 2012).

Large-scale deployment of physical sensor networks must also consider sensor size, cost, and configuration, and may be slow or impractical to implement depending on the application, limiting the availability of physical sensor data within a community (Rawat et al. 2014). Physical sensor networks for structural health monitoring are often designed to monitor single civil structures, as seen in studies such as Hackmann et al. (2014) and Pakzad et al. (2008), rather than groups of structures over larger distributed areas. Single physical sensor types alone are thus limited in the information they can reveal about community conditions affecting infrastructure and people.

Other methods for infrastructure damage assessment include image-based approaches such as satellite imagery processing (Chen and Hutchinson 2007), laser image processing (Tsai and Chatterjee 2018), and image-based three-dimensional reconstruction (Zhou et al. 2015). Despite new technologies and initiatives, CIS still lack comprehensive, integrated, and continuous monitoring and damage detection across systems.

2.3 Social Media as a Data Source

Social media users offer a new type of sensor – social sensors – which takes advantage of the constant information stream present today. Social sensors can fill in gaps of information about infrastructure or environmental conditions where physical sensors are not deployed or do not exist. Social media data is valuable for gathering large amounts of real-time information from the public for event detection, situational evaluation, or sentiment analysis and can enhance situational awareness in disaster events (Kankanamge et al. 2019).

Research has been conducted using social sensors to examine community responses to manmade events such as mass shootings (Vieweg et al. 2008) and to detect critical infrastructure failures (Tien et al. 2016) and natural disasters such as earthquakes and landslides (Sakaki et al. 2010; Musaev et al. 2015). Other works have also investigated the use of social media for monitoring environmental conditions (Boddula et al. 2016), monitoring urban environments and populations (Klotz et al. 2017), and mapping human mobility (Wang and Taylor 2015). These studies have shown that, in addition to providing information on conditions of the built and natural environments, social media can provide

data that physical sensors cannot measure, such as information about societal conditions or public perception regarding major events.

Social sensor data collection and analysis present their own set of challenges. Data from social sensors is noisy and unreliable, containing rumors or misinformation that are both intentional and unintentional (Alexander 2014; Zhang et al. 2019). The majority of users do not share their locations (Leetaru et al. 2013). For instance, less than 1% of data streamed from one social media platform, Twitter, contains geotagged information (Martin et al. 2020). This makes it difficult to determine event locations and whether or not the user is truly observing the event of interest and reduces the amount of useful information that can be extracted from social sensors. Moreover, for social sensors to provide significant new information, specific topics need to garner enough attention from the crowd. While many physical sensors can provide continuous or near-continuous data, social sensors may vary in the amount of information provided on a topic at any given time.

In this dissertation, the focus is on identifying information relevant to infrastructure damage in text-based social media data. The benefit of using text-based social media data instead of image data is that large quantities of the former are available in real-time, including as potential hazard events occur. While most social media platforms also allow multimedia attachments in user posts, the focus on text-based data in this dissertation is to easily include a broad range of damage detection for multiple systems and to support extensibility to other data sources.

Next, previous approaches to categorize the texts of social media posts for event detection, and more specifically, CIS damage identification are described. These include

methods for filtering posts for keywords or hashtags related to disaster and damage events, crowdsourcing to label and categorize posts, and a variety of supervised and unsupervised machine learning models. To improve existing models and approaches for extracting information about CIS from social media data, there is the need to identify specific damages, rather than generally assessing damages in a disaster. The approach described in this dissertation identifies current and ongoing damages according to definitions that are not specific to a single disaster event. In this dissertation, the generalizability of specific classification models is also evaluated in a real-time monitoring scenario, which is also absent from analyses of previous approaches (Zhang et al. 2019).

The following studies identify hazard events and CIS damages through filtering and manual labeling methods. De Albuquerque et al. (2015) filter tweets based on keywords, manually inspecting and labeling tweets as relevant or irrelevant to flood events, then categorize relevant tweets into subgroups, e.g., tweets relating to traffic conditions or infrastructure damage during a flood event. This manual classification, however, is inefficient and cannot be used to increase situational awareness because large amounts of relevant information cannot be identified in real-time. Smith et al. (2015) also use Twitter data to support flood risk management and select tweets based on a predetermined set of fifty-five key terms related to flood events. Filtering posts in this way does not consider other words or terms in a tweet that can change or specify its context and can disregard relevant tweets or include irrelevant tweets without further evaluation.

Tien et al. (2016) search Twitter data for posts specifically related to damages to bridges, highways, power systems, and gas lines. Tweets are collected with a keyword and stop word filtering process. However, the accuracy of the filtering is not evaluated. Yuan

and Liu (2018) determine the relevance of tweets to a specific disaster event to predict the likelihood of infrastructure damage. The tweets are again selected based on keywords pertaining to the specific hazard event investigated, and the method does not identify specific events from the tweets.

The following studies use machine learning classification models to predict hazard and damage events. This dissertation builds on these works to identify CIS damages regardless of the hazard event that caused them. Imran et al. (2013) use crowdsourcing tools to categorize tweets by whether or not they are informative, what type of information they provide, and the subtype of information they provide. While identifying types of tweets that are helpful during disaster events, the study does not show how well the model performs in classifying new sets of tweets not filtered by crowd workers prior to classification (i.e., in real-time monitoring).

Yin et al. (2012) use machine learning classifiers to identify information about critical infrastructures to improve emergency response. The classification is binary for whether or not a tweet contains information about disaster impact on infrastructures. While Yin et al. (2012) demonstrate the success of machine learning classifiers to detect infrastructure damages during an emergency event, the classifiers are trained and evaluated on a single dataset from the same event, the 2011 Christchurch earthquake and does not address generalizability. Similarly, Yuan and Liu (2020) use machine learning classifiers to analyze social media for damage assessment after Hurricane Matthew in 2016. Different models, including those presented in this dissertation (i.e., Naïve Bayes, support vector machine, and decision tree models), are compared, but the models are again trained and

evaluated for a specific hazard event. While a sentiment analysis of collected data is included in the results, the models do not identify damages for a specific CIS.

Musaev et al. (2015) use classifiers to predict the relevance or irrelevance of social media posts to landslides. To conduct real-time detection, posts referring to past or future time periods, i.e., those that contain years or months that do not match the evaluation period, are removed. The remaining posts, however, may still refer to noncurrent landslides. To increase situational awareness of CIS, the classification in this dissertation identifies posts referring to current or ongoing damages labeled as relevant depending on the presence of temporal words that provide context, e.g. “today,” “now,” “yesterday,” etc.

Pereira et al. (2017) use machine learning-based classifiers to identify transportation-related tweets. The classifiers are trained on manually labeled tweets and tested on a separate, random set of tweets that are not collected during any particular hazard or event. Non-travel-related tweets in the training set are selected randomly and then manually verified. Compared to the work in this dissertation, the classifiers do not identify specific CIS damages to support decisions to increase community resilience.

Finally, the following approaches use integrated machine learning methods for mapping disaster-related hazard and damage events identified from social media data. Fan, Jiang, and Mostafavi (2020) identify critical tweets related to CIS disruptions with a graph-based approach and identified tweets are later combined with image and geographical data. The graph-based approach represents tweets as nodes and semantic similarities between the tweets as edges in a weighted graph. This approach identifies clusters of tweets related to CIS disruptions, but does not identify damages to a specific system. Fan, Wu, and

Mostafavi (2020) propose a “pipeline” for mapping events in social media data that also includes supervised and unsupervised learning algorithms. The classification models identify eight different classes related to disaster events. However, they do not focus on damages for a specific CIS. Compared to these studies, the approach proposed in this dissertation also integrates different methods for extracting useful information from social media data but focuses on identifying damage in specific CIS, rather than a general damage assessment. Moreover, the generalizability of models built in this dissertation are evaluated in a specific disaster event, but the approach is not specific to a single event and is tested on an event not part of the training dataset.

2.4 Data Integration for Heterogeneous Sources

Due to the limitations and challenges in using both physical and social sensor sources for monitoring CIS and their surroundings, data from individual sensor networks alone may not truly describe the causes or impacts of a damaging event on a community. In this dissertation, the ability of data integrated from multiple sensor sources to enhance resulting analyses, event detection, and overall situational awareness is investigated. Recent studies show potential in integrating different sensor sources to detect hazard events. For example, Musaev et al. (2015) maps social media posts and data from physical sensor sources, including rainfall and earthquake data, to calculate the probability of detecting a landslide given a grid-based location. While this study and others, e.g., Jongman et al. (2015), use data from across sensor types, they do not provide a framework through which heterogeneous likelihoods of sensor observed data can be integrated to calculate a posterior probability of an event given that data. In addition, the data integration framework in this dissertation considers prior probabilities of event occurrence to account

for uncertainty in the disaster or failure events themselves. This enables the assessment of changing evaluations of risk as more data from different sources is collected

2.5 Critical Infrastructure System Parameters and Analyses

In addition to improving monitoring capabilities, another aspect of ensuring CIS resilience is assessing the impact of network parameters, including topology, system redundancies, and interdependencies across different systems. In this study, assessments are conducted for parameter variations in a single network and for multiple networks with varying topologies and redundancies.

There is a large body of work focusing on the resilience assessment of infrastructure systems. For example, Reed et al. (2009) characterize the behavior of CIS under natural hazards, including resilience and interdependency measures. The focus is on electrical power and telecommunications systems. Attoh-Okine et al. (2009) develop a resilience index, focusing on an urban highway infrastructure system. Pant et al. (2014) focus on evaluating economic resilience of interdependent infrastructure, while Poljansek et al. (2012) and Guidotti et al. (2016) specifically investigate the role of system dependencies and interdependencies on infrastructure risk and resilience. Rather than assessing infrastructure resilience, however, the focus of the work in this dissertation is to evaluate and compare varying approaches to increase CIS resilience. The objective is not to provide particular assessments of CIS resilience. Instead, this study seeks to investigate what measures and actions can be taken to increase infrastructure resilience, with a focus on evaluating the impacts of varying network parameters on CIS resilience.

CIS have a wide range of network topologies and component parameters depending on the types of infrastructure, surrounding environments, and communities they serve. Both system- and component-level network parameters affect CIS behaviors and resulting resilience. System-level parameters include connectivity between components and system redundancies. Component-level parameters include component capacities and probabilities of failure. Previous studies looking at the relationship between network parameters and resilience include Zhang et al. (2015), which looks at the effect of system topology measures on the resilience of transportation networks. The results from that study provide insights on the types of topologies that may be more resilient. However, the connectivity between specific components in the system (e.g., a critical node's reachability to other components in the network) is not studied, nor is the effect of specific improvements to an infrastructure system.

In this study, the objective is to evaluate and compare the impacts of potential changes in a CIS on overall resilience and to investigate the effect of network topology, especially in regard to supply connectivity, on overall system vulnerability. In addition, while Zhang et al. (2015) focuses on topological measures for a single transportation system, this study includes a comparison of topological measures across multiple networks and an investigation of the impacts of component-level improvements and adding redundancies in multiple interdependent networks.

Genge et al. (2012) focus on communication and control logic implementation parameters that influence the outcome of attacks on industrial control systems. The study connects communication and control systems to look at the impact of network parameters on the effectiveness of potential cyberattacks. Rather than looking at control logic-based

parameters as in Genge et al. (2012), the focus in this study is on physical CIS parameters that influence system outcomes. The parameters investigated in this study include physical improvements to CIS, specifically repairing or retrofitting components and building new parts of the system such as additional backups or links. In addition, in Genge et al. (2012), a simulator is used and results aggregated across runs of the simulation for each set of parameters. In contrast, the analyses presented in this study are based on a series of probabilistic inferences conducted over a network, resulting in a probabilistic assessment of CIS behaviors (i.e., component vulnerabilities).

Focusing on physical improvements to CIS, previous work includes Ouyang et al. (2012), which quantifies the impacts of improvements at different stages on resilience. The effect of varying ordering of actions is investigated. In contrast to that work, the scenario of interest in this dissertation is a decision that is to be made at a point in time, where the need is to compare and evaluate the effectiveness of varying actions to increase resilience. The objective is to provide information that facilitates prioritization of infrastructure resources for a given action based on the quantified impacts on CIS outcomes. In addition, the focus in Ouyang et al. (2012) is on a single infrastructure system of the power transmission grid, while this study includes interdependencies across multiple CIS.

Panteli and Pierluigi (2017) also focus on a single infrastructure system, electrical power, to evaluate the resilience of electrical power infrastructure to extreme weather events. The effects of constructing more robust lines and towers and adding parallel lines are considered. The study measures impacts by overall network loss of load frequency and loss of load expectation. Compared to that study, the work in this dissertation measures impacts at the individual distribution component level, providing a more granular

assessment of impacts. The granular approach in this study provides a detailed measure of the impacts of varying network parameters and leads to insights on predicting component-level impacts based on specific component characteristics. In addition, rather than the four specific case studies considered in Panteli and Pierluigi (2017), this study compares system performance across a range of component states and parameter values. The analyses in Panteli and Pierluigi (2017) are also specific to a high wind load hazard, while this study looks at how parameter variations in interdependent CIS affect performance regardless of hazard type.

Interdependent CIS are considered in Duenas-Osorio et al. (2006), which includes investigation of the effect of mitigation actions on increasing functionality. The study considers the addition of bypasses or alternative routes around congested nodes, with actions selected based on congestion in flow to create new edges. This action most closely relates to the link configuration parameter investigated in this study. However, the actions investigated in this study are based on failure scenarios rather than congestion. In addition, effects in Duenas-Osorio et al. (2006) are measured by overall network connectivity loss. In contrast, this study evaluates network impacts based on the effects at individual distribution components. The component-level approach in this study provides analysis results in terms of specific service location disruptions under varying scenarios. In addition, the evaluation of impacts as a function of different component characteristic enables future performance-based system designs that seek to target specific levels of performance for particular components.

2.6 A Bayesian Network Model for Critical Infrastructure Systems

To assess the impacts of network parameter changes and evaluate the performance of CIS under varying scenarios, inferences in this study are conducted over a Bayesian network (BN)-based model of CIS. A BN is an acyclic directed graph in which nodes represent random variables and edges represent the probabilistic dependencies between nodes (Jensen and Nielsen, 2007). The nodes and edges of a BN are defined by conditional probability tables (CPTs), which give the probabilities of nodes being in varying states conditioned on the states of their parents. The CIS BN modeling approach is suitable for the work in this dissertation to evaluate the impacts of varying network parameters and different network connectivities on CIS performance because many successive inferences can easily be conducted over a range of parameters and parameter values. In the BN, CIS components are nodes, and connections between components within a CIS and interdependencies across CIS are edges. A binary representation of CIS, where nodes are in survival or failure states, is considered. The CPT of each component contains the probabilities of failure of the component given the states of its parents, including the occurrence of a hazard.

Three comprehensive types of interdependencies are included in the model. Service provision interdependencies occur when a CIS component depends on the outputs of another component in a different system to operate. These are modeled as direct dependencies in the BN. Geographic interdependencies occur when components are collocated and thus subject to the same hazard risks. These are modeled with common hazard node parents across components. Access for repair interdependencies occur when a damaged component depends on access from another component for repair operations. These are modeled by tracking component states over time and defining dependencies on

access nodes in the BN. This last type of interdependency occurs in post-disaster situations when a component failure has been identified, and its inclusion in the model allows for the analysis of recovery efforts and downtimes (Johansen and Tien 2018). The model distinguishes the functionality of two types of components in the main CIS: supply components (e.g., water treatment facilities) and distribution components. Supplies provide the infrastructure resource to distribution components through minimum link sets (MLSs), which are minimum sets of components that link from a supply to a distribution component and will disconnect if any link is removed. A distribution component is defined as any non-supply component. With the nodes defined, the interdependent CIS BN model is constructed and exact inferences are conducted over the network using the algorithms described in Applegate and Tien (2019).

To assess the impacts of varying network parameters on CIS outcomes in this study, inferences are run for each value in the range of parameter values of interest. The inferences are easily conducted within the BN framework. For the network parameters evaluated, component probabilities of failure are efficiently varied at the nodes, system redundancies are added as additional nodes in the model, and changes in network connectivity and structure are made through altering the edges of the model. These variations are described in more detail in Chapter 6. For general CIS, component locations and additional dependencies can be altered as well. The inferences result in component marginal failure probabilities or conditional failure probabilities based on evidence, such as observed states of component or hazard nodes representing failure or hazard scenarios of interest.

CHAPTER 3. FRAMEWORK FOR PROCESSING AND CLASSIFYING SOCIAL MEDIA DATA FOR CRITICAL INFRASTRUCTURE MONITORING

3.1 Introduction

This chapter describes the proposed framework for processing and classifying social media big data for CIS monitoring and enhanced situational awareness. Including social media data in obtaining situational awareness for CIS provides information regarding public perception, which is essential for overall community resilience. An application of this framework is presented in Chapter 5. The application is to build and evaluate different supervised learning classifiers to identify transportation and power system damages. To further evaluate the generalizability of a classifier built in the application, a second example and analysis is provided that simulates use of the classifier for a real-time CIS damage detection during a hazard event.

3.2 Objectives and Data Collection

This study aims to address several major challenges in efficiently and accurately identifying relevant information in social media text data to assess the state of critical infrastructures for increased situational awareness. The objective is to provide guidance for building text-based machine learning classifiers to increase the availability of data for monitoring CIS and their surroundings. Applications of the framework result in (1) training and testing sets of labeled social media text data, (2) machine learning-based classifiers that automate the identification of current or ongoing damages for different infrastructure

systems in social media posts, and (3) evaluations of those classifiers' performance and generalizability.

To begin the framework, the user must have a database of texts downloaded or streamed from social media platforms. These can include tweets from Twitter, Instagram captions, status updates from Facebook, and more. Each individual post is considered a datapoint. This study focuses on classification based on the words in each post alone; however, the user's database can also include additional attributes for each post, such as date, social media platform, and author. Text data from social media platforms can often be streamed based on search or filter terms (e.g., hashtags). With this in mind, the goal in creating the database is to include datapoints referring to different events and locations while still containing terms referring to critical infrastructures. This creates more comprehensive training and testing sets such that the models are able to generalize to identifying damages in unlabeled data.

3.3 Proposed Framework

The proposed framework for processing and classifying social media big data begins with the development of training and testing sets. The classification task is to identify CIS damage in a social media post, and a different classifier must be built for each system of interest. For the purposes of increasing situational awareness, only posts that refer to a current or ongoing event or system damage state should be considered as relevant information in each post. Ensuring that posts provide information on current, real-time states supports immediate decision-making to increase community resilience, e.g., what

areas need resources or where to send repair crews. Each classifier must have specific definitions of the damages being identified for each system.

To develop the training and testing sets, datapoints must be manually labeled according to definitions of damage for each system. A single datapoint may appear in a training or testing set for multiple systems. The classification of system damage for a post can be binary or multinomial. While separate classifiers can be constructed to predict individual types of damages within a system, this framework assumes the classifiers are constructed to predict all damages for an individual CIS type, e.g., all damages in a power system.

As previously described, the database of social media text data for constructing the training and testing sets should contain a variety of infrastructure-related words and reference multiple events and locations. To label tweets according to damage definition, the database is first searched for datapoints containing specific keywords and terms related to damages in each system and the events or conditions that cause them. The resulting data is a reduced set that can be manually assessed and labeled according to the damage state definitions set by the user.

Once each datapoint is labeled according to damage in a system, it can be split into individual words for classification. From here, if a datapoint contains attributes not useful for classification, it is removed from the training or testing data. In this study, symbols, pronouns, and articles are removed from the text data before classification. Likewise, if a datapoint does not contain a “necessary keyword” or contains a “stop word,” it is thrown out of the training and testing data. A “necessary keyword” is defined as a word or term

that must be an attribute of a post (i.e., datapoint) for that post to be included in the training and testing data for a classifier. A “stop word” cannot be an attribute of a post if that post is to be included in the training and testing data for a classifier. These words will change depending on the classification task and definitions of damage states, and this filtering serves as an additional measure to limit the data for classification by removing noise. For instance, for classifying damage in a power system, the word “outage” may be included as a necessary keyword, and the word “money” may be a stop word. A datapoint without the word “outage” or including the word “money” is assumed to be noise. The remaining manually labeled data for classification contains attributes that filtering alone cannot identify according to defined damage states.

The filtering stage is similar to previous works described in Chapter 2 that use keyword searches and filters to identify CIS information in social media data. The purpose of classification and evaluation after the filtering stage is to predict damage states with a combination of attributes, i.e., words in a post. To ensure suitable sizes for the training and testing datasets, additional datapoints can be searched and manually labeled from the database if a significant amount of datapoints are removed during the filtering process.

Constructing classifiers with the training data is open-ended and depends on the application. For instance, some models are not suitable for multinomial classification. In the application presented in 5.2, three different classifiers are built for two different classification tasks, and the models’ performance is compared. Regardless of the model, hyperparameters should be adjusted using cross-validation to avoid overfitting.

Finally, classifiers constructed through the framework can be evaluated for generalizability on the testing data and other unlabeled data. Typical measures for evaluating machine learning-based classifiers are overall accuracy, precision, and recall. These measures are calculated from the numbers of datapoints correctly and incorrectly classified by the built model on training data, testing data, or through cross-validation. The number of correctly classified datapoints for a particular damage state is the number of true positives, TP , and the number of false positives for a state is FP . The number of datapoints incorrectly classified for a class is FN , i.e., the number of datapoints that are labeled as one state but falsely classified as another.

Overall accuracy is the number of correctly classified datapoints out of the total number of datapoints. Precision, P , is calculated as the number of tweets predicted correctly in a class, or true positives, out of the total number of tweets the classifier predicts in that class, both correctly and incorrectly (i.e., $TP + FP$), as given in Equation (1). Recall or true positive rate, R , is calculated as the number of tweets predicted correctly in a class out of the total number of tweets truly in that class, $TP + FN$, as given in Equation (2).

$$P = \frac{TP}{TP + FP} \quad (1)$$

$$R = \frac{TP}{TP + FN} \quad (2)$$

3.4 Step-by-Step Summary

The steps of the proposed data processing and classification framework are summarized below:

1. Define damage states for the CIS of interest.
2. Search available database for initial keywords and terms related to the CIS of interest.
3. Manually label posts for relevance to each damage state for system of interest.
4. Identify individual words as the attributes of each datapoint for classification.
5. Filter out symbols, pronouns, and articles.
6. Remove datapoints whose attributes do not include necessary keywords or include stop words.
7. Separate the labeled and filtered datapoints into training and testing sets, based on user-defined training and testing ratios.
8. Train classifiers and tune hyperparameters with cross-validation data.
9. Evaluate classifiers on testing data and other unlabeled data, depending on the application.

Figure 1 shows a schematic of the full process of the framework, with steps for training and test set development grouped separately from classification and evaluation steps. The entire framework provides a clear progression to reduce social media big data to relevant information for CIS monitoring and real-time decision-making. The filtering process and identification of current and ongoing damages to develop the training and testing sets before classification reduces the amount of data for classification.

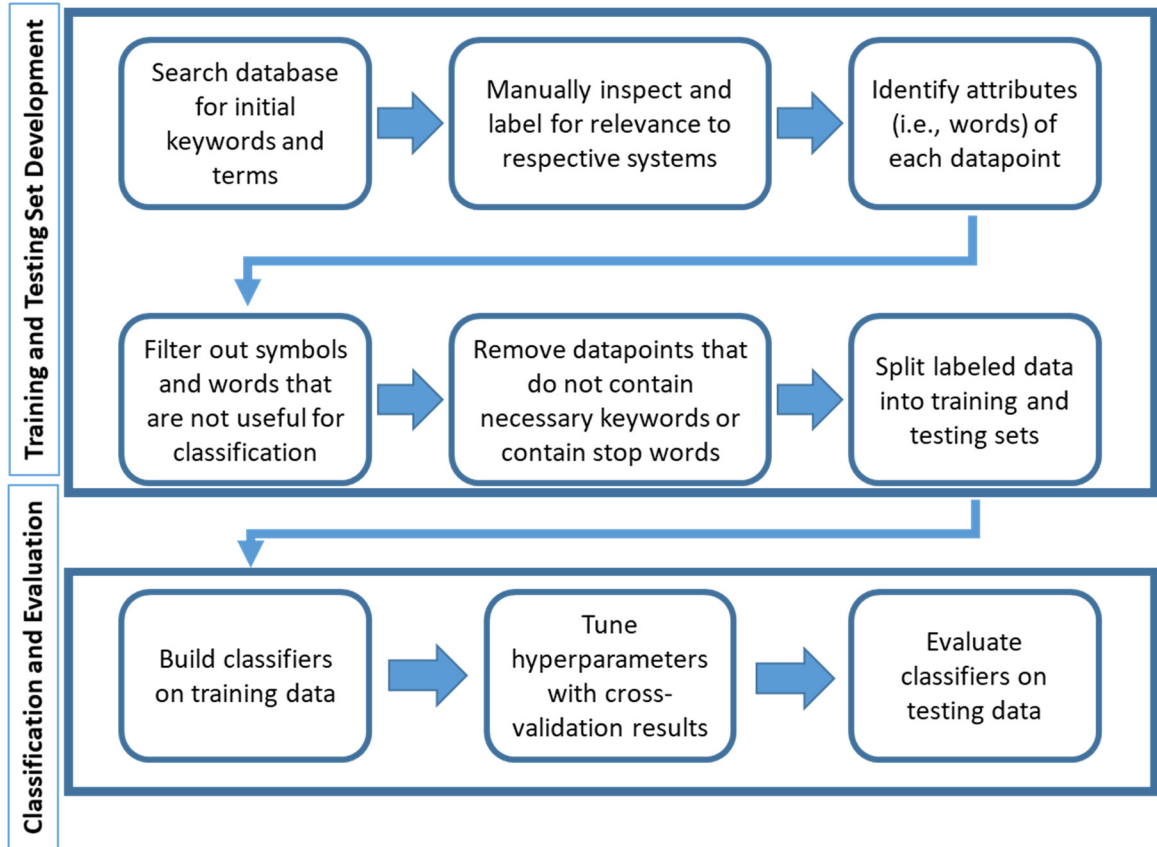


Figure 1. Flowchart for development of text-based classifiers for CIS monitoring

To use any classifiers built from this framework for real-time processing of social media text data, the data must be filtered in the same way as the training and testing set data. An application example of a real-time detection scenario is provided in Chapter 5.

CHAPTER 4. FRAMEWORK FOR DATA INTEGRATION OF HETEROGENEOUS SOURCES

4.1 Introduction

This chapter describes the proposed framework for probabilistically integrating data from multiple, heterogeneous sources in a single time frame. The approach uses Bayesian inference to update prior understandings of risk with new information, as it is collected and can be applied to estimate a variety of events, including CIS damage and hazard event occurrence. In this chapter, definitions and variables used in the framework are first provided, followed by a detailed description of the proposed framework. This data integration framework is then summarized in a step-by-step list, and the process for extending the framework to include additional time periods is described. Two applications of this data integration framework are presented in Chapter 5.

4.2 Variables and Definitions

For the remainder of this chapter and in Chapter 5, the term “sensor source” is used to describe each type of sensor measuring a unique parameter, whether physical, social, or other, and “sensor network” is used to describe a group of sensors from the same source. In the proposed framework, data from sensors within one network is integrated first, and then data from across different sources is integrated. Let θ represent the occurrence of a disaster or failure event at a specified location or affecting a particular system component, and let s_1, s_2, \dots, s_k each represent data from a different sensor source, for k total sources. The prior probability of θ is denoted $P(\theta)$, with the prior probability of nonoccurrence of

the event equal to $1 - P(\theta)$. The framework is applicable for systems or events with multiple states, as long as the states are mutually exclusive and collectively exhaustive. θ is defined as the occurrence of the primary event of interest to be detected, $P(\theta)$ as the assessment to be updated, and $\bar{\theta}$ as the nonoccurrence of θ . Prior probabilities of θ can be based on historical data, physics-based analyses, previously updated probability distributions, expert judgment, or a combination of the above. These represent a current understanding of risk for the event of interest.

4.3 Proposed Framework

To update prior probabilities, appropriate data sources, s_1, s_2, \dots, s_k , must be identified that indicate θ . As no data collection is perfect, these sources indicate θ with some uncertainty. In the proposed approach, the data sources can be physical sensors, social sensors, a combination of the two, or others, and sensors that collect data continuously or provide information only at specific points in time. The data integration across sources, s_1, s_2, \dots, s_k , is ultimately performed through Bayesian updating. Bayesian updating for data fusion combines data from sensors or experiments with prior probabilities of event occurrence to compute posterior occurrence probabilities (Khaleghi et al. 2013). The resulting posterior probabilities from the analysis are the conditional probabilities of events given observed data from the combined sources. Bayesian updating of θ using this data is shown in Equation (3).

$$P(\theta|s_1, s_2, \dots, s_k) = \frac{P(s_1, s_2, \dots, s_k|\theta)P(\theta)}{P(s_1, s_2, \dots, s_k)} \quad (3)$$

From Equation (1), $P(s_1, s_2, \dots, s_k | \theta)$ is the joint conditional probability of observing data from all sensor sources for updating given event occurrence, representing the likelihood of all sources, and $P(s_1, s_2, \dots, s_k)$ is the joint probability of observing data from all sources. The resulting posterior probability, $P(\theta | s_1, s_2, \dots, s_k)$, is the probability of event occurrence given observed data from different data sources.

Calculating the likelihood of all sources and the joint probability of all data requires several intermediary calculations. First, the integrated likelihood of data given θ from each individual network, $P(s_i | \theta)$, must be calculated, where s_i represents the i^{th} data source out of k , by combining the observations indicating θ within each network. These calculations are dependent on the specific application of interest, the nature of the data, any possible relationships between the data (including dependence or independence of observations within a network), and the accuracy or reliability of each source. The observed data from each source must correspond in terms of date, time, and/or location to the event of interest. Out of the sources considered, the number of sources k available for updating may vary in time. For instance, if on a day, only one of the considered sources provides data indicating θ , then $k = 1$, and k is subject to change the following day. The application example in Chapter 5 demonstrates how these calculations can be performed for two specific data sources. The calculations of $P(s_i | \theta)$ are intentionally left open and flexible so that the framework can be applied to different sensor sources outputting varying types of data with varying likelihood calculations for θ .

The next step of the proposed framework is to compute the joint likelihood of observing data from all sources, $P(s_1, s_2, \dots, s_k | \theta)$, from the integrated likelihoods of data

within individual networks. The joint likelihood calculations are based on the assumption that individual source likelihoods are independent of each other when conditioned on the same event, location, or system component of interest, i.e., likelihoods are conditionally independent given θ . From this, the joint likelihoods are computed as the products of the individual source likelihoods for each event state, as in Equation (4).

$$P(s_1, s_2, \dots, s_k | \theta) = \prod_{i=1}^k P(s_i | \theta) \quad (4)$$

The proposed framework next requires calculation of the joint probability of observing data from all sources, $P(s_1, s_2, \dots, s_k)$, using the total probability of data. This is possible because all states of the event are mutually exclusive and collectively exhaustive. Equation (5) shows this calculation for two states of the event, θ and $\bar{\theta}$, i.e., occurred and not occurred. $P(s_1, s_2, \dots, s_k | \theta)$ is previously calculated from Equation (4). The likelihood of data for the nonoccurrence of θ , $P(s_1, s_2, \dots, s_k | \bar{\theta})$, is computed in the same way, i.e., based on the likelihood of data from each source and assuming conditional independence. The likelihood of data from each source indicating $\bar{\theta}$, $P(s_i | \bar{\theta})$, depends on the application and source, just as $P(s_i | \theta)$ depends on them.

$$P(s_1, s_2, \dots, s_k) = P(s_1, s_2, \dots, s_k | \theta)P(\theta) + P(s_1, s_2, \dots, s_k | \bar{\theta})P(\bar{\theta}) \quad (5)$$

Finally, the posterior probability of occurrence of the event of interest, θ , is updated given the data across sources by inputting the results of the previous calculations into Equation (3). This posterior represents an updated understanding of risk for the disaster or failure event in real time as data becomes available.

4.4 Step-by-Step Summary

In summary, the steps of the proposed approach are as numbered and described below:

1. Compute prior probability or probabilities of occurrence for the event(s) being considered, $P(\theta)$.
2. Identify the sensor sources available that indicate θ .
3. Determine which observations from each source indicate θ , with some uncertainty.
4. Integrate likelihoods of data from sensors within each sensor network to calculate the overall likelihood of data from each source, $P(s_i|\theta)$, $i = 1, \dots, k$, where k is the number of sources. This process varies depending on the data output from each source. Repeat this for all sensor sources and states of the disaster or failure event.
5. Calculate the joint likelihood of data from all sources, $P(s_1, s_2, \dots, s_k|\theta)$. Assuming the sources are conditionally independent on θ , the joint likelihood is the product of all source likelihoods. Repeat this for all states of the event.
6. Calculate the joint probability of observed data, $P(s_1, s_2, \dots, s_k)$.
7. Calculate the final posterior probability of event occurrence updated given data from across sources, $P(\theta|s_1, s_2, \dots, s_k)$.

4.5 Sequential Updating

Additionally, disaster or failure events in a community are often dynamically evolving in time. To consider data over multiple sequential time periods to estimate events,

sequential Bayesian updating is employed in the steps above. Equation (3) becomes Equation (6) to compute a posterior probability using sequential data over two time periods. s represents the first set of data from sensor sources $1, \dots, k$, and t , the second set from sources $1, \dots, m$. More time periods can be added as data is available, and it is not necessary to collect data from all of the same sources at every time period, as the availability of data from each source may change over time. Data output from across sensor networks is assumed to be conditionally independent from one time period to the next, with the joint likelihood of data from multiple time periods found by multiplying all available likelihoods. The purpose of sequential updating is to include data from multiple observed periods for updating events that develop and occur over time. Sequential updating in the proposed framework ends when updated probabilities approach 1, suggesting that occurrence of the event θ has been detected with near certainty, or when data indicating θ is no longer available. After obtaining posterior probabilities from updating or sequential updating, the proposed framework can continue to be used to detect θ in the future, with prior probabilities that remain the same or are re-evaluated based on the results of updating and the nature of the event of interest, e.g., if events occur as Poisson processes or with cumulative effects. For example, if damage is detected in an infrastructure component with a certain probability and no action is taken to remediate that damage, the updated probability can be assumed as the new prior. In contrast, if action is taken to repair that component, the prior probability of damage for future updating may be lower than the original prior. The same or additional data sources can be used.

$$P(\theta|s_1, s_2, \dots, s_k, t_1, t_2, \dots, t_m) = \frac{P(s_1, s_2, \dots, s_k|\theta)P(t_1, t_2, \dots, t_m|\theta)P(\theta)}{P(s_1, s_2, \dots, s_k, t_1, t_2, \dots, t_m)} \quad (6)$$

CHAPTER 5. APPLICATION EXAMPLES OF PROPOSED FRAMEWORKS

5.1 Introduction

This chapter presents four applications of the frameworks described in Chapters 3 and 4. In sections 5.2 and 5.3, applications and results for the social media processing and classification framework are presented. Twitter data is selected as the social media source for both applications. 5.2 is an application to explicate the framework's use and describes the construction, evaluation, and comparison of three different classifiers built to identify damages in two CIS: transportation and power systems. In 5.3, a classifier constructed and evaluated in 5.2 is further evaluated on a set of unlabeled data collected during a hazard event to simulate a real-time detection scenario. The results and implications of the classifier's performance for real-time detection are discussed in detail.

In sections 5.4 and 5.5, applications of the data integration framework are presented. 5.4 is an application for estimating flood risk in Louisiana during August 2016. In this application, each step in the data integration is presented in detail to explicate the framework's use. The results are validated with true post-event mapping and flood modeling. In 5.5, the data integration framework is used to estimate probabilities of transportation damage in Atlanta during Tropical Storm Irma. This application includes the results from 5.3 and is an example of how multiple data sources can be used to update the inputs of a CIS model. Specific validation of the application in 5.5 is not included; rather,

the implications and benefits of the probabilistic data integration framework for CIS situational awareness and resilience is discussed.

5.2 Application of Proposed Text-Based Classification Framework for Building Supervised Learning Classifiers to Identify Infrastructure Damage

Classifiers are built to identify damages in two CIS: transportation and power systems. Twitter data is selected in this application for CIS monitoring because of its open availability, accessibility as a free application, and relatively low costs for data collection across communities. All tweets considered in this study are posts of up to 140 characters. While photos, videos, and other multimedia attachments can be added to tweets, only the texts of tweets for application of the proposed framework. The text-based classification can easily be extended across multiple social media platforms as described in Chapter 3.

The classifiers are trained on data collected over multiple days, referencing multiple events and conditions, and only labeled relevant if the referenced damages are current or ongoing based on temporal words. Classification is based on the words of a tweet alone, without any additional semantic analysis to determine relevance to CIS damages, and the method and results presented are not limited to a single type of damage or hazard event. Further discussion is provided in 5.3 on the classifiers' ability to facilitate real-time decision-making for the protection and recovery of CIS and increasing resilience of communities that depend on them.

5.2.1 Twitter Data Collection and Database

Two methods are used to build the tweet database. First, to collect tweets in real time, tweets are streamed from Twitter using Twitter's APIs (application program interfaces). Each tweet includes the date and time the tweet was created, the text posted, information about the poster, and if the poster has selected to enable his or her location to appear, the location at which the tweet was posted. Twitter limits streaming to about 1% of the total number of tweets published in a given time window (Kumar et al. 2014). The second method is to scrape information directly from Twitter's online search page. This method enables searches for openly available tweets posted at any time in the past. For the remainder of chapter, these are referred to as "historical" tweets. Though the downloaded data for these tweets contain less information than for those streamed in real time, they contain the major aspects of a tweet object: the poster's username, text of the tweet, hashtags within the tweet, date and time the tweet was published, number of retweets, and location if available. To train the classifiers, the individual words of each tweet are defined as the attributes of each datapoint.

In both the real-time streaming of tweets and reading of historical tweets, datapoints are collected containing potential words of interest within the texts. For instance, tweets are filtered for keywords that may relate to the CIS of interest or hazard events causing damage, including "outage," "street," "flooded," "detour," etc. Data is collected from multiple time periods, including from dates on which a known damaging event has occurred. For example, tweets from August 13-21, 2016, covered dates of storms in southern Louisiana causing significant and damaging floods. These tweets are filtered to contain keywords such as "Louisiana," "flood," "flooded," "power outage," and "blocked

road” to ensure they refer to the event or condition of interest (i.e., the August 2016 floods and damages caused by them). As described in Chapter 3, goal is to create a database referring to different events and locations while still containing terms referring to critical infrastructures or events impacting them. This creates more comprehensive training and testing sets such that the models are able to generalize to identifying damages in unlabeled data. Tweets streamed during Tropical Storm Irma as it hit the metro Atlanta area on September 11, 2017, are used as previously unseen and unlabeled tweets for final predictions using the built classifiers in section 5.3 and are not included in the training and testing sets.

In total, 352,514 tweets are streamed or scraped from 27 different days from August 13, 2016, to November 7, 2017, and are used for developing the training and testing sets for relevance to transportation or power damages. The tweets reference a variety of events, locations, and potential damages to critical infrastructures.

5.2.2 Training and Testing Set Development

Application of the framework for classifying social media data begins with damage state definitions and training and testing set development. In this application, the focus is on identifying the relevance of tweets to transportation and power system damages. These systems are selected because of their prevalence in social media posts compared to other systems. For the purposes of increasing situational awareness, only posts that refer to a current or ongoing event or system damage state are considered relevant in the damage event definitions. References to repaired damages or components back online (i.e., past damages) and warnings or advisories (i.e., future or potential damages) are considered

irrelevant. Ensuring that posts provide information on current, real-time states supports immediate decision-making to increase community resilience, e.g., what areas need resources or where to send repair crews.

In this application, two damage states for each system are defined for a binary classification. Each tweet is either relevant to damage or irrelevant. Damage to a transportation system is defined as: any closed, blocked, or flooded roadway or highway (either fully closed or partially closed); physical damages to roadways or roadway structures including potholes, sinkholes, or bridge damages; detours that are required from a closed roadway or highway; and any unusual and significant increase in traffic or travel times. Power system damage is defined as an outage or a downed or damaged power line.

Oftentimes, relevant tweets refer to damages occurring in one or both systems at multiple scales. Many tweets relevant to transportation damage refer to specific intersections, e.g., describing the closest point to a blocked road. In comparison, most of the tweets relevant to power damage refer to outages on a larger scale, e.g., for a metropolitan area. Tweets referencing both local and system-level damages are considered relevant.

Table 1 shows examples of the texts of relevant tweets within the database for transportation and power system damages according to these definitions. Table 2 shows examples of irrelevant tweets in the database that contain some keywords that are typically present in relevant tweets.

Table 1. Example tweets that are relevant to transportation and power system damages given definitions

	Text
Transportation	“City of Brookhaven reports fallen tree blocking road at Colonial Drive and Pine Grove Avenue”
Power	“No electricity in Dinwiddie and surrounding areas since 18H00, would EMM please explain?”

Table 2. Example of tweets that are irrelevant to transportation and power system damages given definitions

	Text
Transportation	“Reminder: Driving past barricades onto flooded roads = BAD IDEA. It doesn’t take much water to total a vehicle: bit.ly/2lnJhcc ” (image attached)
Power	“#MysteryAI 16-hour power outage yesterday. Still trying to catch up and get back into some sort of routine.”

In Table 1, the tweet relevant to transportation system damages refers to local damage within the larger transportation system in a city in the United States. The tweet relevant to power system damages refers damage at the system level: a power outage in a city in South Africa. In Table 2, while the first tweet refers to flooded roads, it is posted as a warning to others about how to react when driving during flood events rather than providing information about the current state of a specific road. The second tweet refers to a power outage, but one that has occurred previously, as evidenced by the word “yesterday” and other context in the post. These tweets are examples of how closely related relevant and irrelevant tweets can be. Rather than just identifying a reference to an infrastructure as in previous studies, these definitions of relevant and irrelevant information support detection of true, real-time damage. Many examples similar to those shown in Table 1 and Table 2 are included in the training and testing sets to introduce the models to additional attributes that may indicate the relevance or irrelevance of a tweet.

The training and testing sets to build and evaluate the classifiers are formed from tweets manually labeled as relevant or irrelevant to damages for each system. A single tweet may appear in a training or testing set for both systems. To represent the binary classification of a tweet, relevance to system damage is labeled 1, and irrelevance to system damage is labeled 0. While separate classifiers can be constructed through the proposed framework to predict individual types of damages within a system, here, a single classifier for each CIS type is used to predict all relevant damages in that system.

As previously described, the database of tweets for constructing each training and testing set contains a variety of infrastructure-related words from multiple time periods. To label tweets, the database is first searched for tweets containing specific keywords and

terms related to damages in each system and the events or conditions that cause them. The resulting tweets are either related to the CIS of interest or not related. Tweets that are not related contain keywords or terms that could make it appear related. Table 3 shows example search keywords and terms.

Table 3. Example keywords/terms for searching database

	Keywords/Terms
Transportation	Road(s), street(s), highway, lane, closed road, blocked road, flood(s), flooded road, tree down, accident, major delays, pileup, bridge, collapse
Power	Power, power outage, no power, electricity out, no electricity, electrical outage, live wire, down power line, tree down, storm

From the results of the search for each system, tweets are then manually labeled as relevant or irrelevant to damage. No database can provide the full variety of information in tweets published daily that are potentially relevant to CIS damages, with millions of tweets continuously available from around the world on evolving topics and trends. However, the wide range of search terms and periods used in data collection for this application creates

more representative training and testing sets than data from fewer searches and time periods.

Ultimately, 1,065 of the searched tweets are labeled as either relevant (265) or irrelevant (800) to transportation system damages, and 301 tweets as either relevant (157) or irrelevant (144) to power system damages. The labeled tweets cover 14 and 10 days (i.e., out of the 27 represented in the database) for transportation and power systems, respectively. The labeled tweets include different examples of damage, types of events causing damage (e.g., flooding, storms, vehicle accidents, etc.), and sources of potential noise (i.e., irrelevant tweets) to reduce bias towards a single event or type of damage in the classifiers. For transportation systems, the majority of tweets relevant to damage describe blocked or flooded roadways, and the majority of tweets relevant to power system damages refer to power outages. However, selecting a disproportionate number of datapoints referencing the same event is avoided to reduce the bias in labeled data. Thus, there is no clear distribution of events causing damage in the training or testing sets for either system. Additionally, no repeated tweets (retweets) are included in each set, making each of the labeled tweets unique. The results are more sensitive to the definitions of damage for each system than to the size of the training datasets. This is discussed in subsequent sections, including an analysis of the effect of the size of the training sets on results.

The proposed framework focuses on the texts of the tweets as attributes for training and testing the classifiers; therefore, the attributes of each datapoint are the individual words in that tweet. The full set of attributes are the words found in all tweets. 75% of tweets for each system are used for training. The remaining data for each system is reserved for testing and evaluation. In both training and testing sets, the ratio of relevant and

irrelevant tweets for each system is preserved. The training and testing sets for relevance to transportation system damages have 2,340 and 1,763 attributes (words), respectively, and the training and testing sets for relevance to power system damages have 1,233 and 578 attributes, respectively.

Prior to building the models, the filtering step removes any attributes or datapoints that are found to be unnecessary or insignificant to the classification. Words within a tweet are delineated by spaces. However, this can result in erroneous attributes, which are then removed. For example, hashtags, URLs, symbols (including conversion of emojis to symbols), and users mentioned within a tweet (with the “@” symbol) are considered as single words. In addition, the same words with different capitalization are considered different attributes, e.g., “ROADS,” “road,” and “Road.” To address these cases, hashtags are separated into individual words where possible (e.g., #HurricaneIrma becomes Hurricane Irma), all words are converted to lowercase so that repeated words are considered as single attributes, and symbols and URLs are removed. Pronouns and articles are assumed to be insignificant for determining the relevance of a tweet to system damages and remove them from the training and testing data.

Finally, the filtering step removes datapoints, i.e., tweets, that contain stop words or lack necessary keywords. These tweets are automatically denoted irrelevant and not useful for building the classifiers (i.e., they are removed as noise).

Table 4 summarizes the necessary keywords and stop words used for filtering. The necessary keywords overlap with the example keywords used to search the database shown in Table 3. The stop words may change depending on the time of data collection (e.g., the data collected for this study spanned the 2016 United States presidential election, and thus, stop words include “trump,” “obama,” and “clinton”). Requiring at least one of the necessary keywords in Table 4 to appear in a tweet removes irrelevant tweets that contain a search word from Table 3 but ultimately do not contain any words that identify specific components or damages. It is assumed that any leftover misspelled words or grammatical errors do not significantly affect the models. The proposed framework includes this final filtering process to reduce noise that may appear in the data that are unrelated to any of the CIS damages of interest. In this way, the classifiers are trained and tested on data for which a significant amount of noise has already been filtered out. The resulting sets for identifying transportation damage contain 1,878 and 1,644 attributes for the training and testing sets, respectively, and 1,133 and 709 attributes, respectively, for power system damage. Table 5 summarizes the final sizes of the training and testing sets after this filtering.

Table 4. List of filter terms for training and testing set development

	Necessary Keywords:	Stop Words:
	Tweets must contain at least one of these	Tweets cannot contain any of these
Transportation	Road(s), street(s), highway(s), lane(s), flood(ed), detour, accident, closed, blocked, down(ed), traffic	Politic(s), pray(er), love, thoughts, memory(ies), friendship, devastate, devastating, election, trump, obama, clinton
Power	Power, outage, electricity, electric, electrical, wire, down, line	Fight, money, politic(s), election, recovery, dollars, trump, obama, clinton

Table 5. Final training and testing set sizes

System	Training Relevant	Training Irrelevant	Training Total	Testing Relevant	Testing Irrelevant	Testing Total
Transportation	209	551	760	53	189	242
Power	113	107	220	42	32	74

5.2.3 *Classification and Evaluation*

With these datasets, three classifiers are built for each system: support vector machines (SVM), naïve Bayes models, and decision trees. These three classifiers are selected because of their relatively high explainability, representation of different classification algorithms, and prevalence in machine learning applications. The models are trained and implemented with Weka, a data mining and machine learning software that is implemented in Java (Hall et al. 2009). The hyperparameters are selected as linear kernels for support vector machines, Gaussian distributions for attributes per class for naïve Bayes, and information gain rankings for determining decision tree nodes.

The classifiers are built on the training datasets described in the previous section. The results of evaluating the models on classifying the training data and of 10-fold cross-validation on the training set are shown in Table 6 and Table 7, respectively. Cross-validation results from varying the hyperparameters of each model are used to confirm that the selected inputs create the highest performing models without overfitting the training

data. For each system and classifier, Table 6 and Table 7 display the percentage of tweets correctly classified overall and the precision and recall for each class. The classes being predicted are the relevance or irrelevance of a tweet to CIS damage.

Table 6. Results of models in classifying training data

	Classifier	Percent Correctly Classified	Precision		Recall	
			Relevance	Irrelevance	Relevance	Irrelevance
Transportation	SVM	99.9	0.995	1.00	1.00	0.998
	Naïve Bayes	88.4	0.78	0.93	0.80	0.92
	Decision Tree	93.4	0.94	0.93	0.81	0.98
Power	SVM	99.5	1.00	0.99	0.99	1.00
	Naïve Bayes	90.9	0.90	0.91	0.92	0.90
	Decision Tree	79.5	0.75	0.86	0.89	0.69

Table 7. Results of classification evaluated using 10-fold cross-validation

	Classifier	Percent Correctly Classified	Precision		Recall	
			Relevance	Irrelevance	Relevance	Irrelevance
Transportation	SVM	89.2	0.90	0.89	0.68	0.97
	Naïve Bayes	86.2	0.75	0.91	0.76	0.90
	Decision Tree	86.3	0.80	0.88	0.67	0.94
Power	SVM	70.0	0.70	0.70	0.74	0.66
	Naïve Bayes	75.0	0.77	0.73	0.74	0.77
	Decision Tree	61.4	0.61	0.62	0.69	0.53

The results show that SVMs have the highest accuracies when evaluated on the training data for both systems. In cross-validation, the models decrease in performance, with SVMs having the highest accuracy for predicting each class for transportation damage and naïve Bayes having the highest accuracy for predicting each class for power system damage.

Evaluating the results along varying measures for each system provides insights into classifier performance. For transportation systems, the cross-validation recall is higher for irrelevance than for relevance. This shows that the models are more successful at identifying noise in the data than detecting damage. While the models may disregard many relevant tweets with a low recall for predicting relevance, the high precision of the models for relevance show that when the models classify a tweet as relevant, they have a high accuracy. In real-time monitoring of CIS for increased situational awareness, of the two measures, having high precision is more valuable than having high recall as it reduces the probability of the models incorrectly classifying an irrelevant tweet as relevant. False alarms from low precision models for transportation or any other system damage can be detrimental for decision-making in repair and recovery and can impede emergency response during crises.

For power systems, the classifiers are less successful in correctly distinguishing between relevance and irrelevance to damage. Although predictions on the training sets show high accuracies comparable to the models predicting relevance to transportation damage, the models predicting relevance to power system damage do not generalize as well when evaluated using cross-validation. The cross-validation results also show that, in contrast to the transportation classifiers, the power classifiers have higher recall for relevance versus irrelevance. This indicates higher performance for identifying damages compared to noise in the data. The models' abilities to accurately identify relevance and irrelevance are due to the different types of damage defined for each system, as discussed later in this section.

Each classifier is now evaluated on the testing sets, with results shown in Table 8, to assess the models' predictive performance on unseen tweets from the same database as the training data. The purpose of using the models to classify the testing data is to further assess their generalizability to predict the relevance of unseen data to infrastructure damage.

Table 8. Results of models in classifying testing data

	Classifier	Percent Correctly Classified	Precision		Recall	
			Relevance	Irrelevance	Relevance	Irrelevance
Transportation	SVM	88.4	0.90	0.88	0.53	0.98
	Naïve Bayes	85.5	0.65	0.93	0.76	0.88
	Decision Tree	86.4	1.00	0.85	0.38	1.00
Power	SVM	66.2	0.70	0.61	0.71	0.59
	Naïve Bayes	59.5	0.64	0.53	0.67	0.50
	Decision Tree	62.2	0.64	0.58	0.76	0.44

As expected, the overall accuracies are the same or lower than the results from cross-validation because the testing data introduces the models to larger sets of unseen data and results are not averaged across folds. As in cross-validation, the models classifying relevance of tweets in the testing set to transportation damage identify noise with high recall and precision. In selecting between models, the precision of both the SVM and decision tree models for predicting relevance is high. For the naïve Bayes model, while precision is lower, its recall for predicting relevance is consistent with the results from cross-validation and higher than the other two models evaluated on the testing set. However, with a lower precision, more false alarms are present than in the results from the other two models. To avoid this, high precision is preferred to high recall for predicting relevance. Moreover, the SVM has the highest percentage of correctly classified tweets in the test data, and the decision tree model has a very low recall despite having perfect precision in predicting relevance. With these results, the SVM is selected as the model with the highest performance for detecting both tweet relevance and irrelevance to transportation damage.

In evaluation on the testing data, the power system classifiers also have reduced performance compared to the cross-validation results. In this binary classification, 50% correct classification of data is equivalent to the results of randomly guessing the class of a tweet, and the models perform around this threshold for detecting relevance or irrelevance to power system damage. Among the models, the SVM has the highest and most consistent performance and has similar precision and recall results for both relevance and irrelevance. However, while the models performed well in classifying the training data directly and in evaluation with cross-validation with high accuracies as shown in Table 6 and Table 7,

their performance in predicting the test data suggests that they do not generalize well to larger sets of unseen data.

To explain the difference in performance requires a closer examination of the types of damage identified for each system. The models for predicting the relevance of tweets to transportation damage are most successful at identifying noise. On the other hand, the models for predicting tweet class for power system damages are more successful at identifying relevance, although they do not generalize well for either class. In addition to the examples provided in Table 1 and Table 2, Table 9 provides examples of tweets relevant and irrelevant to damage in each system according to the definitions. The classifiers' different levels of performance for the transportation and power system data highlight their sensitivity to the definitions of damage for each system used in developing the training and testing sets.

Table 9. Examples of relevant and irrelevant tweets for damages in transportation systems

Class	Text
Relevant	“#BREAKING: The overpass on I-85 just collapsed near Piedmont Road”
	“Wisconsin Street Closure in Hobart for Water Main Break Cleanup”
	“Power lines down in the street, endless amount of debris. Scratching the surface of the damage in #Albany” [photos]
	“@n_izzah pothole at Sprint Highway/ Jalan Maarof getting bigger and dangerous. Can plse help get DBKL to look into it? many TIA.”
	“The winner is... the Forth Road Bridge! After discovery of a fracture in 2015 engineers repaired & reopened it in 3 weeks #ICEpeopleschoice” [link attached]
Irrelevant	“Cheers for the Georgia state troopers who evacuated the #I-85 bridge before it collapsed. You guys saved lives today (again). Thank you”
	“Update on North West flooding repair-around 80% of closed roads now reopened & 750metres of flood defences repaired. infrastructure-ni.gov.uk/news/update-no... ”

Table 10. Examples of relevant and irrelevant tweets for damages in power systems

Class	Text
Relevant	“Power still out but it’s not 98 degrees like the last time there was a sustained power outage.”
Irrelevant	“3 rd grade made the best out of the power outage this morning by reading informational text w/ flashlights. Nothing stops learning @South!”
	“Hello! Due to the power outage and crazy week we were unable to put up our Halloween photos, so here they are!... fb.me/7u6jGJiKP”

The transportation system classifiers consistently have high precision for relevance because the definitions for damage are distinct, despite having multiple types of damage in the definitions considered. On the other hand, these models have lower recall than precision for relevance and high recall for irrelevance because irrelevant tweets cover a wide range of subjects, including references to noncurrent transportation damages. The initial search words and necessary keywords for transportation systems have multiple meanings and can reference a variety of conditions or unrelated sentiments that indicate irrelevance regardless of the temporal context of a tweet. For instance, the word “accident” can refer to many other events besides vehicular accidents that cause transportation system damage. If a tweet refers to transportation infrastructure, the classifier must decide if it refers to damage and if so, current damage. This increases the difficulty for the classifiers to identify all relevant

tweets. However, most irrelevant tweets in the transportation system training and testing data do not refer to damage, and the classifiers are successful at detecting noise with high precision and recall. In Table 9, tweets relevant to multiple types of damage (i.e., bridge collapse and blocked streets from two different causes) are displayed with irrelevant tweets that refer to transportation infrastructure but do not refer to damage or refer to past damages.

Compared to the transportation system data, the data for power system damages are more nuanced in their distinction between relevance and irrelevance due to the specific definitions of damage. Since damages are either power outages or damaged power lines, there is less noise from various meanings of the search words and necessary keywords. Instead, most irrelevant tweets refer to past events or warnings for future events, and noise is less discernable than it is in the transportation system data. While the distinction between current and noncurrent damage is critical for increasing real-time situational awareness, i.e., only the current state of the system is of interest, the difference between relevant and irrelevant tweets is difficult for the classifiers to distinguish. Table 9 shows this nuance with one relevant and two irrelevant tweets. While the reader is able to distinguish the difference between the current and past power outages in these examples, there are few useful temporal attributes that appear consistently enough in the training data from which the classifiers can learn without focusing on exact events referenced in individual tweets. To improve the performance of the classifiers for detecting power system damage, the training sets need to be increasingly representative of relevant and irrelevant tweets, which may not be possible as a classification task based on the typical definitions of power system damage as used in this study.

The effect of training size on the classifier performance, specifically its recall and precision, is now analyzed. Results are shown for the SVM for detecting transportation system damage as it generalizes well to evaluate the testing set. Figures 1 and 2 show the performance metric of the classifier evaluated through 10-fold cross validation as training size increases, where the training size is varied by sampling 10% to 100% of the full training set without replacement. In many machine learning applications, the performance of a classifier increases as the amount of training data increases. In Figure 2 and Figure 3, an initial improvement in performance is seen with increases in training dataset size. However, the performance then levels off as the number of datapoints in the training set increases. This indicates that the performance would not significantly improve with more training data from the database of tweets. Rather, adding tweets from new databases to increase representativeness of types of damages within each system or changing the definitions of relevance or irrelevance would improve classification performance for unseen data.

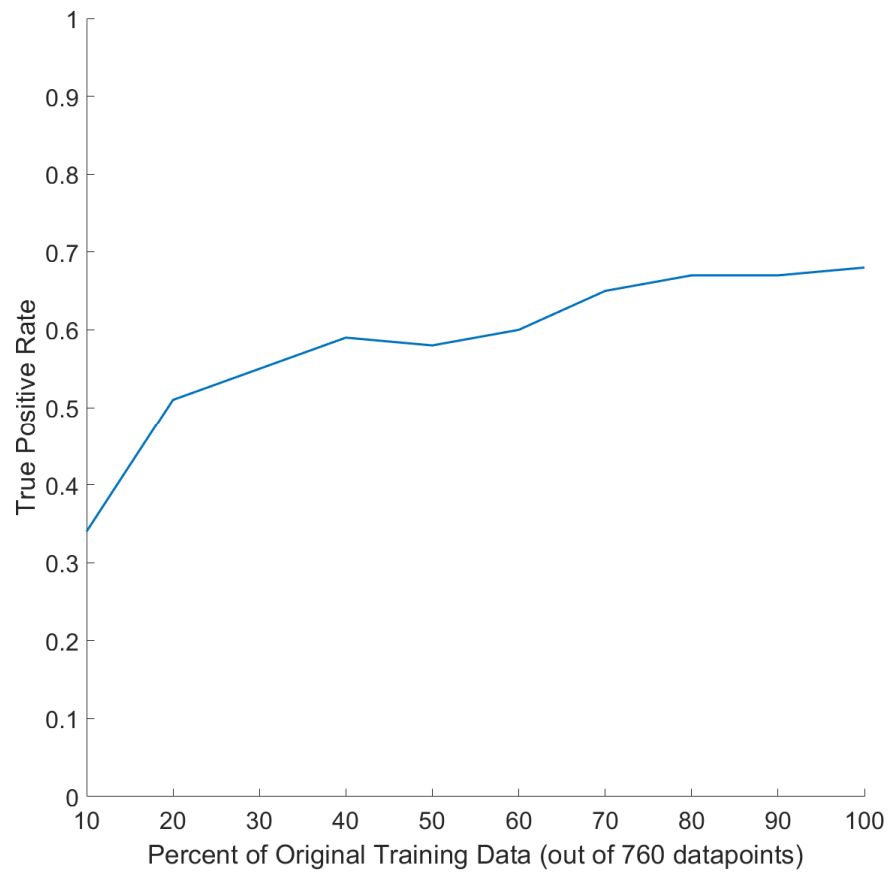


Figure 2. Effect of training size on recall for relevance

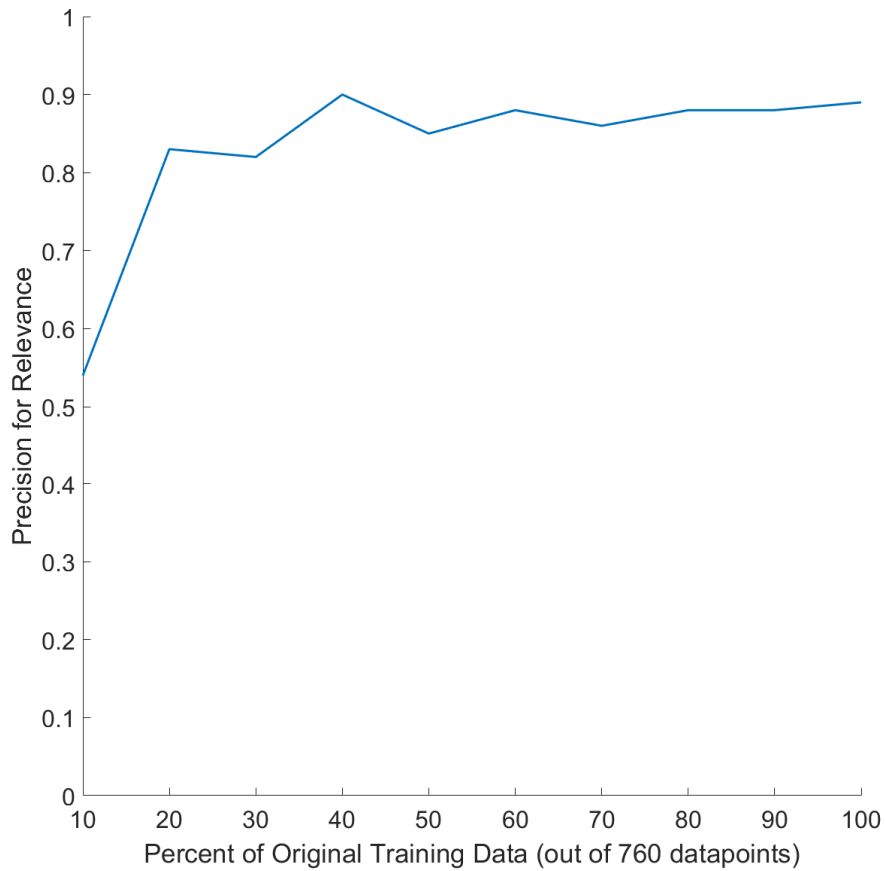


Figure 3. Effect of training size on precision for relevance

In the next section, the same filtering steps from the framework – summarized in Figure 1 – is used to apply the classifier to predict the presence of transportation damages in a separate, unlabeled dataset of tweets. The model’s capabilities to predict the classes of new data are assessed. Given the poor performance of all three classifiers for predicting relevance to power system damage in the testing set, focus is on detecting transportation system damage in the next section.

5.3 Application of Social Media Classifiers for Identifying Critical Infrastructure Damages During Tropical Storm Irma in Atlanta, Georgia

The original training and testing datapoints manually labeled in section 5.2 were selected from the same overall database of tweets from 26 different days of historical and streaming data collection. The goal of the framework for processing and classifying social media data is to increase monitoring capabilities and situational awareness for current and previously unidentified system states. In this section, a model is used to classify a set of unlabeled tweets collected in real-time during a hazard event for relevance to transportation system damages. The new data is separate from the original training and testing data and collected during a hazard event. The SVM classifier for transportation system damage is used because of its high precision and accuracy for both classes. Performance is evaluated based on the ability to predict the relevance or irrelevance of tweets retrieved during a hazard event.

The hazard event of interest is Tropical Storm Irma as it hit the Atlanta area. In the analysis, 188,267 tweets are streamed from September 11, 2017, when the storm reached Atlanta, filtering for words related to infrastructure damages, the Atlanta area, and Tropical Storm Irma. Tweets are streamed from around the world, but for this data collection in a hazard event, tweets are filtered for references to Atlanta as CIS damages were anticipated during the event in the city and surrounding area. The Federal Emergency Management Agency (FEMA) declared a Major Disaster Declaration in Georgia on September 15, 2017, with all counties covered for Individual and/or Public Assistance (FEMA 2017).

Data processing of the unlabeled tweets follows the flowchart as shown in the Training and Testing Set Development portion of Figure 1. As in first developing the training and testing data for transportation systems in section 5.2, the unlabeled tweets are searched for words and terms relevant to potential transportation damages. Due to the importance of the hazard in this event and to simulate real-time detection, the search terms for the unlabeled data also include words referring to Tropical Storm Irma and the Atlanta area as well as the terms related to transportation damages. Tweets that do not contain necessary keywords and contain stop words are also removed, but tweets are not manually evaluated for relevance. Instead, the transportation system SVM is used to automate predictions of relevance in the unlabeled data. After filtering, the dataset for analysis consists of 16,106 tweets, again demonstrating the reduction of a large dataset to a more manageable dataset for assessment.

These unlabeled tweets are classified with the built SVM model. The SVM predicts the class for 16,106 datapoints in 5 seconds on an 8 GB RAM computer with 1.70 GHz processor, so it is efficient for real-time monitoring. The classification results in 2,130 tweets (13.2%) relevant to transportation damage and 13,976 irrelevant tweets. After discounting retweets, 1,192 relevant datapoints remain. To evaluate the performance of the classifier, the 1,192 unique tweets predicted relevant by the SVM are then manually assessed to determine which are truly relevant to transportation damage based on the definitions and which are irrelevant. 813 are found to be relevant and 379 are irrelevant, for a precision of 0.68. While this is lower than the 0.90 precision in evaluating the classifier on the testing data, as expected for this unlabeled dataset, the model is still able

to classify unique tweets with better than 50% precision for data that is compiled completely separately from the original training and testing data.

Evaluating the correctly and incorrectly classified relevant tweets reveals the types of tweets the model classifies well and those it mistakenly classifies as relevant. The correctly classified tweets explicitly reference closed or blocked roads or roadway structures across a variety of causes, including floods and accidents. In addition, the model correctly classifies types of damage not included in the original definitions of damage for transportation systems. For instance, closed rail tracks are not included in the original definitions but appear in the tweets classified as relevant by the model. As the objective is to identify current damage in transportation systems, these are considered correctly classified tweets from the model. The SVM identifies these damages, not originally defined in the model, from the attributes of the tweets. The tweets incorrectly classified as relevant include references to school and facility closings, warnings regarding weather and crisis preparation, and references to repaired damages (e.g., reopened roads after closings).

These insights from the classification results inform adjustments to the definitions of damage, the development of training and testing sets, and the sets of necessary keywords and stop words for filtering prior to creating the model. Because trends, topics, and events discussed on social media platforms are constantly evolving, models for increasing situational awareness for CIS must be trained on data that are representative of current events and language. Therefore, training and testing sets should be kept up-to-date to reflect time-appropriate context. It is important to add data from new databases to the training data to continually maintain or increase their representativeness for real and current damage

events and prevalent social media language to use the classifiers for increased situational awareness and community resilience.

5.4 Application of Proposed Data Integration Framework for Flood Event Estimation in Louisiana, United States

In this section, a specific example to illustrate application of the proposed data integration framework is presented. The application is to update flood risks in the state of Louisiana during August 10-13, 2016, by integrating data from both physical and social sensor sources using the proposed framework. Flood risk around the world is increasing due to climate change and other environmental factors (Hirabayashi et al. 2013), and Louisiana is known to be subject to high flood risk given its low elevation and coastal proximity (Groves et al. 2016). Flood events are distributed over large geographic areas that cannot be completely and continuously monitored in real time. In addition, uncertainty exists in terms of event timing and location based on a combination of factors in the natural and built environments (Morss et al. 2005).

Therefore, there is the opportunity to integrate data from multiple sensor sources to increase situational awareness for these events. This application example is also chosen due to the availability of post-event data for validation of the approach for the flood events impacting Louisiana in August 2016. Flooding during this time resulted in a FEMA disaster declaration in 26 parishes on August 14, 2016, an estimated \$30 million in relief efforts from the American Red Cross, and over \$110 million of estimated losses for Louisiana's agriculture industry (Van Der Wiel et al. 2017).

The data sources selected for integration for event estimation are stream gages (a physical sensor type) from the United States Geological Survey (USGS) and tweets (microposts) from Twitter (a social sensor type). Prior probabilities of flood events in each parish are derived from FEMA flood risk maps. The results are validated by comparing the dates and locations of updated flood risks to the true flooding that occurred in Louisiana in August 2016.

In the subsections that follow, each step in obtaining updated estimated flood risks based on collected data is explained within the structure of the proposed framework. Steps referred to specifically correspond to the steps of this data integration framework summarized in Section 4.4.

5.4.1 Prior Probabilities and Data Source Identification

For the application, let θ be defined as flood occurrence in a single parish (county) in Louisiana and $\bar{\theta}$ indicate nonoccurrence of a flood in the same parish with $P(\bar{\theta}) = 1 - P(\theta)$. FEMA Flood Insurance Risk Maps are used to derive a prior probability of risk, $P(\theta)$, and its complement for all 64 parishes. These maps, which are part of the National Flood Insurance Program, designate zones of the United States that are likely to be inundated in a flood event (Burby 2002). The base-level flood considered is the 100-year flood, or flood with a 1% probability of occurrence in a given year. The regions that will be inundated during the base flood event are named Special Flood Hazard Areas (SFHA) and are highlighted on the flood risk maps.

Figure 4 shows an example flood risk map for Ascension Parish in Louisiana. This screenshot is from a publicly available interactive map developed by the Louisiana State

University Agricultural Center, which allows users to view flood risks in Louisiana by parish (LSU AgCenter and LADOTD 2017). 64% of Ascension Parish is covered by a SFHA (shaded in the figure) and will be inundated in a 100-year flood event. The remaining area of each parish outside of the zone is expected not to be inundated in such an event.

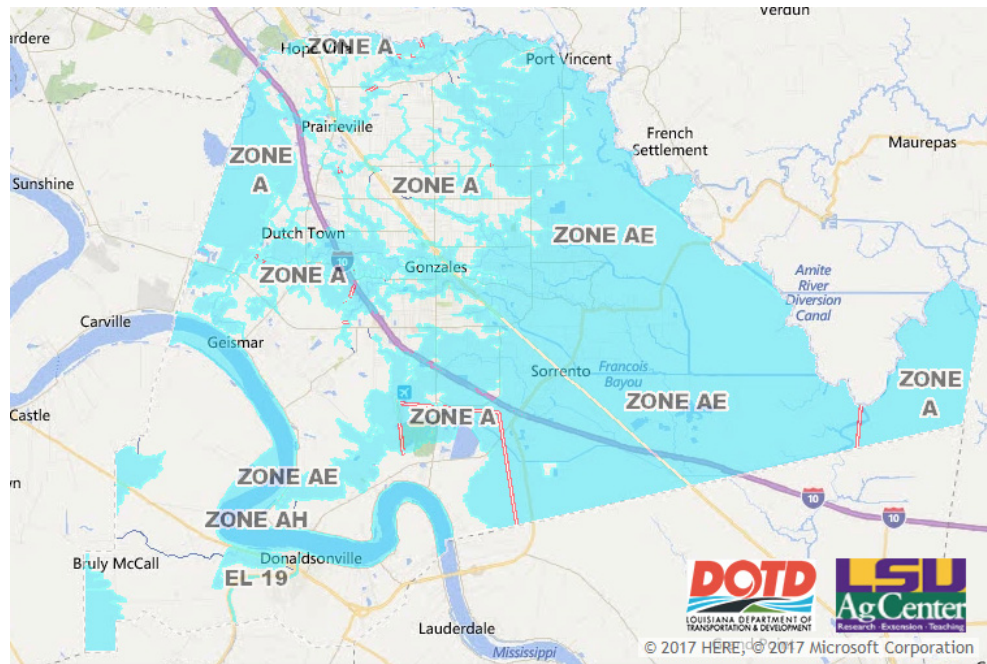


Figure 4. Example flood risk map with 100-year floodplain shaded for Ascension Parish, Louisiana (LSUAgCenter and LADOTD 2017)

An example event of interest $\theta_{Ascension}$ is flood occurrence in Ascension Parish. The prior probability of this event is assumed to be 1% (the probability of flood inundation in a given year in a SFHA) multiplied by the fraction of the parish area covered by the SFHA. For Ascension Parish, this results in a probability of 0.64% for a prior knowledge of flood risk, and this is defined as $P(\theta_{Ascension})$. This process is repeated for all 64

parishes in Louisiana resulting in a range of prior probabilities of 0.05%-0.88% across all parishes. As with all probability estimations, there are potential errors in defining these priors. However, these represent typical assessments of flood risks in the United States, so they are considered sufficient to set an initial estimation of risk for each parish.

For data integration, data from two publicly available sources ($k = 2$) are selected that, with some uncertainty, indicate a flood within each parish. The first data source (s_1) is physical sensor data from USGS, which monitors the conditions of the nation's streams and rivers with near-real-time data from stream gages. A sudden increase in gage height is selected as the measurement of interest to provide updating information on the probability of flood occurrence for each parish. The second data source considered (s_2) is social media big data from Twitter. The data collected from Twitter includes the texts of tweets and metadata such as date, time, and tweet location, if available. Tweet relevance to flood events is selected as the metric of interest, which is determined using a machine learning classifier, built using the social media data processing framework as described in Chapter 3.

These sources are also selected to demonstrate how the data integration framework can be used to integrate data from unrelated sources. By integrating information within each individual network before integrating data from across sources, this framework is able to consider unique data likelihoods for θ from each source.

5.4.2 Data Collection and Indications of Flood Events

Next, data is collected from the two sources and determine which observations from each source indicate a flood event for each parish.

5.4.2.1 Gage Height from USGS Stream Gages

From the USGS stream gages, data from all gages outputting gage height in Louisiana from June to August 2016 is collected. 235 stream gages in Louisiana provide downloadable gage height data during the period of interest. Gage height is reported at every 15- or 30-minute interval, depending on the gage. Figure 5 provides an example of available gage height data from a stream gage in Louisiana during a week in August.

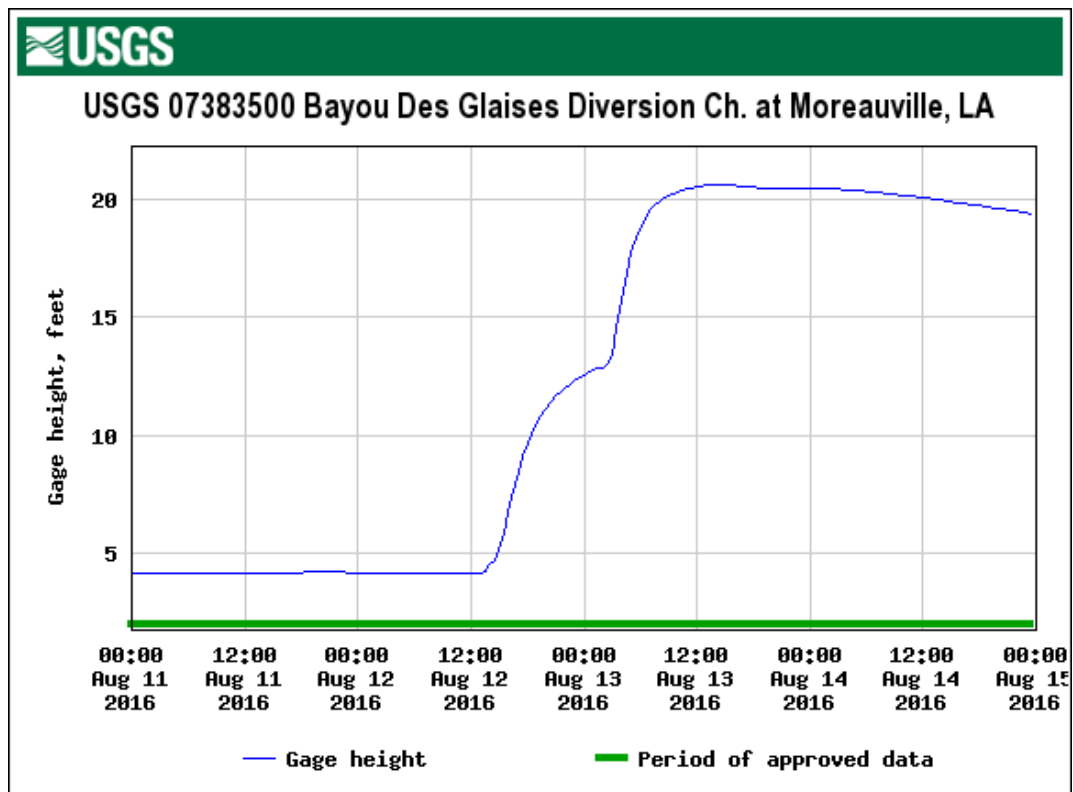


Figure 5. Example gage height data for a stream gage in Louisiana (Data from USGS 2017)

Full flood predictions are based on a combination of topology, rainfall, streamflow, and gage height data, as well as other hydrological and meteorological measurements. The objective of this study is to investigate the updating of prior estimations of event risk with data from multiple sources, rather than precise hydrologic modeling of flood systems. Therefore, a sudden increase in gage height is taken as the indicator of flood occurrence.

To find sudden increases, daily average gage heights are first computed over the period of data collection. From these averages, the percentage increase between each day is calculated. y_j represents the percentage increase from the previous day to the current day for the j^{th} stream gage. A percentage increase in daily average gage height y_j over 100%, i.e., the average gage height at least doubled from one day to the next, indicates a potential flood event at that stream gage. With this indicator of flood risk in the area surrounding a stream gage, the data is binarized to indicate a flood event in an area on a particular day if $y_j > 100\%$. For this example, data that does not indicate a flood event, where $y_j < 100\%$, is not considered, but such data can be incorporated if the likelihoods based on such an indication can be calculated. Here, the focus of this application is on the differences in updating from unrelated data sources, and only the $y_j > 100\%$ indication is chosen for the stream gage sensor source. The uncertainty of this indicator is accounted for when computing the overall likelihood of the sensor network, and the proposed framework allows for additional measurements and uncertainties from the same source indicating θ , such as multiple thresholds for a sudden increase in gage height. Using other thresholds for indicating a flood event (e.g., $y_j > 150\%$) would add or remove datapoints from the analysis, with the overall process to calculate posterior probabilities remaining the same.

From the data, 66 stream gages indicated a flood at least once during the three-month data collection period from June to August 2016, with several gages indicating floods multiple times within the period. These are the observations that indicate occurrence of the event of interest θ and are used for updating in the following steps. Figure 6 shows the locations of these stream gages in Louisiana and the dates on which they indicated a flood for the period August 10-13, 2016.

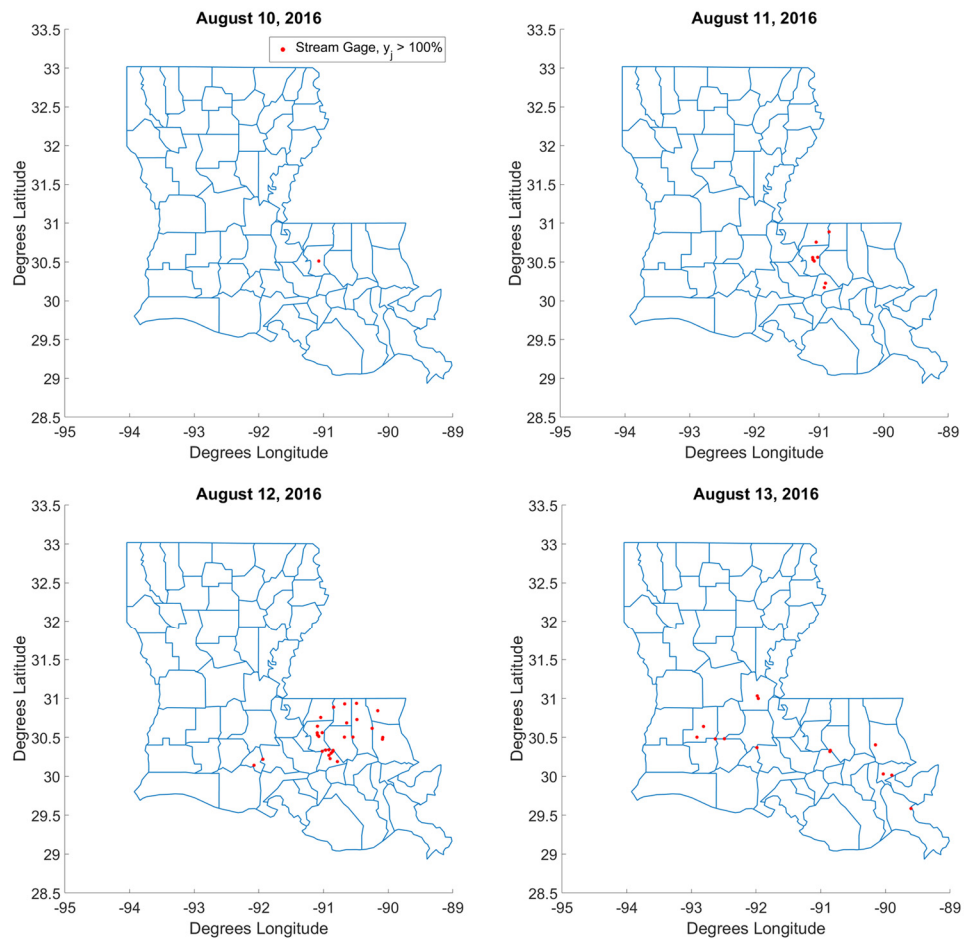


Figure 6. Stream gages in Louisiana indicating a flood, August 2016

5.4.2.2 Tweet Relevance to Flood Events

The Twitter data for this application is downloaded using Twitter's Streaming API and by scraping historical tweets from Twitter's search page. To determine which observations from this data source indicate a flood event in each parish, a machine learning classifier is built to predict a tweet's relevance to a flood event, defined z . The prediction is binary: a tweet is either relevant ($z = 1$) or irrelevant ($z = 0$) to a flood event. The classifier is then used to determine which tweets from the period of interest are relevant to flood events. After the tweets are classified, they are sorted by location.

The classifier is a support vector machine (SVM) model built using Weka. To build the model, a small training set is first compiled by manually labeling tweets as relevant or irrelevant to flood events. This training set is assembled with a diverse set of tweets that refer to many different flood events and also unrelated phenomena or events to ensure the model generalizes and classifies relevance for tweets regarding any flood event. The full database is comprised of tweets downloaded to contain initial search terms related to CIS damages and hazard events and is described in Section 5.2.1. In this application, tweet relevance to a flood event is defined as tweets referring to current flood events, excluding updates about flood recovery efforts and expressions of sympathy from others. For this training and testing set, tweets collected during the analysis period, August 10-13, 2016, are not included. Just as tweets collected during Tropical Storm Irma in Atlanta are used as unlabeled data for a detection scenario in Section 5.3, tweets collected during the data integration period are left unlabeled for this classification task, i.e., flood event identification.

The final training set for this application consists of 496 tweets. In the training set, 125 tweets are relevant and 371 are irrelevant. A small testing set is also compiled to evaluate the model's performance at classifying new tweets, which is next used to determine this sensor source's likelihood. The testing set contains 214 tweets, manually classified with 55 relevant and 159 irrelevant tweets. Table 11 shows the confusion matrix for the model evaluated on the test set. The model's classification of a tweet is denoted by z , and the true value of a tweet's relevance is denoted by λ , where 1 indicates true relevance and 0 indicates true irrelevance.

Table 11. Confusion matrix for support vector machine classifier tested for tweet relevance to flood events

Classified as $z = 1$	Classified as $z = 0$	True Class
34	21	$\lambda = 1$
7	152	$\lambda = 0$

The confusion matrix shows the performance of the model for evaluating the test set. The top row ($\lambda = 1$) shows 34 tweets were correctly classified as relevant, while 21 tweets were incorrectly classified as irrelevant. For $\lambda = 0$, the second row shows that 7 tweets were incorrectly classified as relevant although they were truly irrelevant. True positives are the tweets correctly classified as relevant by the model (the top left corner of the confusion matrix). Recall is calculated as 0.618.

This represents a measure of accuracy expressing the likelihood of a correct tweet classification given true relevance to a flood, also represented by the conditional probability, $P(z = 1|\lambda = 1)$. The recall of the model in automatically classifying the testing set is used as the measure of the model's reliability. This value is used to calculate the likelihood of tweet relevance collected from Twitter in the next step of the updating process.

Using the built classification model, unlabeled tweets from August 10-13, 2016, filtered for the word “flood” are classified as relevant or irrelevant with the accuracy discussed above. As the event of interest, θ , is flooding in a parish, tweets are next filtered by location, if that metadata is available, to categorize them by parish in Louisiana. If location is not available, the attributes (i.e., words) of each tweet are compared to a comprehensive list of municipalities and their respective parishes in Louisiana (Smith 2005). If the tweet contains one of the cities or towns on the list, it is considered a tweet relevant to a flood event for that parish. Of course, this means some tweets that are found may not truly be located in Louisiana (e.g., Iowa, Louisiana, is a town in Calcasieu Parish, but tweets found mentioning Iowa typically refer to the state of Iowa). Uncertainty in the data is accounted for by calculating the sensor source likelihood in the following section.

This process ultimately results in a list of tweets relevant to flood events for each parish. Table 12 shows an example of one of the tweets found relevant to a flood event in East Baton Rouge and Tangipahoa Parishes on August 12, 2016. There is no location attached to the metadata of this tweet, so its associated locations are determined by searching the text for municipalities from the aforementioned list. The text of the tweet indicates Baton Rouge, LA (East Baton Rouge Parish), and Hammond, LA (Tangipahoa

Parish). The tweet is therefore categorized to update the state estimations for these two parishes.

Table 12. Example of tweet classified as relevant and containing location indicators for Louisiana

Date	Tweet Text
August 12, 2016	“Move to higher ground! Flash Flood Warning continues for Baton Rouge LA and Hammond LA until 1:00 PM CDT”

5.4.3 Likelihoods of Data from Each Source

5.4.3.1 Likelihood of Gage Height Data

The daily likelihoods of $y_j > 100\%$ given a flood event in a parish θ , $P(y_j > 100\% | \theta)$, are now calculated, where y_j is the percent increase in daily average gage height for the j th stream gage and $j = 1, \dots, n$ for n stream gages indicating a flood event on the day in question. There is little validated empirical data that can be used to estimate the likelihoods of observing the data $y_j > 100\%$ given a flood event in a parish θ . Therefore, these likelihoods are calculated with a decaying function for $P(y_j > 100\% | \theta)$ that decreases in likelihood with each gage’s distance from the parish being considered. Figure 7 shows the function used in this application, where distance is

expressed in degrees of latitude and longitude. $P(y_j > 100\% | \theta)$ is assumed to be 1 when stream gage j is in the parish considered by θ , i.e., a stream gage will certainly show a daily average gage height that doubles from one day to the next given a flood in a parish if it is located in that parish. As the distance between the parish and stream gage increases, $P(y_j > 100\% | \theta)$ decreases. The distances are measured from each stream gage to the nearest point on the border of the parish considered by θ .

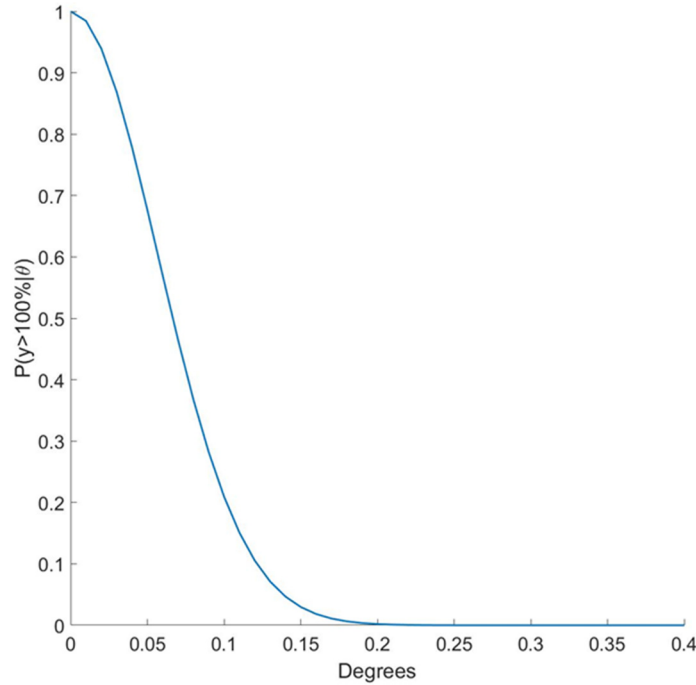


Figure 7. Decaying probability function with distance for stream gage sensor likelihood

As an example, on August 10, 2016, only one stream gage in Louisiana reported $y_j > 100\%$ where $j = 1$ as shown in Figure 6. That stream gage's distance away from

every parish was calculated and input into the function in Figure 7. The results are the likelihoods of observing $y_1 > 100\%$ given flood events in each parish. For all stream gages, it is assumed that any $P(y_j > 100\%|\theta)$ less than 0.01 is insignificant for updating and can be expressed as 0. Ultimately, there are $64 \cdot n$ likelihoods for each day, one from each stream gage, out of n , for each of the 64 parishes.

Next, the likelihood of observed data for all stream gages, $P(y > 100\%|\theta)$, is computed by combining the daily likelihoods of all stream gages indicating a flood in a parish. There is no assumption of independence between these likelihoods and the total probability theorem is used to integrate them. To do this, a variable, g_j , is introduced to represent the j^{th} stream gage out of all stream gages indicating a flood per day. $P(g_j)$ is simply $\frac{1}{n}$, so that each stream gage indicating a flood event is weighted equally. Let g_j be independent of θ and y_j , so $P(y_j > 100\%|g_j, \theta) = P(y_j > 100\%|\theta)$. Independence is created for this variable to facilitate the integration of likelihoods of gage height data. To eliminate this assumption of independence, more information on the combined likelihood of data from stream gages indicating a flood for each parish is needed. The likelihood of observed data for all stream gages is calculated as shown in Equation (7).

$$P(y > 100\%|\theta) = \sum_{j=1}^n P(y_j > 100\%|g_j, \theta)P(g_j) \quad (7)$$

If no data from other sources is available, as is the case for some parishes on August 10, 2016, the probability $P(y > 100\%)$ is computed, and the need to compute joint probabilities between multiple data sources in Step 6 is eliminated. To calculate $P(y > 100\%)$, first, $P(y_j > 100\%)$ is calculated for each stream gage empirically by

dividing the number of days the gage read $y_j > 100\%$ by the total number of days on which data was collected. These values are then combined with total probability using g_j to result in $P(y > 100\%)$.

The likelihoods of the stream gages indicating a flood given no flood occurrence in a parish are computed as shown in Equation (8). The expression is derived using total probability.

$$P(y_j > 100\%|\bar{\theta}) = \frac{P(y_j > 100\%) - P(y_j > 100\%|\theta)P(\theta)}{1 - P(\theta)} \quad (8)$$

In assessing this sensor source, including more or fewer datapoints by changing the threshold y_j from 100% to another value would change the overall likelihood depending on the likelihoods from the individual stream gages added or removed. For instance, if more stream gages are included in the analysis, but some of those stream gages are farther away from the parish of interest (i.e., with lower likelihoods), the overall likelihood will decrease even though more information is available. In other cases, decreasing the number of stream gages available may increase the final posterior probabilities if those that remain are in or very close to that parish, making the overall likelihoods of stream gage data at or close to 1. Due to the nature of the data collected, the individual likelihoods of each stream gage are important in calculating this specific sensor source likelihood. This is in comparison to the calculation of the Twitter data source likelihood as discussed in the following section.

5.4.3.2 Likelihood of Tweet Relevance to a Flood Event

Tweets indicating a flood event in a parish during August 10-13, 2016, are collected from Twitter and classified by the built SVM flood model. The likelihood of tweet relevance to a flood event, $P(z = 1|\theta)$, is calculated using the number of tweets indicating a flood in each parish. $P(z = 1|\lambda = 1)$ is the recall calculated by the performance of the classifier. The number of tweets available is an indicator of classification accuracy (Musaev et al. 2014). Therefore, in Equation (9), the recall accuracy metric is factored by the number of tweets indicating a flood in the considered parish N_P to compute the probability of tweet relevance given θ . This uncertainty is added to the likelihood calculation because most of the tweets are geolocated based on the presence of Louisiana city or town names in their texts, which does not guarantee correct categorization of indicated tweets by parish. Moreover, there are many fewer tweets available for each parish compared to the total number of tweets for each day. The classifier is therefore assumed to be less likely to predict the relevance of these specific tweets of interest correctly. In reducing the accuracy metric by $\frac{N_P}{N_P+1}$, the likelihood of tweet relevance in a parish is higher when there are more relevant tweets in that parish. That is, if there are more relevant tweets mapped to a parish, the higher the source likelihood, $P(z = 1|\theta)$, will be. In contrast with the overall likelihood calculations for stream gage data previously described, the likelihood of this social sensor source given a flood event in a parish does not depend on varying likelihoods of individual tweets due to the nature of the data collection and observations.

$$P(z = 1|\theta) = P(z = 1|\lambda = 1) \left(\frac{N_P}{N_P+1} \right) \quad (9)$$

To calculate the probability the model will classify any tweet as relevant, $P(z = 1)$, the total number of tweets classified as relevant on a day is divided by the total number of tweets collected on each day, regardless of location. $P(z = 1)$ is taken as the same value for all parishes. This value is necessary on August 10, 2016, when several parishes are referred to by tweets classified as relevant to a flood, but no stream gage data is available because those parishes are too far away from the only available stream gage indicating a flood on that day. Therefore, Twitter is the only available data source for those parishes, and the joint probability calculation in Step 6 is replaced by $P(z = 1)$, as is the case when only stream gage data is available and only $P(y > 100\%)$ is needed. Finally, $P(z = 1|\bar{\theta})$ is derived from total probability just as $P(y > 100\%|\bar{\theta})$ was calculated in Equation (8) to calculate the likelihood of data for the complement of θ .

The calculations and integration of data likelihoods within each network in this application is used to demonstrate the required information to use in Step 4 of the proposed framework. The specific calculations for this step will vary depending on the nature of the data collected for each source. For the example, the gage height data represents a source for which the likelihoods of individual observations can be integrated, while Twitter data represents a source for which the number of indications of θ and the accuracy of classifying individual observations can be used to obtain the overall likelihood of the source. It is acknowledged that the assumptions and simplifications made in the analysis of both data sources for indicating flood events may not include other factors used in more comprehensive flood modeling and detection. The data collection and integration of information from these sources in the example explicates use of the framework for data with different likelihoods.

5.4.4 Probabilities of Data, Integration of Data Likelihoods, and Final Updating

The data from each source, gage height and tweet relevance, are conditionally independent given the event of interest θ . The joint likelihood of data from both sources on each day is calculated using Equation (10).

$$P(y > 100\%, z = 1|\theta) = P(y > 100\%|\theta)P(z = 1|\theta) \quad (10)$$

The joint probability of observing the data from both sources is calculated using Equation (11), with information included on both states of a flood event in a parish: occurred, θ , and not occurred, $\bar{\theta}$.

$$\begin{aligned} P(y > 100\%, z = 1) &= P(y > 100\%|\theta)P(z = 1|\theta)P(\theta) + \\ &P(y > 100\%|\bar{\theta})P(z = 1|\bar{\theta})P(\bar{\theta}) \end{aligned} \quad (11)$$

For this application, data is also sequentially updated from August 11-12, 2016, and August 12-13, 2016. The probability of data from the two-day intervals is calculated using Equation (12). Subscript 1 refers to likelihoods calculated on one day, and subscript 2 refers to likelihoods calculated on the following day.

$$\begin{aligned} P(y_1 > 100\%, z_1 = 1, y_2 > 100\%, z_2 = 1) \\ &= P(y_1 > 100\%, z_1 = 1|\theta)P(y_2 > 100\%, z_2 = 1|\theta)P(\theta) + \\ &P(y_1 > 100\%, z_1 = 1|\bar{\theta})P(y_2 > 100\%, z_2 = 1|\bar{\theta})P(\bar{\theta}) \end{aligned} \quad (12)$$

For the application, the final posterior probabilities for each parish in Louisiana are computed using Equations (13) and (14), with Equation (14) used for sequential updating. The sequential updating here is limited to two days because with the amount of data

available during this time period, updating using more than two days of data results in posterior probabilities of nearly 1 in several parishes (i.e., nearly 100% probabilities of flood occurrence in a parish) with the remaining parishes receiving little to no updating.

$$P(\theta|y > 100\%, z = 1) = \frac{P(y > 100\%|\theta)P(z = 1|\theta)P(\theta)}{P(y>100\%,z=1)} \quad (13)$$

$$P(\theta|y_1 > 100\%, z_1 = 1, y_2 > 100\%, z_2 = 1)$$

$$= \frac{P(y_1 > 100\%, z_1 = 1|\theta)P(y_2 > 100\%, z_2 = 1|\theta)P(\theta)}{P(y_1>100\%,z_1=1,y_2>100\%,z_2=1)} \quad (14)$$

In cases when data from one source is completely unavailable (e.g., if there are no relevant tweets in a parish), the prior risks are updated with information from only the other source. When there is no data available from any source, the prior risk remains unchanged. In other applications with more than two sources, the joint probabilities of different combinations of sensor sources must be computed.

5.4.5 Results

5.4.5.1 Results Over Time

Using the proposed framework for integrating data across multiple sources, the resulting updated probability distributions for flood risk by parish are mapped in Figure 8 for each day of the four-day period of investigation. The lowest probabilities in light yellow show the parishes that had little or no data with which to update their prior probabilities of flood events. From Figure 8, few updated probabilities are computed on August 10. The highest updated probabilities of flood occurrence in a parish on each day from August 10-

13, 2016, are 0.05 (Orleans Parish), 0.45 (Ascension Parish), 0.73 (St. Tammany Parish), and 0.81 (Livingston Parish), respectively. The largest changes in prior to posterior risks occurred on August 12.

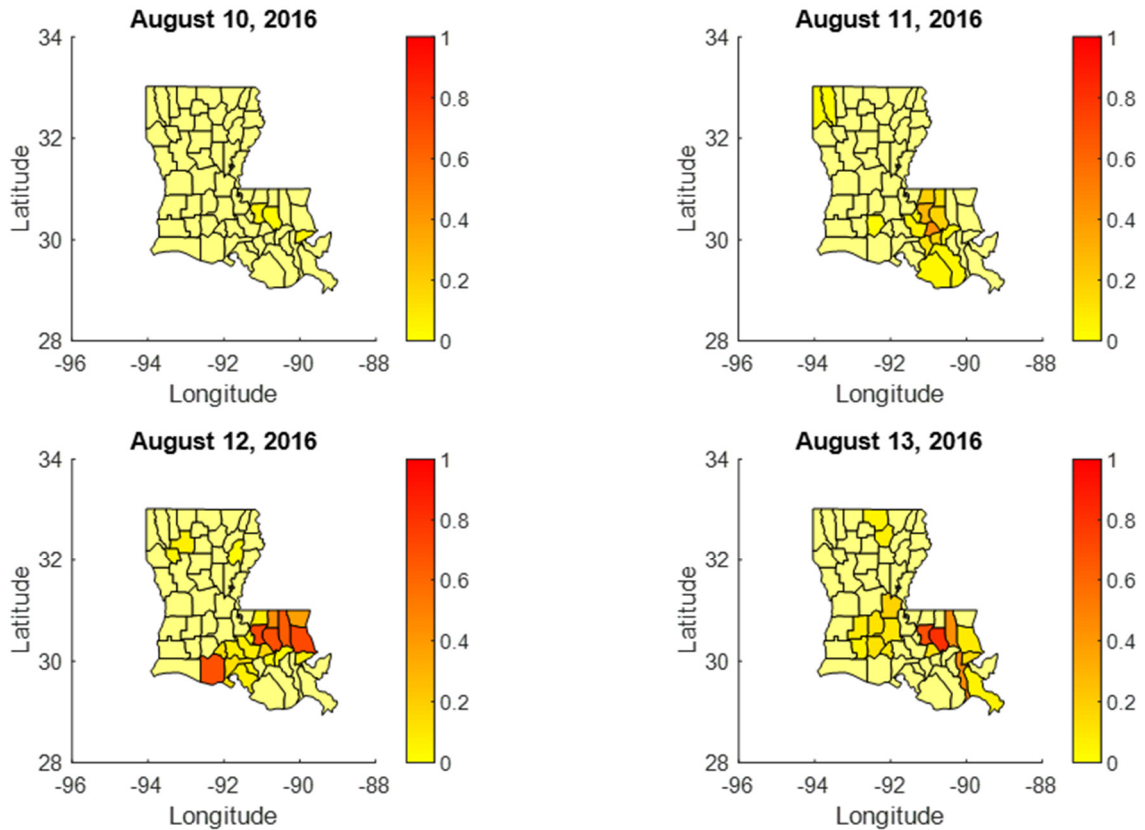


Figure 8. Resulting updated probability distributions for August 10, 11, 12, and 13

Figure 9 shows sequential updating results, from August 11 to 12 and from August 12 to 13, and the effect of combining data from multiple days. The largest updated probability of flood occurrence in a parish after sequential updating from August 11 to 12

was 0.998 (East Baton Rouge), and for sequential updating from August 12 to 13, 0.999 (East Baton Rouge and Livingston).

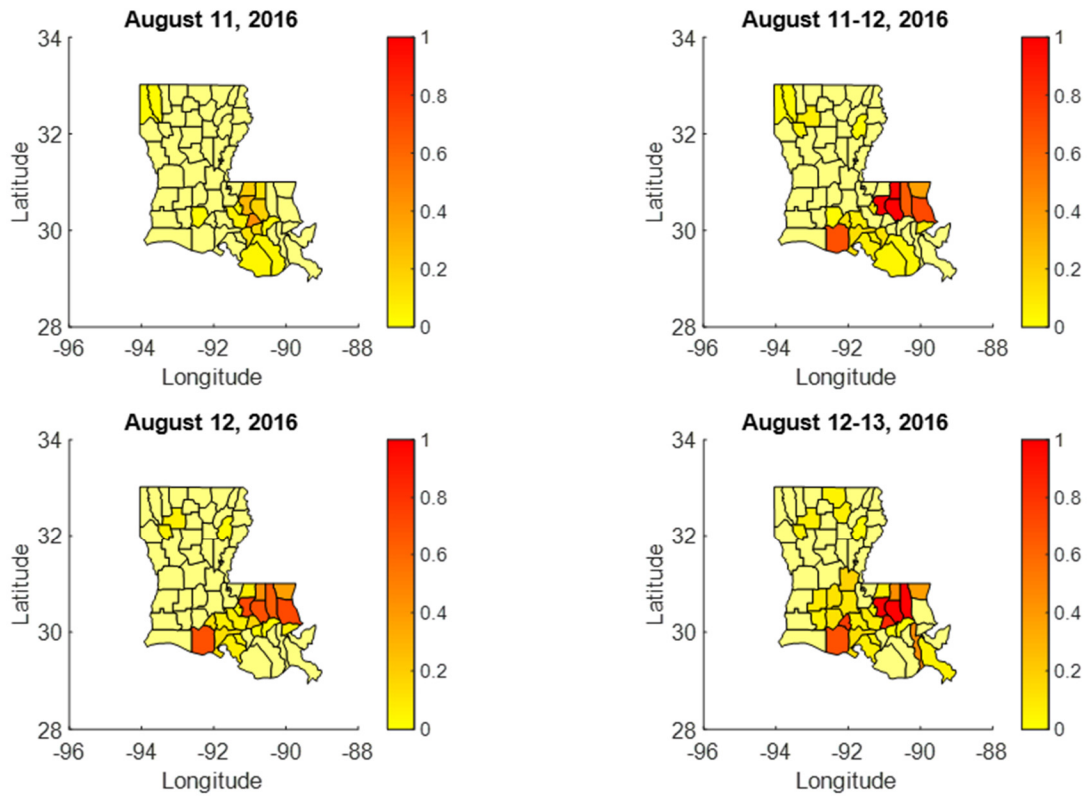


Figure 9. Updating and sequential updating results with sensor observations from August 11 to August 11-12 and August 12 to August 12-13

5.4.5.2 Results by Sensor Type

As an objective of this study is to integrate multiple data sources for event estimation, the effect of additional data on the estimation is investigated, specifically looking at the results using data from single sensor sources compared to combining the

chosen datasets. Figure 10 shows the results of updating probabilities of flood risk on August 12, 2016, by parish based on only stream gage data, only Twitter data, and then using information both sensor sources.

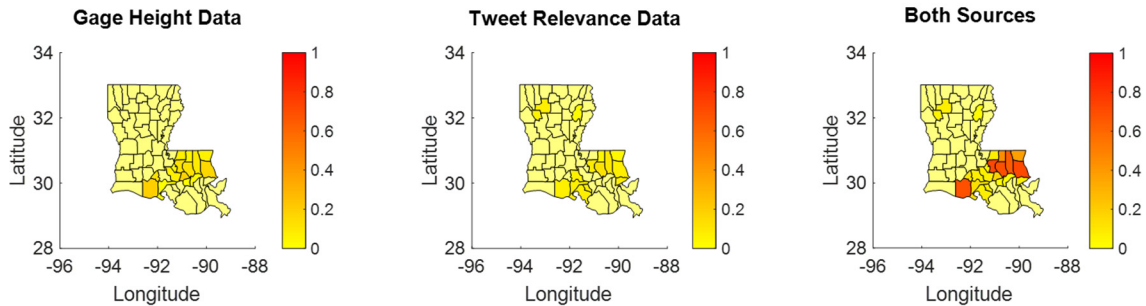


Figure 10. Resulting updating probability distributions for August 12, 2016, using gage height and tweet relevance data sources alone compared to estimation from integrating both data sources

Using information from both sources significantly increases the updated probabilities. For instance, in St. Tammany Parish, three stream gages indicated a potential flood event based on increased gage height and two tweets were classified as relevant to a flood event. The prior probability is computed to be 0.0065. The resulting updated probabilities are 0.17 and 0.08 for updating with gage height data and tweet relevance data alone, respectively. When data from both sources are included in the inference, the updated probability of flood risk is 0.73. This is seen for all parishes that have data with which to update their prior probabilities of flood risk; updating with an individual data source does not result in a posterior probability higher than 0.20 on August 12, while both data sources combined update probabilities up to 0.73. This is due to the joint probabilities of observing

data from multiple sources being smaller than the probabilities of observing data from one source alone. Therefore, even as more data is available from one source, the probabilistic estimations experience the most updating when multiple data sources indicate the same event θ .

5.4.5.3 Validation and Data Availability

The results of the approach and analyses are compared to the true flooding that occurred in August 2016 in Louisiana for validation. The floods caused a Major Disaster Declaration from FEMA, and 26 parishes were designated for Individual or Public Assistance, shaded in Figure 11 (FEMA 2016). The results in Figure 8 and Figure 9 show increased flood risks after updating in most of the parishes listed in the Major Disaster Declaration. Of the 26 parishes with a declared Disaster Declaration, 17 are updated based on the data from August 12, and an additional six parishes not updated on August 12 are updated based on data from August 13. Figure 12 shows parishes with updated probabilities in dark blue and parishes without data to update priors in green for August 12 and 13, and after sequential updating from August 12-13.

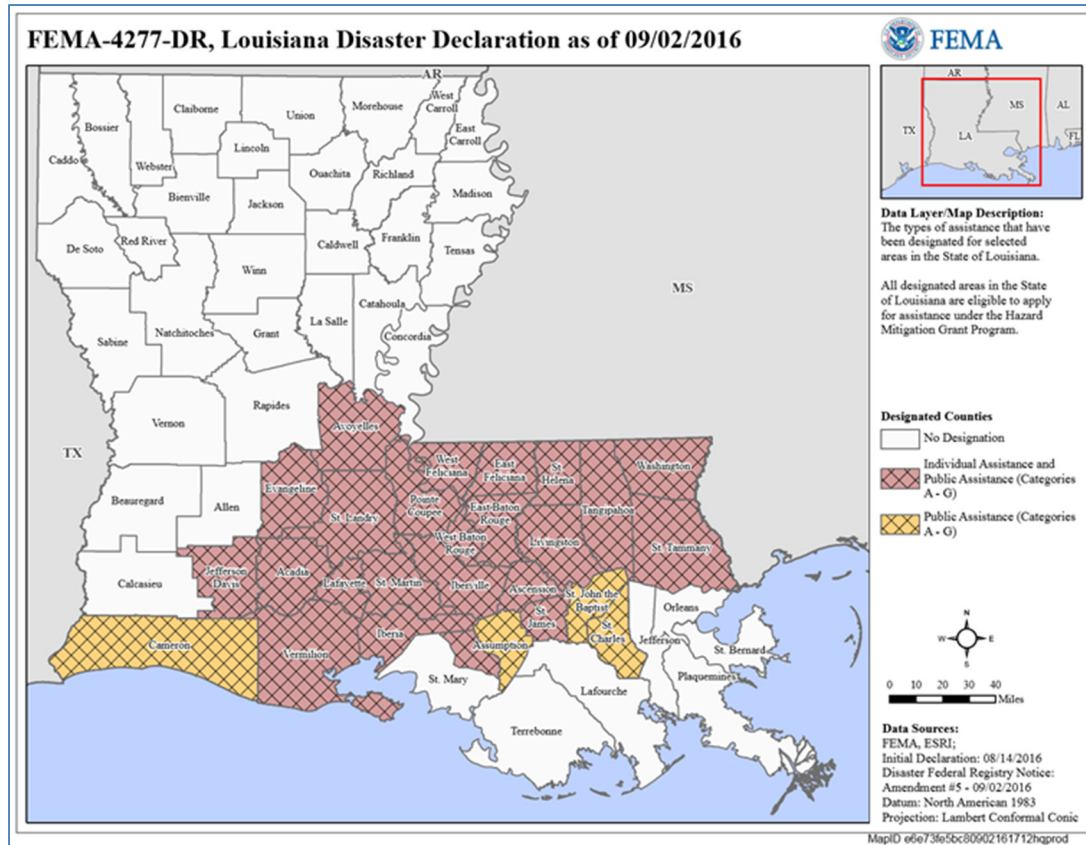


Figure 11. Louisiana Disaster Declaration for flooding event in August 2016, FEMA-4277-DR (Reprinted from FEMA 2016)

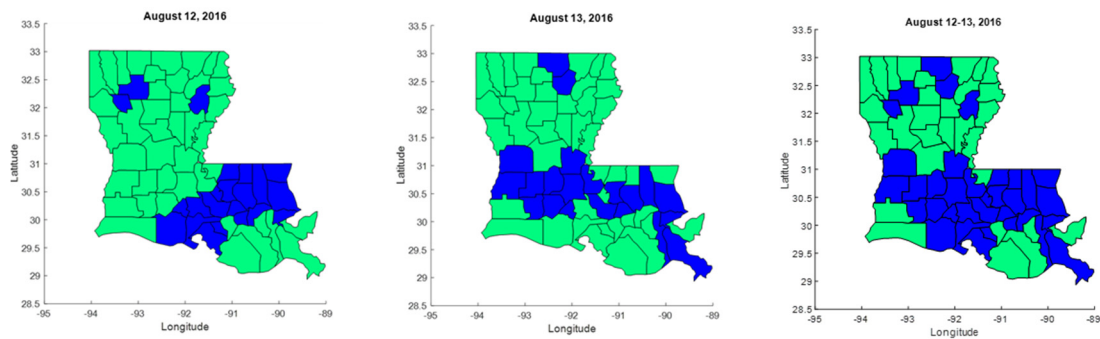


Figure 12. Parishes with updated probabilities of flood risk (dark blue) from integrated data sources on the dates shown

The updated probabilities also include nine parishes not in FEMA's Disaster Declaration. This can be explained by a number of factors. FEMA considers numerous variables when deciding on areas for Disaster Declarations. Some of these factors include localized impacts, insurance coverage in force, estimated cost of assistance, and other federal assistance programs (FEMA 2017). Ultimately, flooding may have occurred in parishes updated by data, even if they are not part of the Disaster Declarations.

For those parishes included in the Disaster Declaration that did not have updated probabilities of flood risk from the application, this is due to a lack of availability of data indicating a flood event for those particular parishes. The presence of data is essential in the framework to estimate the event of interest. For any application, different sensor sources or different measures for which sensors indicate the event of interest would create different results. Figure 13 shows the availability of data indicating θ over time for each parish. The number of available observations is shown, for counts of stream gages with sudden increases in gage height, number of tweets classified as relevant, and total daily datapoint counts from combining both sources.

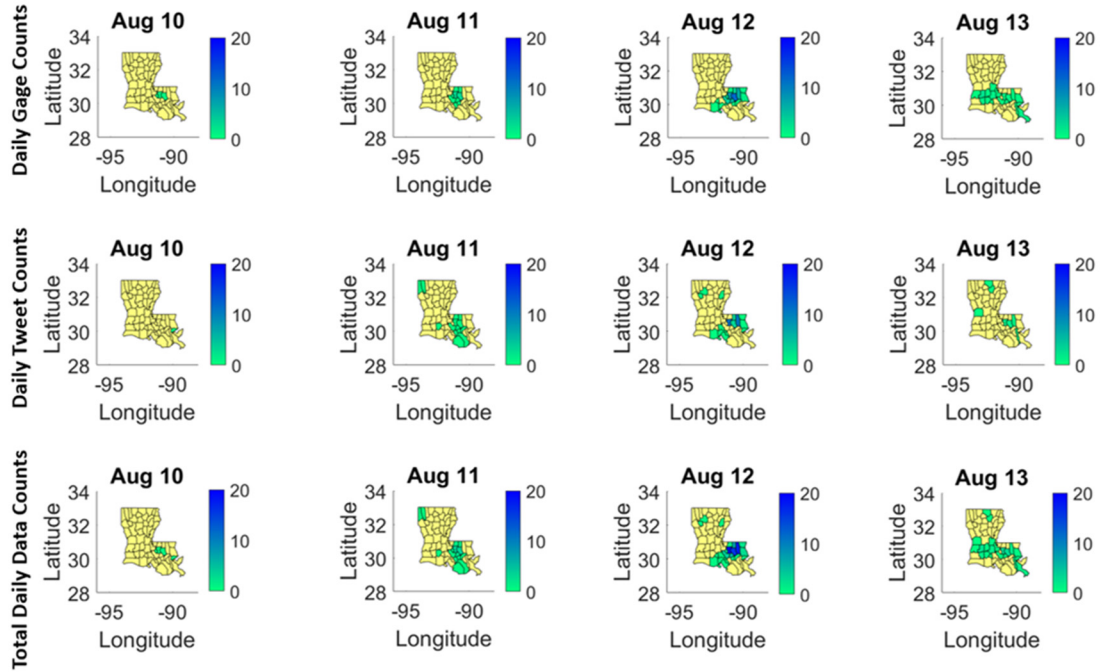


Figure 13. Stream gage and Twitter data availability by parish from August 10-13, 2016

To further validate the results, USGS post-event maps are examined to qualitatively evaluate the results of applying the framework to the application example. After the flood events in August 2016, USGS created a report to summarize the flooding and developed several flood inundation maps based on high-water marks. The report also included a map of cumulative rainfall across the state during August 11-14, 2016. This and the inundation map for Louisiana with shaded areas indicating inundation is shown in Figure 14 (Watson et al. 2017).

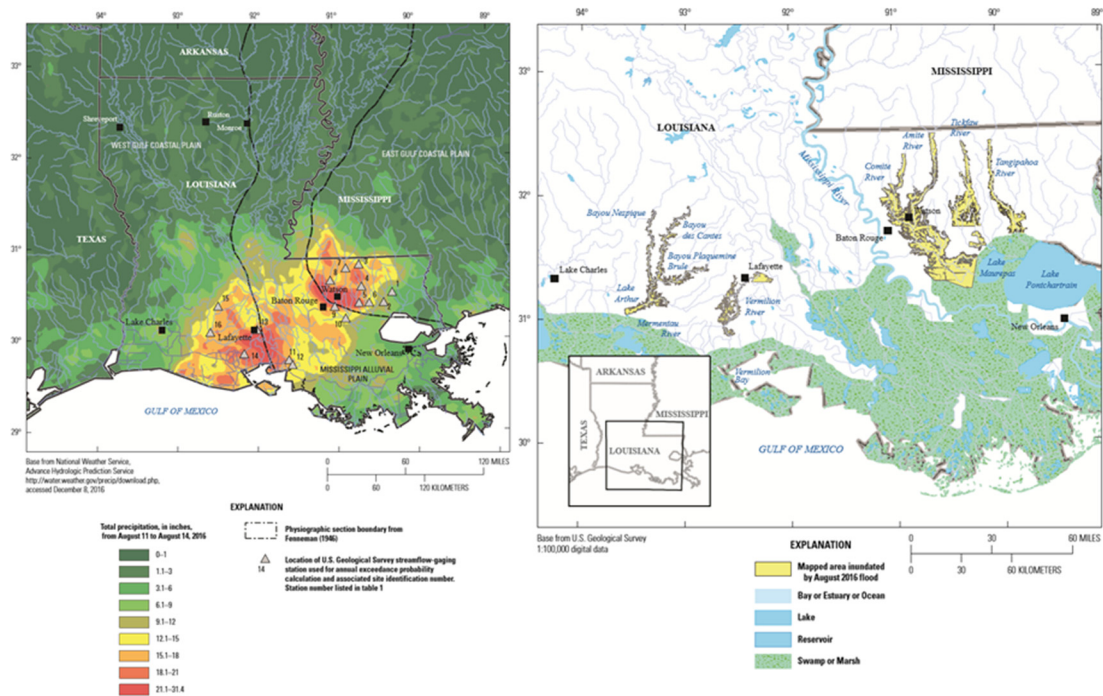


Figure 14. Precipitation in Louisiana from August 11-14, 2016 (left) and inundation map for Louisiana (right) (Reprinted from Watson et al. 2017)

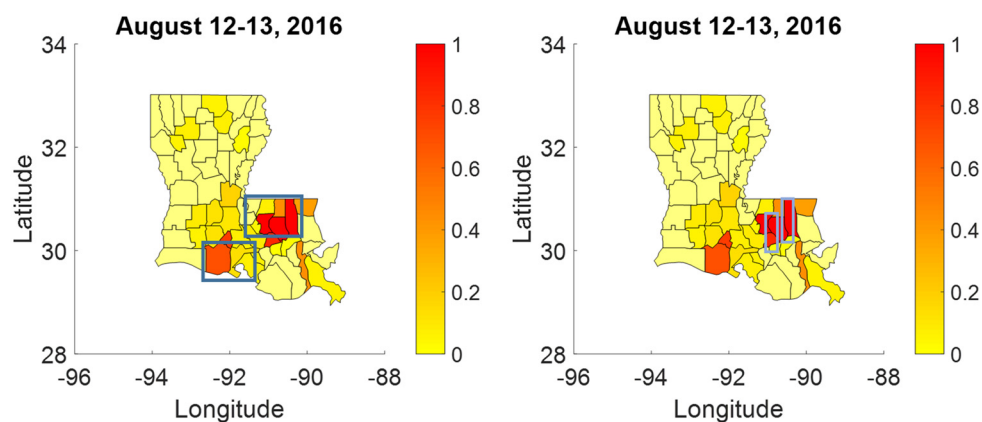


Figure 15. Map of sequential updating from August 12-13, 2016, with regions of highest precipitation (left) and confirmed inundation (right) from Figure 14 highlighted

Specifically, extensive inundation around the Amite and Tangipahoa rivers affected parishes in the northeastern portion of Louisiana’s “boot.” These include three of the parishes with the highest updated probabilities of flood occurrence using the proposed approach: East Baton Rouge Parish, Livingston Parish, and St. Tammany Parish. These correspond to the areas of highest precipitation during the time period investigated and with areas of confirmed inundation. The integrated data during August 10-13, 2016, updates the prior probabilities of flood occurrence from 0.0055, 0.008, and 0.0064 to 0.73, 0.81, and 0.73 for East Baton Rouge, Livingston, and St. Tammany parishes, respectively. Figure 15 shows the results of sequential updating for the application from August 12-13, 2016, with regions of highest precipitation (left) and verified inundation (right) from Figure 14 highlighted. The regions of highest precipitation (shown in dark orange and red in Figure 14, left) correspond to the boxed parishes in Figure 15 (left). The inundation around the Amite, Comite, Tickfaw, and Tangipahoa Rivers mapped in Figure 14 (right) correspond to the boxed region in Figure 15 (right).

5.5 Application of Proposed Data Integration Framework for Estimating Transportation Damage During Tropical Storm Irma in Atlanta, Georgia

The proposed data integration framework is also applied to predict transportation damage on an intersection (i.e., at two roadways) in Atlanta, Georgia, during Tropical Storm Irma on September 11, 2017. Here, the event of interest, θ , is this specific roadway in a transportation network being blocked or impassable during the hazard event. This transportation system component is selected because of its proximity to a major water supply for the city, and this same water supply is part of the larger system evaluated in Chapter 6. The purpose of this application is to demonstrate potential both of the social

media data processing framework and the data integration framework to provide new, real-time information for a network model to predict overall system behaviors and vulnerabilities.

The transportation component for analysis is an intersection of two roadways in Atlanta, Georgia. These roadways are shown in bold in Figure 16, among surrounding streets and roads which are not in bold. The intersection's proximity to a water supply component is also shown in Figure 16. The water supply is labeled as Supply 4 and shown as a white circle for consistency with the network analyzed in Chapter 6. The black line represents a link (i.e., pipe) from Supply 4 to another component in the water network. An uninformative prior probability of component failure during the hazard event (e.g., the roadway is blocked or flooded) is chosen as $P(\theta) = 0.01$.

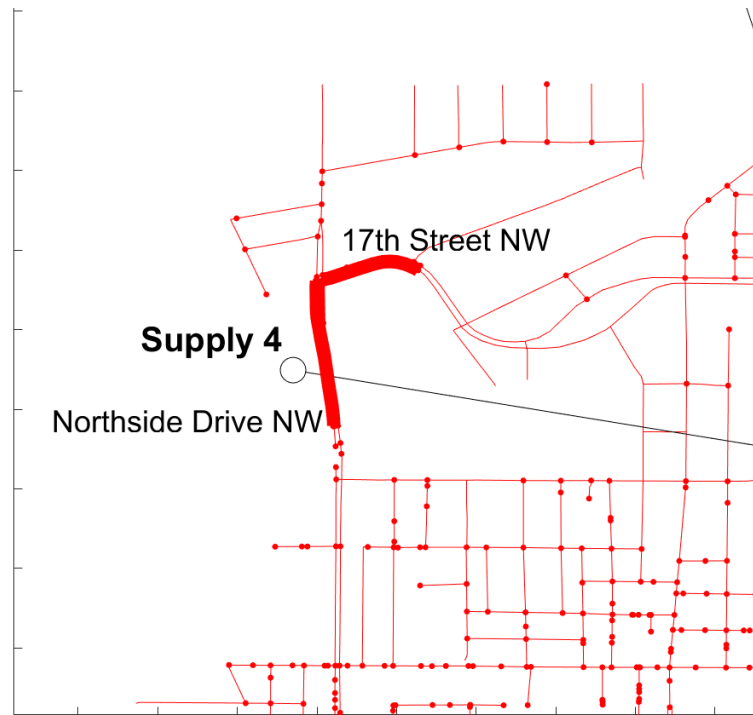


Figure 16. Roadways needed to access a water supply (Supply 4) for repair in Atlanta, Georgia

Twitter data and USGS stream gages are selected as data sources for integration, and similar likelihood calculations as from Section 5.4 and classifier as from Section 5.3 are used for the analysis. From USGS, gage height from a single stream gage on the day of analysis is collected and observed because of its location compared to the CIS components of analysis. The stream gage indicates potential damage to the transportation component by showing a sudden increase in gage height on the day of analysis. As in Section 5.4, a sudden increase is calculated as a greater than 100% increase from one day to the next. While this indication is used to calculate the likelihood of data for a different event, flood occurrence, the same indication is used here to demonstrate how sensor sources can be used to observe different events, including hazard occurrence and CIS damages. More

sophisticated urban flood models can replace this to determine accurate indications of flooded roadways given meteorological or hydrological data. The data downloaded for this stream gage is shown in Figure 17. The likelihood of observing gage height data indicating damage to the transportation component is a function of the stream gage's distance away from the transportation component. The same distance function used in Section 5.4.3 is employed for this application to compute the stream gage likelihood. The location of the stream gage compared to the transportation component is shown in Figure 18.

From Twitter, tweets collected and classified as relevant to transportation damage during this event in Section 5.3 are used for updating. Of these, tweets indicate potential damage for the transportation component if they are located in or specifically reference Atlanta. Ultimately, 14 tweets that are classified as relevant by the transportation SVM model in Section 5.3 contain references or geotags to Atlanta or local Atlanta communities. An example of one such tweet is shown in Figure 19. The likelihood of observing the Twitter data collected is the accuracy of the classifier (i.e., the 0.68 precision calculated in Section 5.3) factored by the number of relevant tweets available, as done for the likelihood of tweets indicating a flood in Section 5.4.3.

The final updated probability of failure for the roadway is 0.25 from the collected data from both sources, which can be directly used as inputs to probabilistic network models, such as the BN model used in Chapter 6. Possible extensions of this work are to collect data from additional sources to include in the updating process and to estimate damage on other transportation components during the same event for comparison.

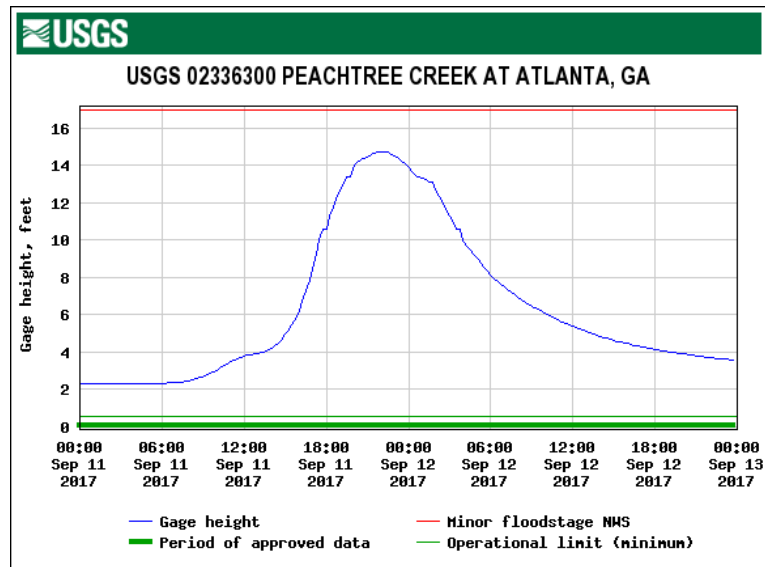


Figure 17. Example gage height data for a stream gage in Atlanta, Georgia (Data from USGS 2017)

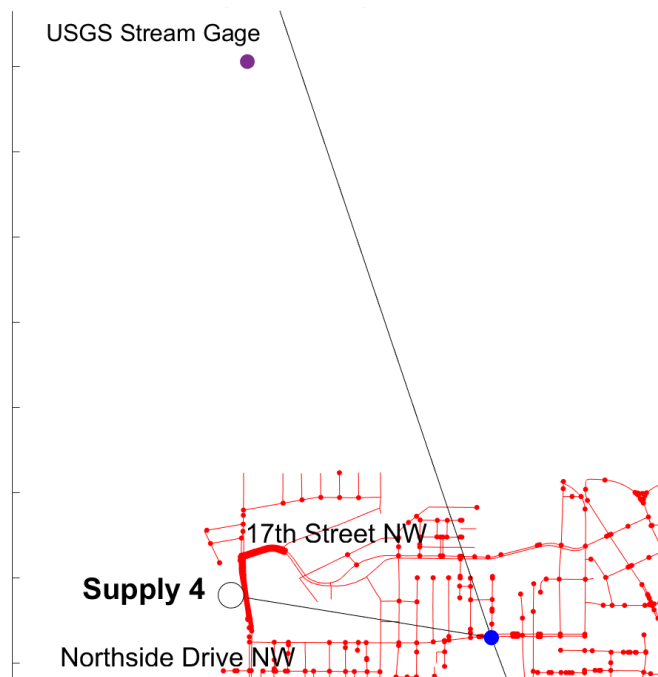


Figure 18. USGS stream gage and transportation component shown



Mike McClain FOX 5 @mikemcclainfox5 · 11 Sep 2017
More **trees down** in Atlanta. This one at Peachtree Park in **Buckhead**
[@FOX5Atlanta](#) [@FOX5StormTeam](#) [#Irma](#)



17



19



Figure 19. Example tweet related to transportation damage in Atlanta, Georgia during Tropical Storm Irma

CHAPTER 6. PROBABILISTIC SYSTEM MODEL ANALYSES

6.1 Introduction

In this chapter, probabilistic system model analyses are presented to evaluate the impact of network parameters and topology on overall vulnerability and resilience. In a single network, Atlanta's water network, parameters are varied and system performance (i.e., vulnerability) is evaluated under those variations. System performance is also compared across multiple, small-size water networks to further investigate the impact of network topologies and connectivities on overall system vulnerability. All analyses are conducted with a Bayesian network (BN) model, described in Chapter 2 and again in this chapter.

6.2 Atlanta Water Network, Varied Parameters, and Network Model

This study first investigates three parameters of Atlanta's water distribution network including service provision and access for repair interdependencies across power and transportation systems. Geographic dependencies are accounted for through common hazard exposure in nearby nodes. The schematic of the CIS and its supply and distribution nodes is shown in Figure 20. Electrical power is needed at the supply nodes and roadways are located across the area, which are required to access the water supply nodes for repair operations. A representation of the BN model is shown in Figure 21, with water supply, water distribution, electrical power, and access for repair (e.g., roadway) nodes indicated. MLS components connecting multiple distribution components (i.e., Distr. Comp 1 and Distr. Comp 2) to supply components 1 and 2 are also shown.

The parameters of a CIS are defined as including aspects of the physical network structure (e.g., the layout of links, or pipes, between water distribution components), service provision interdependencies within and across systems, and parameters of specific components in the system (e.g., probabilities of component failure given hazard occurrence). The effects of varying the following parameters on overall system behavior are evaluated: conditional probabilities of failure for water supplies, power supplies, and access for repair components; the number of power supply nodes in the network; and the configuration of links in the water network. Changes to each of these parameters affect overall system performance. This study focuses on the impacts of network changes on predicted performance at the distribution components of the water distribution system in evaluating the results of the inferences.

Two types of inferences with varied parameter values are conducted for analysis. First, inferences are conducted with no evidence. These inferences result in the marginal probabilities of failure for all water, power, and transportation components represented in the CIS BN. The outcomes are then compared across the varied parameter values to detect trends in component performance. For these results, decreased marginal probabilities of failure for distribution components indicate decreased likelihoods of outages at those points providing CIS services. Increased marginal probabilities of failure correspond with increased likelihoods of service outages. Second, inferences are conducted under a hazard occurrence or specific water or power supply component failure. The outcomes from these inferences correspond with predictions of CIS performance under failure or hazard scenarios. The results show the effects of outages in the power and water systems on the

final states of distribution components and the impacts of varying network parameters on system resilience, particularly the ability of the system to withstand such a disruption.

The component-level results are evaluated according to four characteristics of the distribution components to assess the types of components for which changes have the largest impacts. Distribution components are linked to supply nodes, with MLSs indicating the minimum sets of components needed to deliver the CIS resource from a supply to distribution node. The characteristics considered for each distribution component are the number of water supplies linked to the distribution component, shortest physical distance from a supply to the distribution component in terms of pipe length, number of MLSs for the distribution component (i.e., number of minimum paths from any supply to the distribution component), and minimum number of links from the distribution component to a supply (i.e., number of links in the shortest MLS).

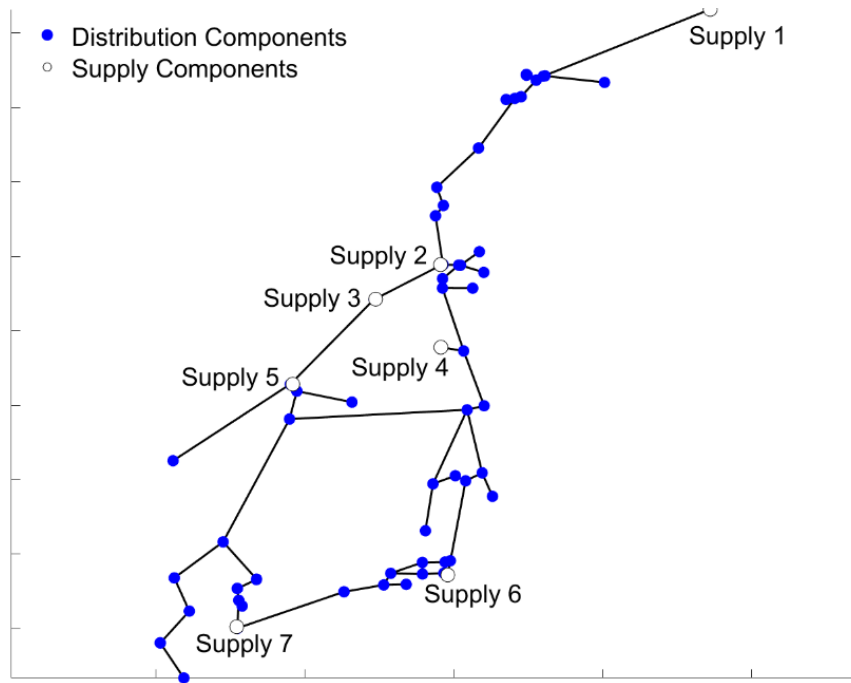


Figure 20. Schematic of Atlanta's water distribution system comprising supply and distribution components

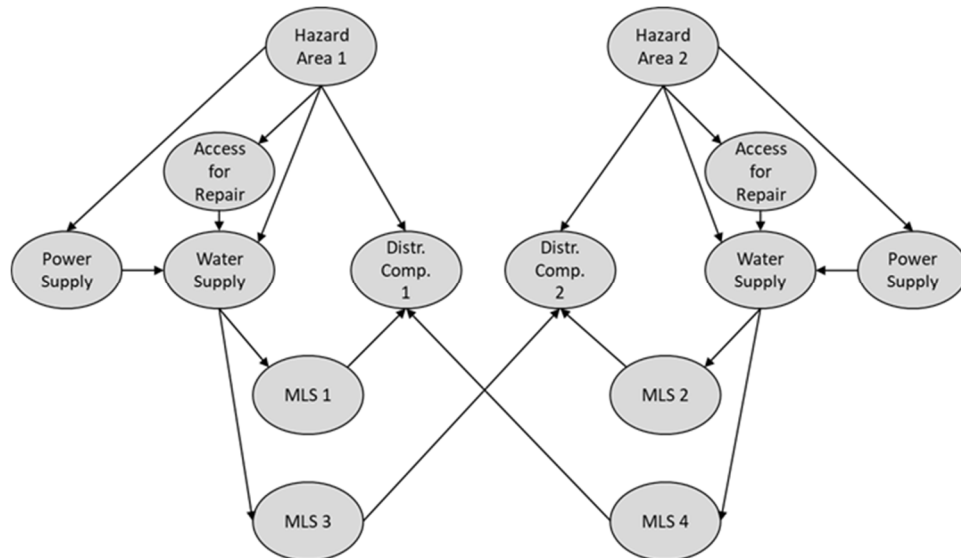


Figure 21. Representative Bayesian network model of interdependent CIS

6.3 Component Conditional Probabilities of Failure

The infrastructure components in the CIS BN comprise water supply, water distribution, power supply, and access for repair transportation components such as roadways required for crews to access failed water system components for repair. The baseline network for comparison assumes a 0.01 conditional probability of failure given hazard occurrence for all water distribution, water supply, power supply, and access for repair components. The consistent 0.01 value enables relative comparisons of performance across components. The baseline network also assumes a 0.01 probability of hazard occurrence across the system. The 0.01 values are within the ranges of typical probabilities of component failure or hazard used in engineering risk applications but can be changed with any additional information at the component or hazard levels. The system outcomes from using the 0.01 probability values are the baseline to which all other results are compared for this network.

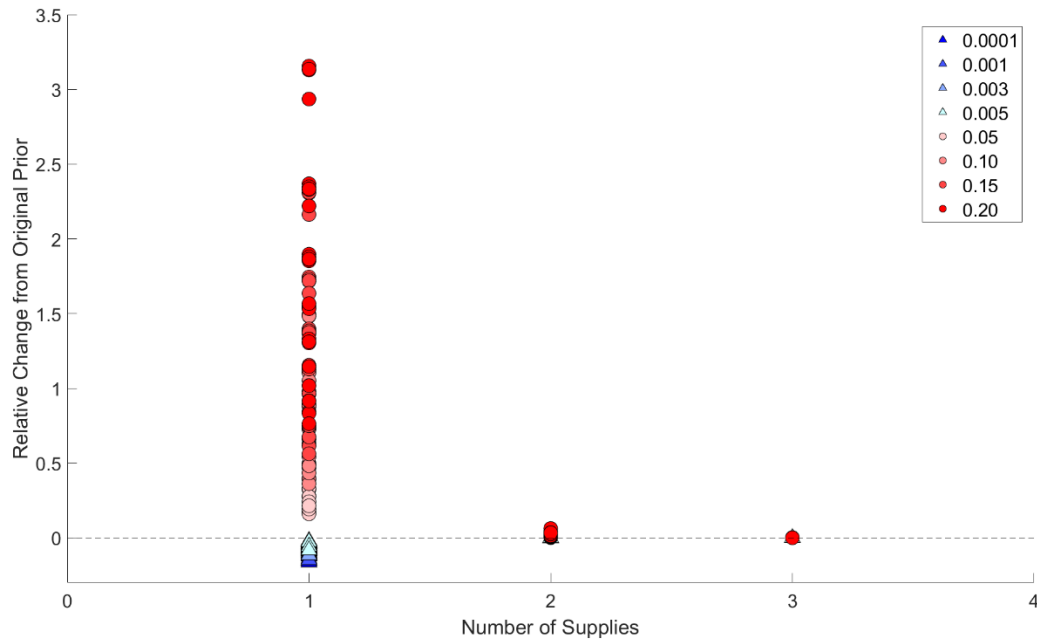
The conditional probabilities of failure are varied given hazard occurrence for the water supply, power supply, and access for repair nodes in the CIS BN. These variations correspond with changes that can occur as the system evolves over time. For example, decreasing the conditional probability of failure of a component can occur from retrofit upgrades or other preventive measures taken to increase component-level resilience. On the other hand, an increased probability of failure given hazard occurrence indicates a damaged or aging component, including a component that has not been upgraded or fully repaired after a less severe hazard event.

In the analysis, the supply and access conditional probabilities of failure given hazard are varied from the baseline 0.01 down to 0.0001, 0.003, and 0.005, and up to 0.05, 0.1, 0.15, and 0.2. The maximum is set at 0.2, as it is assumed that any component with a greater than 0.2 probability of failure given a hazard will no longer be operable, i.e., it will be out of service and repaired. The conditional probabilities of failure are modified by changing the CPTs of the supply and access nodes. Inferences are conducted at each parameter value to calculate marginal probabilities of failure for all components. The relative change, RC , in marginal failure probability from the baseline for a component is computed as in Equation (15). $p_{f_{old}}$ and $p_{f_{new}}$ are the original marginal probability of failure calculated with the baseline 0.01 failure probabilities, and the new marginal probability of failure calculated with the varied component conditional probabilities, respectively.

$$RC = \frac{p_{f_{new}} - p_{f_{old}}}{p_{f_{old}}} \quad (15)$$

Figure 22 to Figure 25 show the relative changes in marginal probabilities from the original baseline results for all distribution components varying the conditional probabilities of failure for power supply components. Each figure gives the relative change against a particular component characteristic. These characteristics are the number of reachable supplies to a distribution component (Figure 22); the shortest physical distance between a distribution component and a supply as measured by pipe length (Figure 23); the number of MLSs indicating the number of minimum paths connecting a distribution component to a supply (Figure 24); and the minimum number of links to a supply, representing the fewest number of components in any MLS for the distribution component

(Figure 25). A line at 0 relative change from the original prior marginal probability of failure is indicated. Inference results for each varied value of component conditional probability of failure from 0.0001 to 0.2 is represented by a different marker. Positive relative changes indicate increases in marginal probabilities of failure.



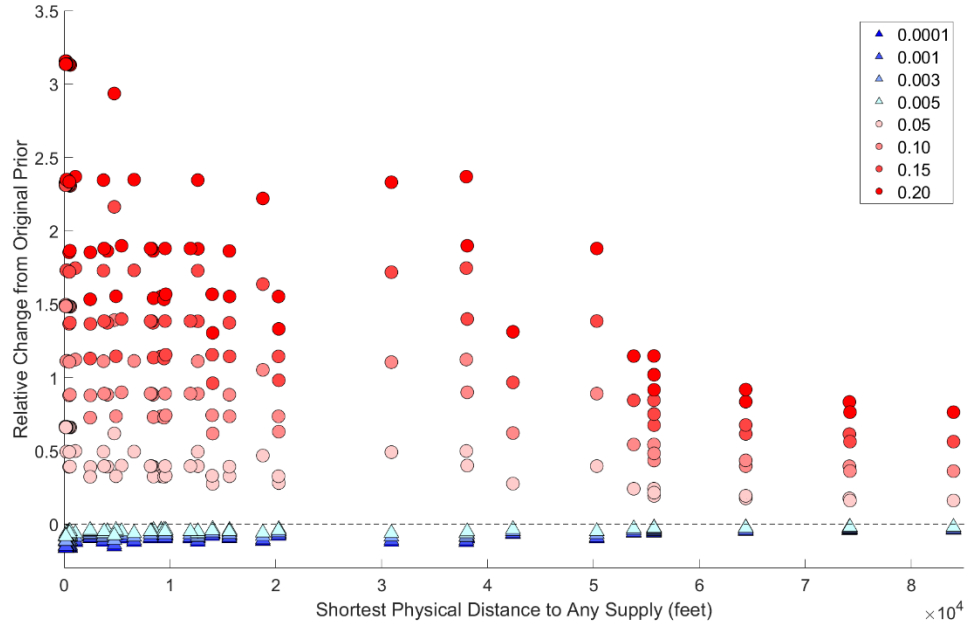


Figure 23. Relative change in marginal probabilities of failure versus shortest distance to supply for components with one reachable supply

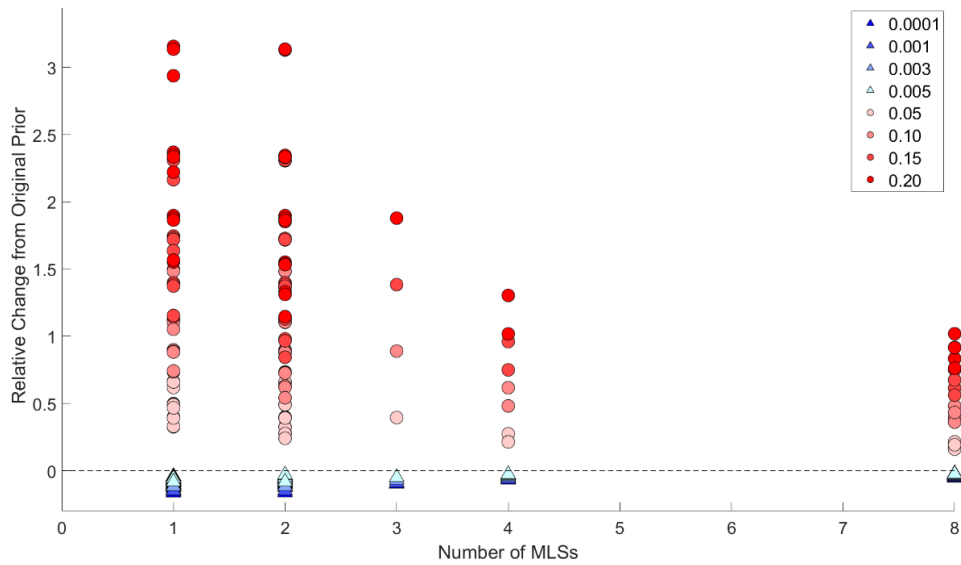


Figure 24. Relative change in marginal probabilities of failure versus number of MLSs for components with one supply

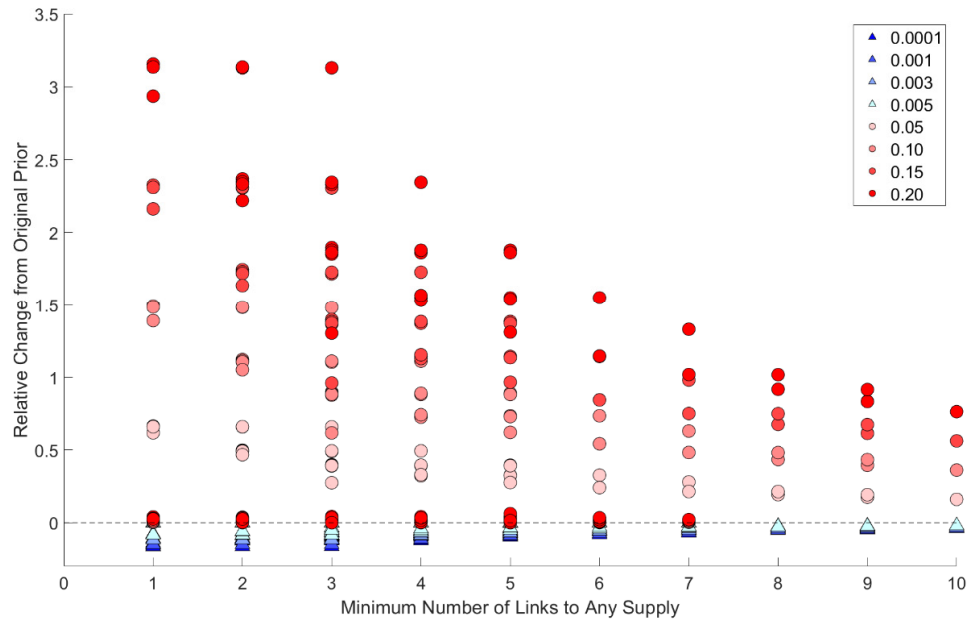


Figure 25. Relative change in marginal probabilities of failure versus minimum number of links from components to supply

From Figure 22, distribution components in the water network have either one, two, or three reachable supplies. The results show that components with more than one supply have little to no change in their marginal failure probabilities regardless of variations in power supply conditional probabilities of failure, while components with one supply have relative changes of up to three times the baseline (i.e., a 300% change). Having multiple reachable supplies is a type of redundancy in the system, limiting the impacts of changes, whether positive or negative, in probabilities of failure of any given supply.

Focusing on one-supply distribution components in Figure 23, the relative changes are found to be a function of each component's distance to a supply, with the relative changes in marginals being highest for components closest to a supply. In Figure 24, the

relative change in marginal probabilities is highest for components with one or two MLSs and decreases as the number of MLSs for a component increases. Components that have multiple MLSs have different paths available to the supply, representing redundancies in the system. With increased MLSs, there is an increased influence on the performance of the distribution component from the states of the links comprising the MLSs compared to the state of the supply nodes. Therefore, components with more MLSs have smaller relative changes with variations in the supply conditional probabilities of failure. Figure 25 shows a similar finding for the minimum number of links in a MLS for a component. In Figure 25, the relative changes in marginal probabilities of failure are largest for those components with few links to a supply and decrease as the number of links increases.

In addition to assessing the trends in changes in probabilities of failure for components of different characteristics, Figure 22 to Figure 25 show how increasing compared to decreasing supply failure probabilities leads to differences in impacts at the distribution components. Increased conditional failure probabilities at the power supplies, representing aging, damage, or lack of repair, correspond with much higher relative changes in failure probabilities at the distribution nodes compared to decreasing the conditional probabilities, representing retrofit upgrades or preventative measures, by the same amount. These results suggest that to increase CIS resilience, there is a greater benefit to repairing aged or slightly damaged components that may have increased conditional probabilities of failure compared to retrofitting components that may have already relatively low failure probabilities.

Looking more closely at the results shown in Figure 23, the decrease or increase in relative change with varying power supply component conditional failure probabilities

appears to follow an exponential trend as distance to a supply increases. A similar trend is seen in Figure 24 and Figure 25, where components with fewer numbers of paths to a supply and fewer links to a supply have higher relative changes. These results are consistent because as distance from a supply increases, components are more likely to have increasing numbers of paths to a supply and increasing numbers of links in their MLSs.

To quantify the increasing or decreasing exponential trend for the relative change in marginal failure probabilities versus distance to supply, exponential functions of the form shown in Equation (16) are fit to the data. The exponential fit function has two coefficients, a and b . The best-fit values for these coefficients are found for each varied value of power supply conditional probability of failure. Figure 26 shows the resulting exponential trendlines for relative change in component failure probability as distance from a component to a supply increases. Individual trendlines are shown for each varied value of conditional failure probability. For both increased and decreased conditional failure probabilities from the 0.01 baseline, the b -coefficient value indicating rate of decay is the same: $-1.665\text{e-}5$. The a -coefficient value varies depending on the value of the supply conditional failure probability. The a -coefficient is negative for decreased conditional probabilities of failure and positive for increased probabilities. The coefficient value increases linearly as a function of the conditional failure probabilities as shown in Figure 27, indicating larger predicted relative changes in distribution component performance with increased variations in supply component conditional failure probability. For decreased conditional failure probabilities, the differences between a -coefficient values also decrease, so that the exponential fit functions become increasingly similar to each other. This result supports the previous finding that decreasing the conditional probabilities

results in much smaller relative changes in distribution component outcomes than increasing them. The system is thus more vulnerable to higher conditional failure probabilities including due to minor damage and not as improved by component upgrades when prior component conditional failure probabilities are low.

$$f(x) = ae^{bx} \quad (16)$$

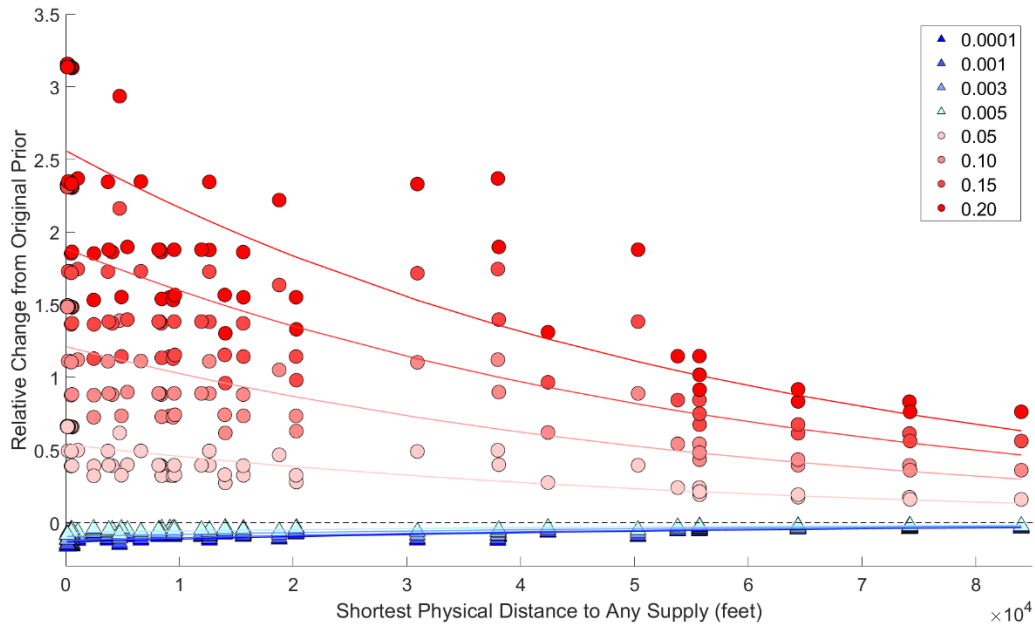


Figure 26. Exponential trendlines for each conditional failure probability for power supplies, for relative changes in component outcomes as a function of shortest distance to a supply

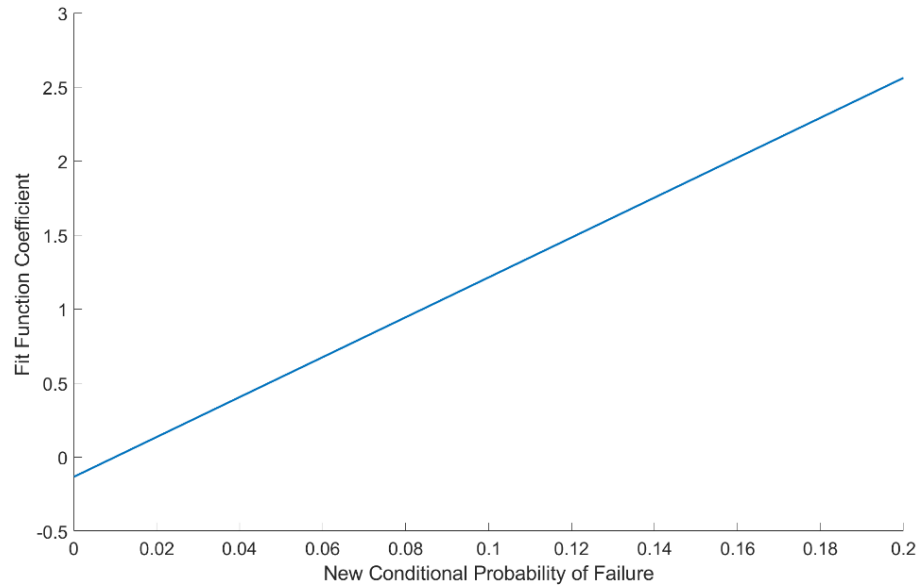


Figure 27. α -coefficient values of exponential fit functions versus varied network parameter

Varying the conditional probabilities of failure in the water supply CPTs produces similar results to changing the power supply probabilities of failure. The results for the relative changes in performance at the distribution components versus component characteristics for changes in water supply conditional probabilities of failure are similar to those shown in Figure 22 to Figure 25 and are therefore not shown here. The water supplies depend directly on the states of the power supplies, supporting the similarity in results.

The states of the access for repair nodes are one step removed from the day-to-day performance of a CIS. The reliability of a node needed to supply access to repair any failed nodes is only relevant for the cases of failures in the network. Therefore, varying the conditional failure probabilities of the access for repair nodes has the largest effect on the

resulting marginal probabilities of failure of the access for repair components themselves, and little to no influence on the marginal probabilities of failure of the distribution components. Relative changes in marginal probabilities for the access for repair components increase up to nine times when increasing access component conditional probabilities of failure from the baseline 0.01 to the maximum 0.20, and decrease up to 0.5 times when decreasing the access component conditional probabilities of failure from the baseline 0.01 to the minimum 0.0001. This difference in effect from increasing compared to decreasing component conditional probabilities of failure is consistent with previous results; increasing conditional probabilities of failure has a higher impact on results than decreasing probabilities of failure including in comparing results from changes of similar magnitudes. Among the conditional probabilities of failure at the component level, varying the failure probabilities for access for repair components has the smallest impact on system behavior. This result is due to the fact that an increase or decrease in failure for an access component in post-disaster recovery does not impact the network's overall performance unless there is an outage occurrence where a water or power component needs repair.

6.4 Redundant Power Supply Components

In this section, backup power supplies are added to the water distribution system to evaluate the effects of redundancies on system resilience. These redundancies are implemented by adding power supply node parents to the water supplies in the CIS BN. This network parameter change addresses vulnerabilities arising from service provision interdependencies in the CIS. The parameter variation reflects actions connecting water supplies to multiple substations or adding backup generators at the water supplies. The redundant power supply works such that a failure of one power supply does not lead to a

water supply failure. Instead, if one power supply fails, the other can be utilized to provide service. The additional power supplies are feasible and practical redundancies to be added to a system and several municipalities and water distribution utilities are implementing such actions to improve CIS resilience.

The lower marginal probabilities of failure for each of the water supplies from adding a redundant power supply at the water supply result in lower probabilities of failure for all distribution components. To quantify these changes, Figure 28 shows the relative change in component marginal probabilities of failures from the original priors with the added redundancy in power supplies as a function of the different distribution component characteristics as in Figure 22 to Figure 25. The negative relative changes show the decreases in marginal probabilities of failure when redundant power supplies are added. These results show similar trends to those found for directly decreasing the conditional probabilities of failure for the water and power supply components. The components with the largest changes in probabilities of failure are those with one water supply, shorter physical distance to a supply, fewer MLSs, and fewer minimum numbers of links to a supply.

The trends also match an exponential relationship, and an exponential function is fit for the relative change in marginal probabilities of failure of distribution components as a function of component characteristics. For a function of the form shown in Equation (16), the a - and b -coefficient values for the shortest distance to a supply, number of MLSs, and minimum number of links to a supply are shown in Figure 28. The resulting negative a -coefficient values are larger in magnitude than any of the a -coefficients in the fit functions for decreasing the conditional probabilities of the power supplies directly, meaning that the

relative changes shown in Figure 28 are larger than any relative changes calculated after decreasing conditional probabilities of failure for the power supplies (i.e., the blue triangular datapoints in Figure 22 to Figure 25). Therefore, adding redundant power supplies has more impact on system performance than decreasing conditional probabilities of failure for individual power supplies.

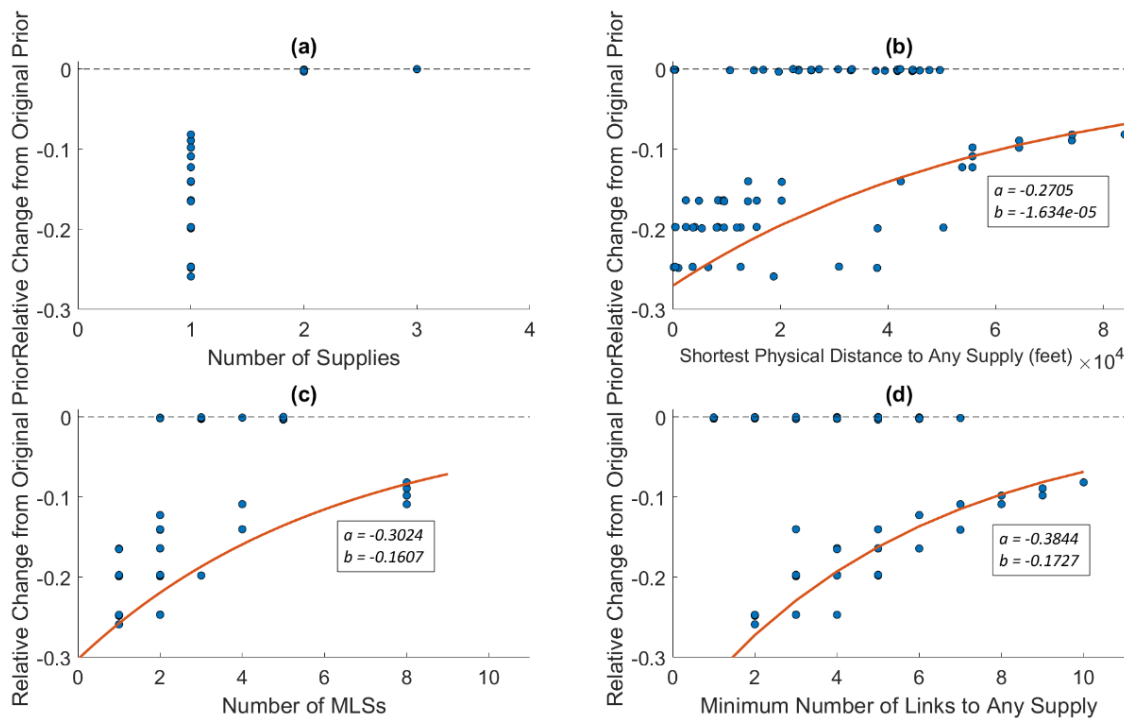


Figure 28. Relative changes in marginal probabilities of failures after adding redundancies at each water supply versus (a) number of supplies, (b) shortest physical distance to a supply, (c) number of MLSSs, and (d) minimum number of links to any supply

6.5 Additional Link Analysis

Finally, the system's link configuration is altered by adding new links to the water distribution system (i.e., new pipes). This network parameter variation corresponds with the action of building out new parts of a CIS. To evaluate the effect of adding new links, potential connection points for new links are searched in the existing system and the effect of adding those links on resilience is assessed. In the analysis, inferences are conducted over a failure scenario in which a failure at Supply 4 leads to the failure of two distribution nodes, indicated A and B as shown in Figure 29. In Figure 29, the marginal probabilities of failure for each component are shown with a color bar. The darker red components show failed components, and the lighter yellow components show that the remaining nodes in the network have significantly lower marginal probabilities of failure, indicating that they are unaffected by a failure at Supply 4.

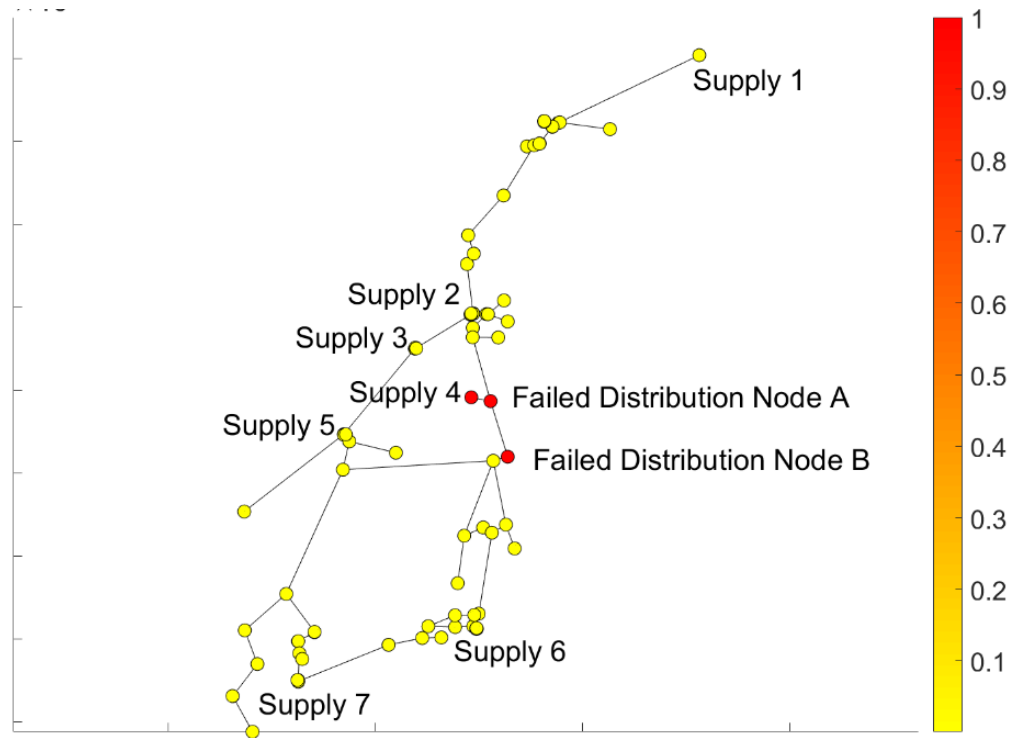


Figure 29. Failure scenario where failure of Supply 4 leads to failure of two distribution nodes

To identify potential new links to add, nearby nodes are searched to identify locations for new links to be built from the failed distribution nodes. New links connect each failed distribution node to a non-failed distribution component where a link does not already exist, starting with the closest component. An inference is then conducted for the new link configuration under the Supply 4 failure scenario. The baseline values for component conditional probabilities of failure are used to compare results across the configurations. Four new system link configurations are shown in Figure 30.

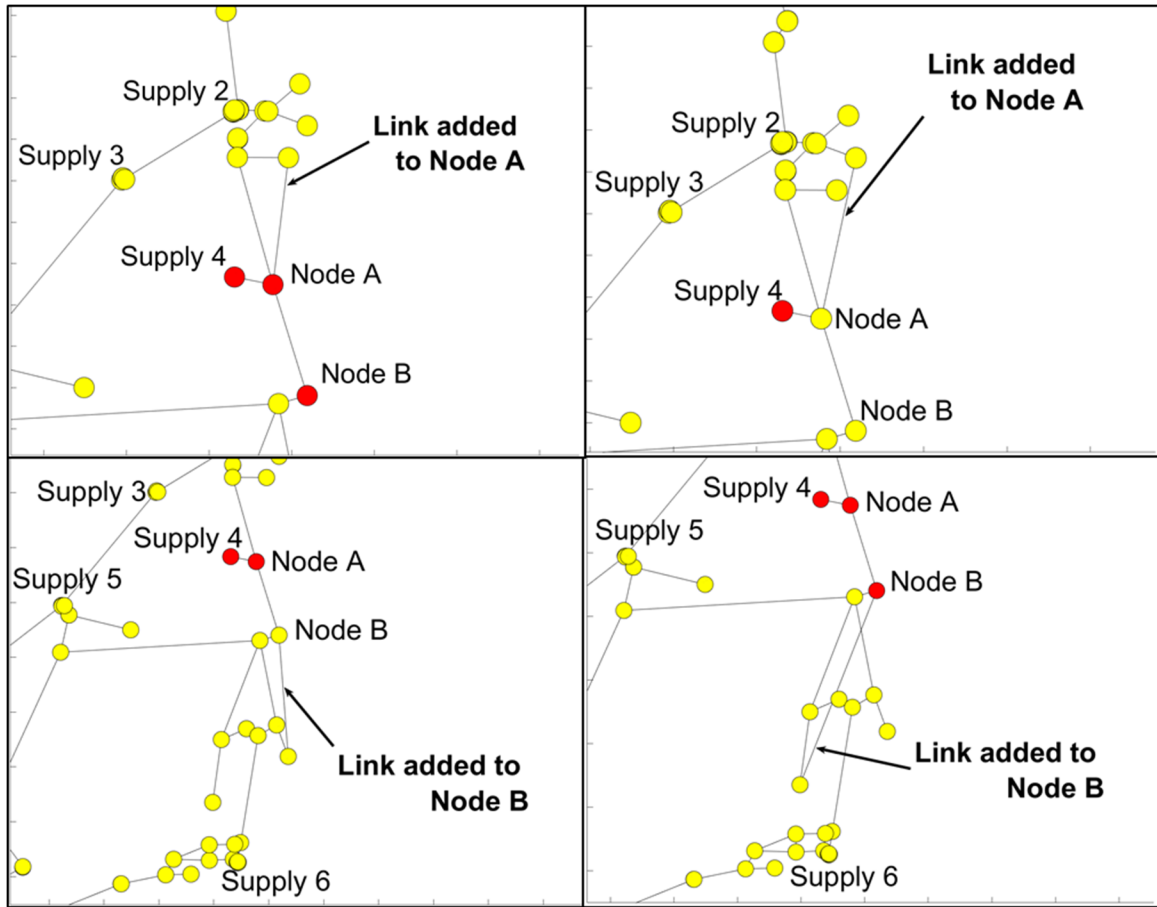


Figure 30. (a)-(d) from top left, right, bottom left, bottom right – New links added to the network and inference results for Supply 4 failure

The inference results show that a new link protects Nodes A and B from failing if it connects them back to another supply through new MLS paths. In Figure 30a (top left), connecting one failed distribution node (Node A) to another distribution node with multiple supplies does not lead to the creation of a new MLS path. Figure 30b (top right) shows the placement of a new link in the system, connecting failed Node A to a different non-failed distribution node, which protects both distribution nodes from failing. In this scenario, the new link connects both Nodes A and B to Supply 2 as well as Supply 4, and neither Nodes

A nor B fail when Supply 4 fails. Figure 30c-d show new links connected from Node B to other distribution nodes. However, when connecting Node B to its closest non-connected, non-failed distribution components, paths to additional supplies are not created to protect either failed distribution node. In this case, adding new links from Node A to nearby non-failed distribution components has a larger effect on improving overall system resilience than adding new links connecting Node B to nearby components. In general, adding links in the system improves performance if they are able to link failed components to additional supplies.

To quantify the effects of the varying system link configurations on resilience, the component failures are mapped to disruptions of service in the surrounding communities. The outcomes in the varying scenarios are compared and evaluated in the context of affected populations, housing units, and critical facilities. The population serviced by each component is estimated using United States Census data (U.S. Census Bureau 2010). Each census block is matched to the nearest water system component, and the number of people affected by a component's failure is estimated as the sum of populations of the census blocks matched to it. Similarly, the number of housing units affected is estimated as the sum of housing units located in the census blocks matched to a failed component. The critical facilities affected are estimated using Open Street Map data (Open Street Map 2018), identifying hospitals, schools, churches, fire departments, and potential evacuation locations within each census block. It is assumed that each component in the water system, including supply components, immediately services the surrounding population. Additional service information can be included if available. An example of the resulting matching between CIS components and populations is shown in Figure 31, where the

colored blocks are associated with three nodes, Supply 4 and the two distribution components subject to failure if Supply 4 fails.

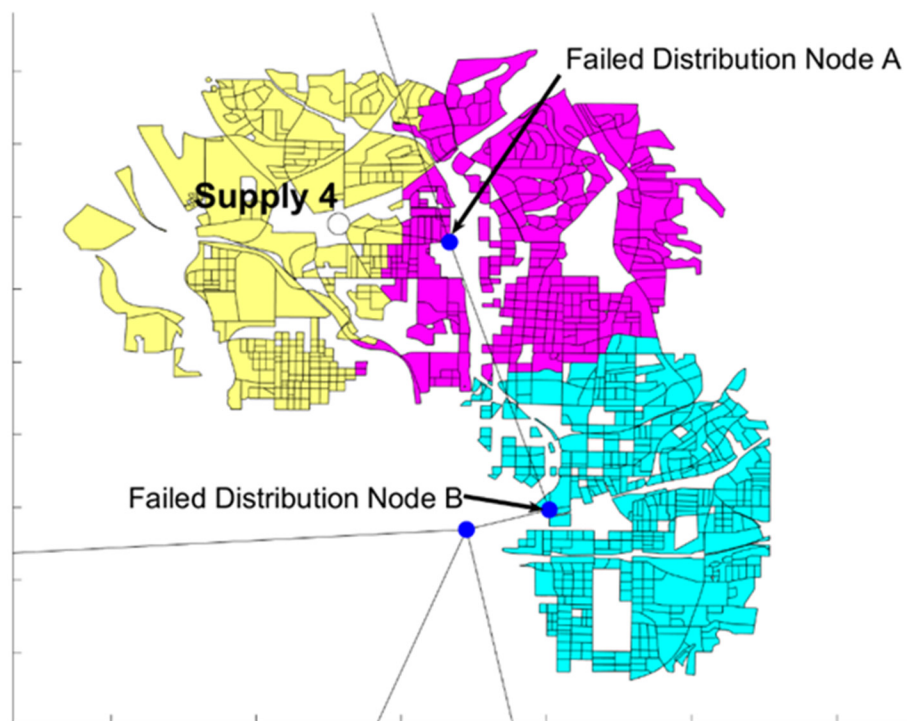


Figure 31. U.S. Census blocks associated with Supply 4 and two distribution components that fail in Supply 4 failure scenario

For the failure scenario in Figure 29, 96,217 people; 57,445 housing units; and seven critical facilities are estimated to experience disruption of service from a failure at Supply 4 and the subsequent failure at Nodes A and B. The same disruptions occur if the power supply at Supply 4 fails and there are no other redundant power supplies. After adding new links from Nodes A or B (e.g., Figure 30a-d) to non-failed distribution components, the new extents of disruptions of service are computed. The results for the

original and the potential new system link configurations are shown in Table 13. The figure numbers corresponding to each new schematic are shown in the table. For the varying system link configurations, the second option (i.e., Figure 30b) results in the lowest disruption of service under the failure scenario. Figure 30d results in the same disruption of service as the original configuration. The last line of Table 13 shows disruption of service if Node B were to survive, but Node A fails.

Evaluating CIS outcomes in terms of estimated populations affected by component failures informs future decision making when implementing retrofit upgrades and new system build-outs based on quantified impacts on increasing resilience.

Table 13. Estimated disruptions of service for Supply 4 failure scenario

Link Configuration with Supply 4 Failure	Population	Housing Units	Critical Facilities
Figure 29: Original Configuration	96,217	57,445	7
Figure 30a: Supply 4, Node A, and Node B fail	96,217	57,445	7
Figure 30b: Supply 4 fails	21,340	12,203	2
Figure 30c: Supply 4 and Node A fail	62,829	37,748	5
Supply 4 and Node B fail	54,728	31,900	4

6.6 Vulnerability Assessment Across Multiple Networks

Finally, inferences are conducted with the BN model and results are compared across 36 small-size water networks to further investigate the impacts of system topology and connectivity on overall performance. This is essential for planning future systems and

identifying the types of networks and the surrounding communities they serve that are most vulnerable.

Eight distinct networks are selected for the analysis. Two of the networks are real water networks, the Atlanta water network (individually analyzed in Sections 6.2 – 6.5) and the Shelby County water network in Tennessee (Johansen and Tien 2018). The remaining networks are synthetic networks (Giudicianni et al. 2018; University of Exeter 2019). As an objective of this study is to compare topology and connectivity with regard to supply-to-distribution component paths, additional networks are added by varying the number of supplies in each network. This creates a total of 36 networks, which are summarized in Table 14. The Atlanta water network is denoted as Network 2 in this section.

Table 14. Descriptions of multiple small-size water networks for vulnerability comparisons

Network	Type	# Nodes, n	# Edges, m	# Versions
1	Real	49	71	9
2	Real	112	122	8
3	Synthetic	25	44	4
4	Synthetic	48	51	6
5	Synthetic	20	21	2
6	Synthetic	32	34	2
7	Synthetic	12	14	2
8	Synthetic	37	58	3

To evaluate overall system performance and vulnerability across all networks, inferences are conducted over the BN model without evidence for each network and its derivatives. This same type of inference is conducted in Sections 6.2 – 6.4 for varying network parameters for the Atlanta water network. The average marginal failure probabilities for all components in each network are compared against the networks' topological characteristics, including density, average node degree, and average path length, and diameter. These are typical metrics used in graph theory to describe graphical structures, which are often used to represent CIS.

Since, this analysis focuses specifically on pathways between supply and distribution components, average path length is computed as the average number of links

in a MLS for each network, and diameter is also computed using supply-to-distribution pathways, i.e., MLSs. The calculations for density q , average path length AL , and diameter D are summarized in Equations (17) – (19). For these calculations, n represents the number of nodes in a network, m is the number of edges, s is the number of supplies, and d_{ij} is the length of a MLS between supply i and distribution component j . Average node degree is the average number of neighbors (i.e., connections) each component has in a network.

$$q = \frac{2m}{n(n-1)} \quad (17)$$

$$AL = \frac{\sum_{i=1}^s \sum_{j=1}^{n-s} d_{ij}}{sn} \quad (18)$$

$$D = \max(d_{ij}) \quad (19)$$

Figure 32 shows these metrics versus the size of the networks in terms of number of nodes. Since density is the same for each distinct network regardless of number of supplies, only the eight distinct networks are plotted in Figure 32, left. As the size of a water network increases, its density generally decreases while its diameter increases, which is consistent with previous works, i.e., Giudicianni et al. 2018. Its average path length will also increase as network size increases, as more components will be further from available supplies.

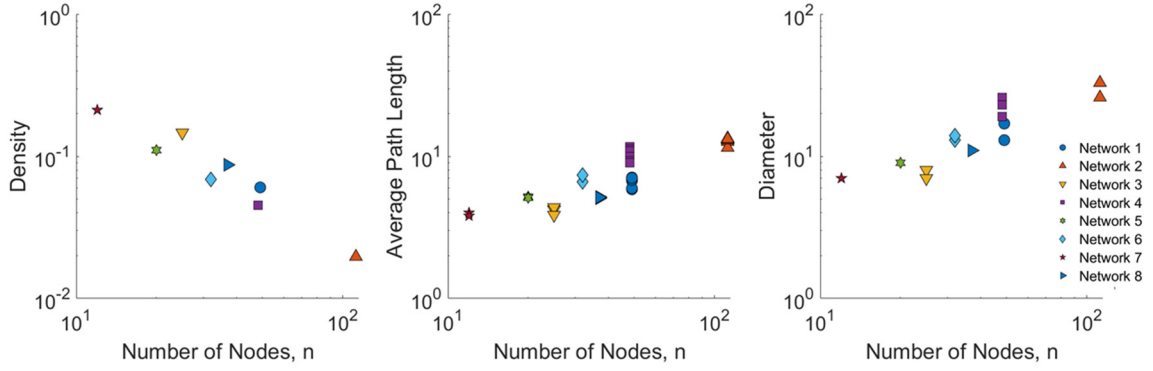


Figure 32. Graph metrics versus network size

The inference results are shown in Figure 33 to Figure 37. In Figure 33, the average component probability of failure for each network is shown to decrease as density increases, and in Figure 34, the same trend is seen for average failure probability as average node degree increases. These trends are expected because higher density and average node degree indicate more connected networks with more pathways. That is, more connected networks should have more MLS redundancies between supply and distribution components. However, Figure 33 and Figure 34 also show that for a distinct network, the average failure probability can vary greatly depending on the number of supplies. For instance, Network 4 is represented by the purple squares in Figure 33 and Figure 34. The number of supplies in the network varies from one to seven. With inferences conducted over all eight derivatives of the network, the average failure probability ranges from 0.0005 and 0.0015. Since the number of supplies has a significant impact on each system's overall vulnerability, the number of reachable supplies per component is evaluated next.

The number of reachable supplies is counted for each component in a network through its MLSs. In this study, MLSs are found through a depth-first search process as in

Applegate and Tien 2019, but different paths can be added to the BN model depending on the physical flow of commodities throughout a CIS. Components across all 36 networks have between one and five reachable supplies, with the average number of reachable supplies for the components in each network ranging from one to three. The number of components with only one reachable supply, i.e., distribution components without redundancies to different supply components, are also counted for each network.

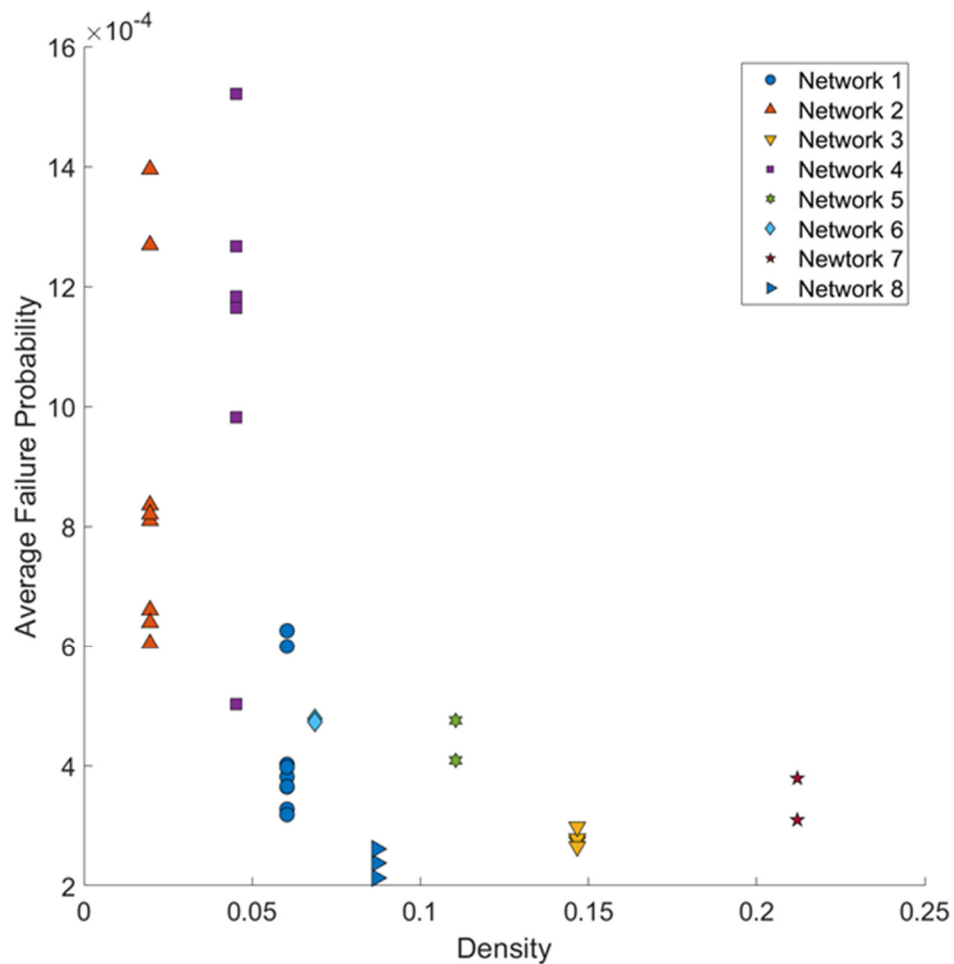


Figure 33. Average failure probability of network components versus network density

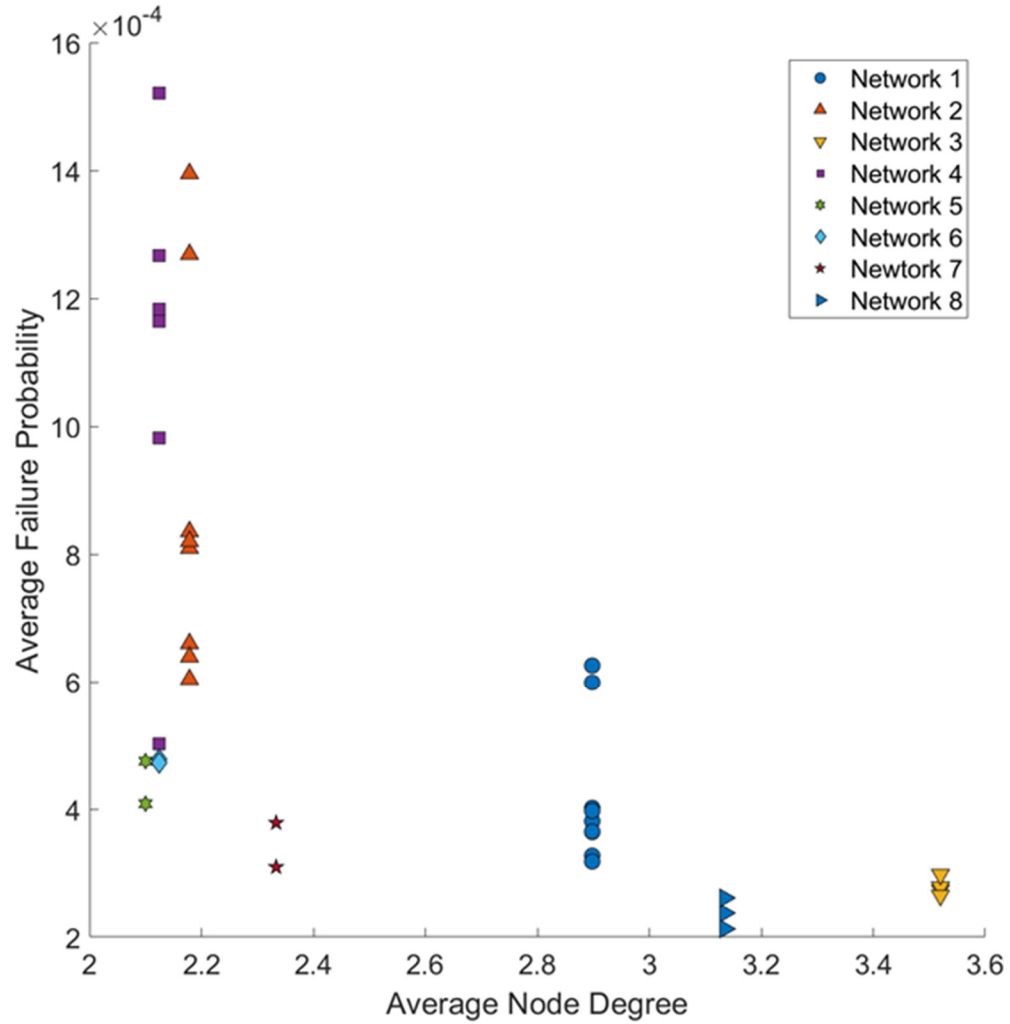


Figure 34. Average failure probability of network components versus average node degree

Overall system vulnerabilities are shown for average supplies per component in a network and fraction of one-supply components in a network in Figure 35 and Figure 36, respectively. In Figure 35, average failure probability decreases as the average number of supplies per component increases. Increasing the number of redundancies per components in a system therefore reduces overall system vulnerability, as expected. However, with more than one redundancy, i.e., more than two supplies per component, overall system

vulnerability levels off and does not continue to decrease. This suggests that it is more valuable for overall system resilience to ensure that all components in that system have at least one redundancy, rather than continuing to add redundancies to specific components.

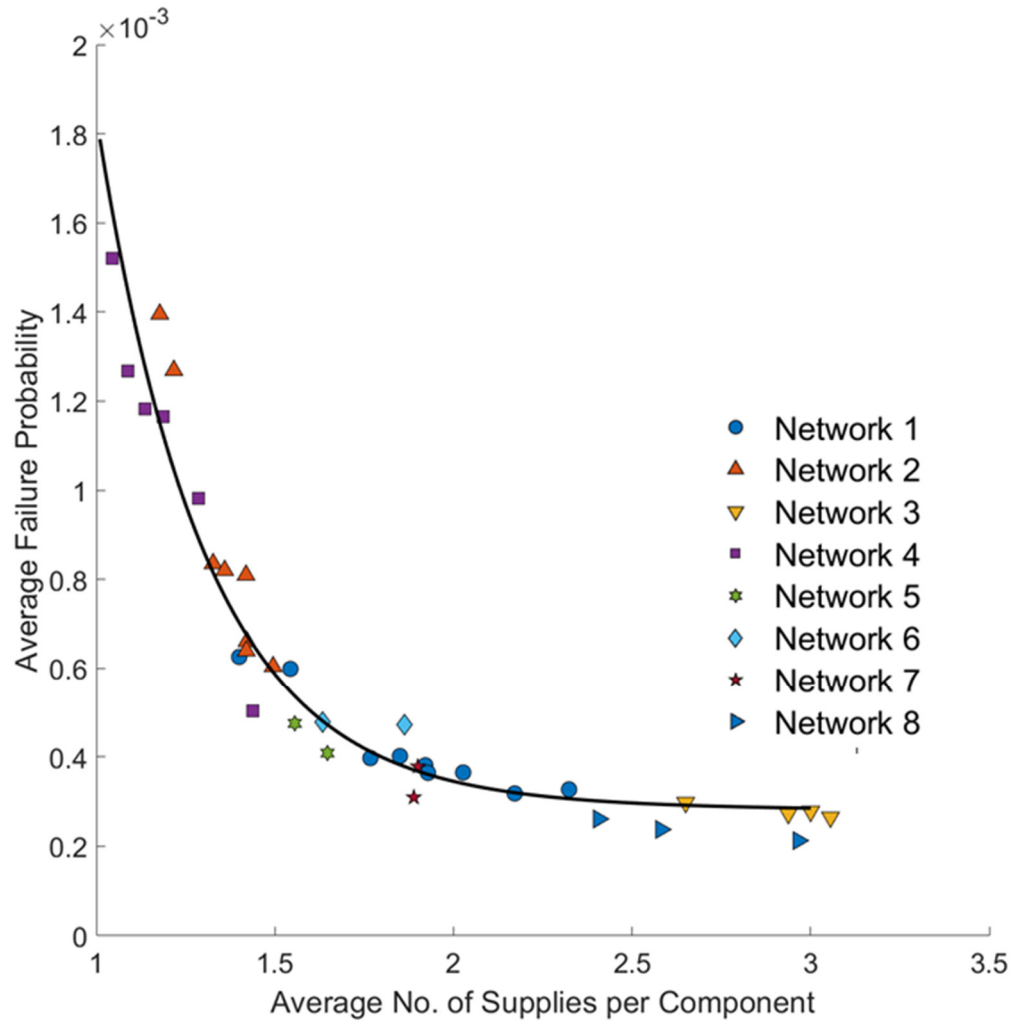


Figure 35. Average failure probability of network components versus average number of reachable supplies per component

In Figure 36, overall system vulnerability decreases as the fraction of one-supply components in a system increases. This is consistent with the results in Figure 35; networks with more uniformly distributed redundancies are less vulnerable. The increase in overall

system vulnerability is not linear, and as the fraction of one-supply components in a network surpasses 50%, the average failure probability in a network increases more quickly.

Lastly, Figure 37 shows overall system vulnerability versus the average supply-to-distribution component path lengths in each network. Here, the same shapes are used to represent each distinct network in the figure, i.e., Networks 1-8, and the color bar represents the number of supplies in each network. This way, the number of supplies in each network can be compared as average path length (i.e., average MLS length) increases. While overall system vulnerability generally increases with average path length, the number of supplies, and ultimately, the number of redundancies in a system is more impactful for reducing system vulnerability. This supports the results in Figure 35 and Figure 36.

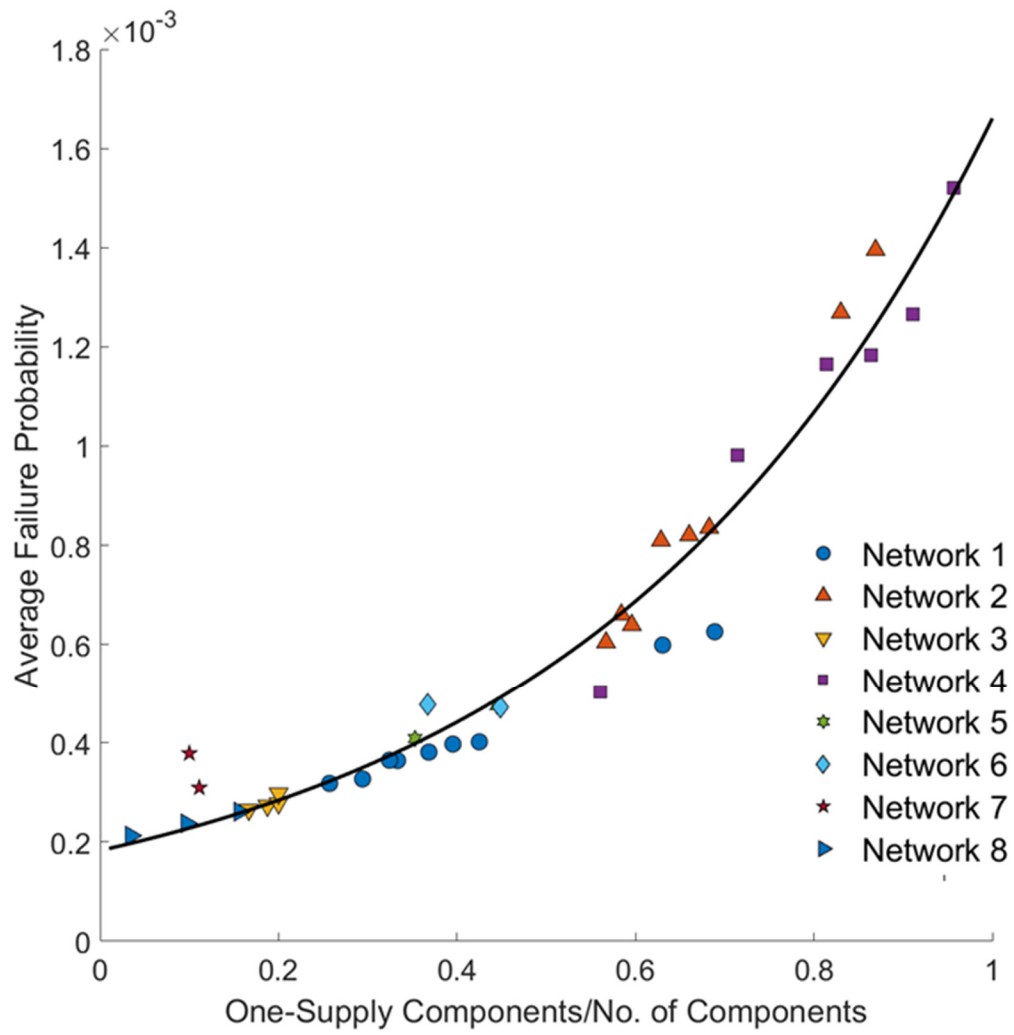


Figure 36. Average failure probability of network components versus fraction of one-supply components in each network

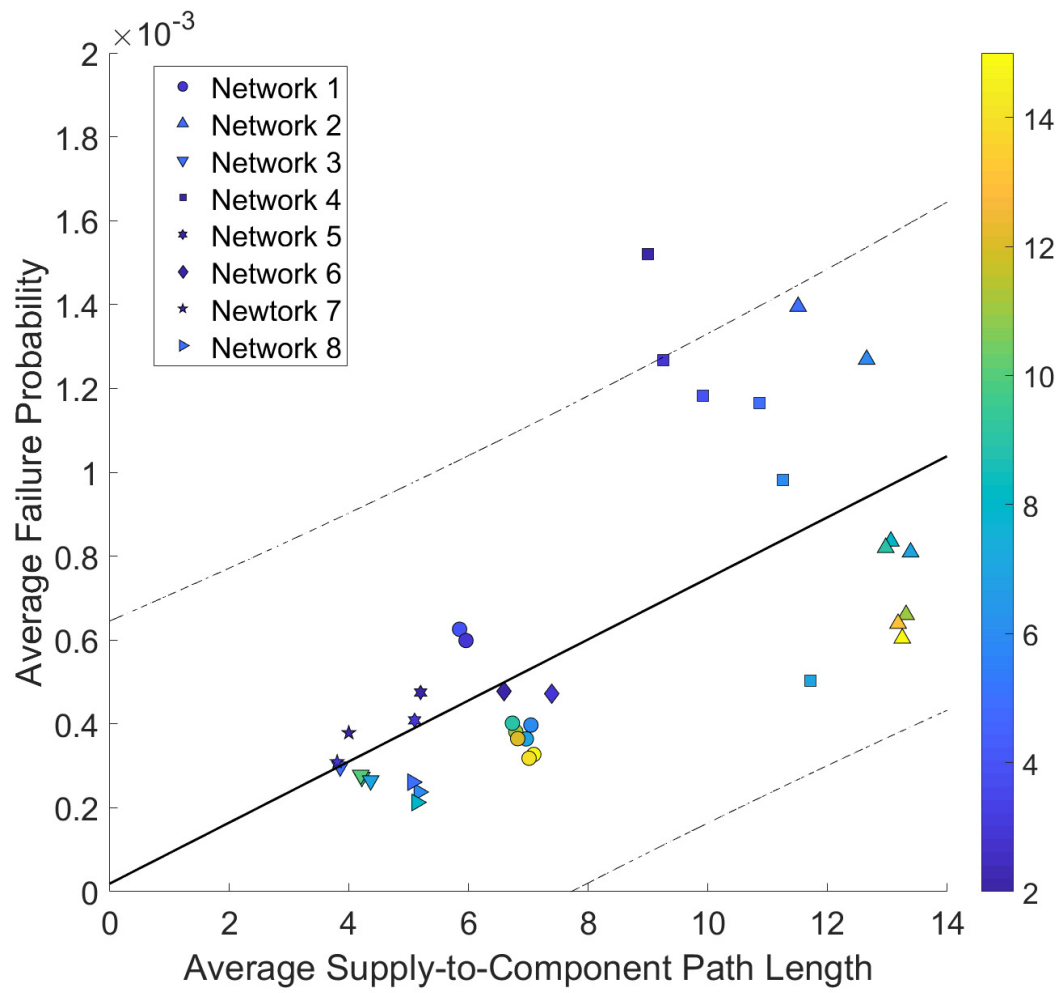


Figure 37. Average failure probability of network components versus average supply-to-distribution component path lengths in each network

CHAPTER 7. CONCLUSIONS

7.1 Introduction

The contributions of this dissertation are summarized in Sections 7.2 – 7.4. Section 7.2 presents the contributions of the framework for processing and classifying social media big data, including implications from applications of the framework. Section 7.3 presents the contributions of the probabilistic data integration framework. Because the steps of the framework are demonstrated in detail in the flood estimation application, contributions are further elaborated on by discussing the flood estimation results. Section 7.4 summarizes contributions of the probabilistic system model analyses conducted. The final section of this chapter is a description of future work and directions for continuing to increase CIS situational awareness and resilience with heterogeneous data and probabilistic analyses.

7.2 Social Media Data Processing and Classification Framework and Applications

This work provides a roadmap and insights to using social media data for CIS assessment and increasing situational awareness. The contributions of this framework are elaborated through the application results. Machine learning-based classifiers are built and evaluated for predicting the relevance or irrelevance of Twitter data to transportation and power system damages in real time. SVM, naïve Bayes model, and decision tree models are built based on training and testing sets developed for each system that represent a variety of damages, causes of those damages, and irrelevant information. A datapoint is only considered relevant if it references current or ongoing events or damage.

In the application, the classifiers for predicting the relevance of tweets to transportation system damages achieve high overall accuracies, high recall for determining irrelevance, and high precision for determining both classes when evaluated with cross-validation and on the testing set. High precision is preferred to high recall for detecting relevant tweets because it reduces the amount of falsely detected damages. Based on this metric, the SVM achieves the highest performance for classifying transportation damage. The classifiers predict relevance of tweets to power system damage with high accuracies for the training data but have significantly lower accuracies when evaluated using cross-validation and on the testing set. In evaluating the use of the classifiers to detect damage in real time among newly collected datapoints, a separate, unlabeled set of tweets collected during Tropical Storm Irma are classified for transportation system damage using the highest performing SVM model. The SVM is able to predict the classes for 16,106 unlabeled tweets in 5 seconds with a precision for detecting damage of 0.68.

These models, evaluations, and predictions on unlabeled data describe the cases and scenarios for which machine learning classifiers are able to predict the relevance of tweets to CIS damages with specificity and accuracy. The contributions of this proposed framework and applications for building and assessing these automated classifiers are threefold:

1. The development and filtering of the training and testing sets for two systems with different definitions of damage is discussed in detail. In one system (transportation), multiple types of damage are considered relevant and the sources of noise compared to relevant information in the data are discussed. In the other system (power), only two types of damage are considered relevant (i.e., power outages and damaged

power lines) and the nuances between relevant and irrelevant information that hinder successful classification are presented, underscoring the need for diversity and representativeness in training datasets. The ability to build classifiers that detect current and ongoing damages is demonstrated. In both systems, this distinction between current and ongoing damages and references to past or future damages are most difficult for the models to detect. While this significantly lowers the generalizability of the models for identifying power system damage, the classifiers for transportation system damage are successful at identifying most of the irrelevant tweets in evaluation with cross-validation and on the testing data. The highest performing model (SVM) is able to classify the test data with an accuracy of 88.4%. Creating classifiers that can automatically distinguish current damage from other references to the CIS of interest is essential for increasing situational awareness using social media data.

2. One of the classifiers is evaluated on a separate, large set (16,106) of unlabeled tweets to show how well it generalizes to predict relevance of data collected in real-time during a hazard event. While the performance of the classifier decreases from the testing set results as expected, the classifier is able to correctly identify relevance with a precision of 0.68 for this new dataset. High precision for relevance, limiting false alarms in detected datapoints for damage, is particularly important for the social media data to contribute to increasing situational awareness in a community and to support emergency response decisions.
3. Insights and recommendations are presented to improve the performance of classifiers for determining social media data relevance to critical infrastructure

damages. It is found that performance is more sensitive to the definitions of damage for each system, including the consideration of only current or ongoing damage as relevant and the specificity of the events of interest. The size of the training set, classifier type, and model hyperparameters have less effect. The classifiers also have the potential to identify CIS damage not explicitly included in the original damage event definitions, as shown in the predictions on the separate, unlabeled dataset.

7.3 Heterogeneous Data Integration Framework and Applications

The data integration framework provides a unique probabilistic approach to integrating data from across sources to estimate the probability of disaster or failure event occurrence given observed data. It updates prior probabilities of event occurrence with both individual and combined data sources. The framework is able to include data from a wide range of sensor types with varied likelihoods and shows how prior risks of an event change as new, potentially anomalous data is introduced. It is applicable to general disaster or failure events, including natural disasters and structural or infrastructure system failures as long as data is available. The Bayesian updating approach for data integration does require the establishment of prior probabilities of event occurrence. If these are unknown, they are initially assumed with the potential use of uninformative priors to limit bias in the results.

In an application example, the framework is applied to estimate flood events in Louisiana in August 2016. Prior flood event risk in a parish is calculated based on FEMA risk map data. For updating in the application, physical stream gage data and social Twitter data are used. Data from other sensor sources can be easily added using the same approach

presented, with new likelihood calculations for each additional sensor type. While these likelihoods will be defined differently depending on the data output by each source, the general framework is flexible such that changing these will not change the implementation process as long as the sensor source likelihoods can be found.

The results from updating prior flood risks in Louisiana from August 10-13, 2016, show that additional data over time and from across both sensor sources increases the amount of updating possible in real-time event estimation. The results are validated by comparing the parishes with highest updated probabilities with FEMA disaster declarations and post-event inundation and precipitation maps from USGS. This showed similar regions of flooding indicated based on updating from the integrated datasets as from the true event.

In another example, the proposed framework is used to update probabilities of damage for a transportation system component during a hazard event. The transportation component is an intersection in Atlanta that is necessary for access to a supply component in the city's water network. The same data sources as for the first application example, USGS stream gage and Twitter data, are integrated to estimate component damage during Tropical Storm Irma. This example demonstrates the framework's use for monitoring CIS components with information from multiple sensor sources measuring different conditions. Here, stream gage data indicates potential weather conditions that cause damage to the intersection, while Twitter data indicates any possible damage on the roadway to make it impassable. Results from this application can be directly input into probabilistic system models, such as the BN model used for analysis in this dissertation.

Probabilistic updating using the proposed framework increases situational awareness and can be used to support community decision-making during and after disaster or failure events. The impacts of the results obtained from integrating data using the proposed framework are in three main areas, described with the flood estimation application:

1. Updating prior risk assessments based on integrated inferences from multiple data sources improves situational awareness, particularly if done in real-time, with updated probabilities indicating locations or components most likely to be experiencing the event at the time assessed. In the example application, communities have a more holistic view of their flood risk from monitoring both conditions from the natural environment and firsthand accounts from community members of a current event.
2. Based on the granularity of the estimation, comparing updated probabilities facilitates prioritization of resources by location and time during a disaster or failure event. In the application, directly comparing flood risks between parishes supports real-time decision-making and resource allocation during emergency response by identifying the most vulnerable parishes.
3. The framework enables assessment of the monitoring capabilities of different data sources. By integrating data first within networks then across sources, the updating approach reveals the availability of data and the updated probabilities from individual compared to combined sources post-event. These results evaluate a community's monitoring capabilities, demonstrate what types of data are available throughout the community, and detail where multiple data sources can effectively

supplement each other. The application shows how stream gage data and Twitter data contribute to updating prior risks, both individually and particularly when combined.

7.4 Probabilistic System Model Analyses

Finally, this dissertation assesses the impacts of varying parameters and topology of a CIS on system resilience. Impacts are evaluated at the individual distribution component level and overall system level, with probabilistic analysis results that include interdependencies between multiple infrastructure systems. A water distribution system is used as an example, including the connections of the water system with power and access components needed for repair. The network parameters varied include the conditional probabilities of failure given a hazard event for power supplies, water supplies, and access for repair nodes; redundant power nodes that are added to service water supply nodes; and new links added (i.e., pipes) to the network to investigate the effects of changing network connectivity in protecting against system failures. Overall system vulnerability is compared across multiple water networks to further evaluate the impact of connectivity on system performance.

A BN approach is used to conduct inferences and evaluate the resulting changes in marginal probabilities of component failure from the parameter variations. Results from the parameter variations are compared to results from the baseline network to quantify the changes in CIS outcomes. Impacts are also measured based on affected populations and critical facilities.

The following conclusions are drawn from the results of the probabilistic analyses. Adding redundant power supplies at the water supplies is the most effective in decreasing marginal probabilities of failure across the system. In assessing the characteristics of components with the largest impacts, distribution components connected to one supply have the largest changes in performance compared to components with multiple reachable supplies for all parameter variations. For one-supply components, the degree of relative change in marginal probabilities of failure for a component is a function of the characteristics of the component, including physical distance to the nearest supply node, number of MLSs, and minimum number of links to a supply. Components closer to a supply have higher relative changes than components farther away. Components with fewer MLSs and fewer minimum links in a MLS also show larger relative changes in probabilities of failure. These results indicate the types of components that are most likely to experience changes in expected functionality with network parameter changes.

Exponential functions are fit to assess the relative change in component marginal probabilities of failure with changes in component characteristics. As a function of the shortest physical distance between a component and supply, the coefficient of the exponential fit function increases linearly as the conditional probability of failure given hazard increases, indicating larger predicted relative changes as the supply component conditional probabilities of failure increase. Increasing component-level conditional probabilities of failure results in larger relative changes compared to decreasing probabilities of failure. Introducing new links in the network can also increase resilience, if the links connect one-supply distribution components to multiple other supplies through new MLS paths. The impacts of network parameter changes are evaluated based on the

relative change in marginal probabilities of distribution component failure and the population affected in supply failure scenarios.

Assessing the sensitivity in CIS outcomes to the varying network parameter changes informs CIS management decisions by facilitating prioritization of actions that will lead to the greatest increases in resilience. For example, results show the benefits of adding redundant power supplies at water supply nodes to increase CIS resilience. Results also indicate that it is more important to repair aged or slightly damaged components, which have higher conditional probabilities of failure, compared to retrofitting components, which may have already low failure probabilities. Locations for new system build-outs should be selected to increase the number of reachable supplies for distribution components. Knowing the characteristics of the components that experience the greatest impacts from network parameter changes informs a performance-based approach to design where decisions for system repairs or upgrades can target performance at particular components to increase resilience.

Lastly, across different water networks, results show that networks that have evenly distributed redundancies throughout all components have the lowest overall system vulnerability. This occurs in networks that have at least one supply redundancy per component. As the average number of supplies per component increases, the average failure probability of all components in a network does not continue to decrease. These analyses again provide insights for prioritizing decisions for adding redundancies in a system and can characterize aspects of less vulnerable and more resilient systems.

7.5 Future Work

The work in this dissertation can be extended in several ways. First, for increasing situational awareness through CIS monitoring capabilities, further investigation can be done of machine learning methods for processing social media big data, and additional studies can be conducted to integrate data within single sensor sources. Second, probabilistic system outcomes can be compared across larger water networks for a fuller analysis of the impact of topology on CIS vulnerability and resilience. Additional network parameters can also be investigated for their impacts.

7.5.1 *Additional Machine Learning Applications*

This study proposes a framework for processing and classifying social media big data to identify CIS damages and hazard events. The framework first filters social media text-based posts to reduce the amount of data for training and constructing machine learning-based classifiers. In applications of this framework, three well-known classifiers are built on manually labeled training data.

The framework can be applied to create additional machine learning classifiers to compare the models' performance and to make recommendations for the most suitable model for predicting damages in a specific CIS. For instance, in this study, the classifiers built were not generalizable to identify power system damages. Additional classifiers and unsupervised learning models can be built through the proposed framework. The performance of these new models can reveal whether the identification of such specific and current power system damages is feasible as a classification task.

7.5.2 Data Integration within Sensor Sources

In the proposed probabilistic data integration framework, data from each source is integrated first before integration across data sources. Integrating data from a single sensor source does not assume conditional independence. Conditional independence is assumed for integrating data across sensor sources. In each application presented in Chapter 5, USGS stream gage data is integrated by creating an additional variable to combine the likelihood of observing data from each individual stream gage. To extend this work in data integration, different combinations of data likelihoods for each source can be applied to evaluate the impact of source likelihoods on the resulting analyses for integration across sources.

For instance, for the applications presented in this dissertation, alternate integrations of stream gage likelihoods and the subsequent changes in results when integrated with other sensor sources can be investigated. Moreover, additional sensor sources can be added to the integration of data across sensor sources, and the impacts of information from a range of sensor sources can be evaluated.

7.5.3 Impacts of Network Topology on the Resilience of Different Networks

Finally, the impact of network topology on overall system vulnerability and resilience can be further assessed by varying prior probabilities of component failure in a single network (i.e., Atlanta water) and across more networks of different sizes and types. To vary the parameters of the Atlanta water network in this study, all conditional probabilities of failure are initially assumed as 0.01, i.e., a 1% probability of failure given hazard occurrence, and conclusions as changes are made to supply vulnerabilities, system

redundancies, and connectivity. Including variation in prior probabilities of component failure has potential for investigating the impacts of components that age and deteriorate at different rates in the system on overall system behaviour. It can also further help in prioritizing actions for decreasing system vulnerability.

Moreover, in this study, eight distinct networks are evaluated with the BN model, and additional networks are created by varying the number of supplies in each network. The networks are all small-size water networks. To draw more in-depth conclusions about the impacts of network topology on overall CIS resilience, a wider variety of network sizes is needed for analysis. The same analyses can be conducted to compare the vulnerabilities of different types of CIS, which may have different trends in topology and connectivity.

REFERENCES

- Alawadhi, S., Aldama-Nalda, A., Chourabi, H., Gil-Garcia, J.R., Leung, S., Mellouli, S., Nam, T., Pardo, T.A., Scholl, H.J., and Walker, S. 2012. "Building understanding of smart city initiatives." *International Conference on Electronic Government*. 40-53.
- Alexander, D. 2014. "Social Media in Disaster Risk Reduction and Crisis Management." *Science and Engineering Ethics*, 20(3), 717-33.
- American Society of Civil Engineers. 2016. "Failure to act: closing the infrastructure investment gap for America's economic future." Economic Development Research Group.
- Amin, M., 2002. "Toward secure and resilient interdependent infrastructures." *Journal of Infrastructure Systems*, 8(3), pp.67-75.
- Applegate, C.J. and Tien, I., 2019. "Framework for probabilistic vulnerability analysis of interdependent infrastructure systems." *Journal of Computing in Civil Engineering*. 33(1).
- Attoh-Okine, N.O., Cooper, A.T. and Mensah, S.A., 2009. "Formulation of resilience index of urban infrastructure using belief functions." *IEEE Systems Journal*, 3(2), pp.147-153.
- Boddula, V., Ramaswamy, L., Pasumarthi, R., Mishra, D. 2016. "Data Driven Analysis of Algal Bloom Activity for Effective Water Sustainability." *IEEE International Conferences on Big Data and Cloud Computing (BDCloud), Social Computing and Networking (SocialCom), Sustainable Computing and Communications (SustainCom)*, 424-430.
- Brakkton, B. 2020. "After quakes, Puerto Rico's electricity is back on for most, but uncertainty remains." <https://www.npr.org/2020/01/13/796018006/after-quakes-puerto-ricos-electricity-is-back-on-for-most-but-uncertainty-remain>.
- Burby, R. 2001. "Flood insurance and floodplain management: The US experience." *Global Environ. Change Part B: Environ. Hazard*. 3 (3): 111–122.
<https://doi.org/10.3763/ehaz.2001.0310>.
- Chen, Z. and Hutchinson, T. 2007. "Urban Damage Estimation Using Statistical

- Processing of Satellite Images.” *ASCE Journal of Computing in Civil Engineering*.
- Coumou, D. and Rahmstorf, S. 2012. “A decade of weather extremes.” *Nature Climate Change: Perspective*.
- de Albuquerque, J.P., Herfort, B., Brenning, A., and Zipf, A. 2015. “A Geographic Approach for Combining Social Media and Authoritative Data Towards Identifying Useful Information for Disaster Management.” *International Journal of Geographical Information Science*, 29(4), 667-689.
- Duenas-Osorio, L., Craig, J.I., and Goodno, B.J., 2006. “Seismic response of critical interdependent networks.” *Earthquake Engineering and Structural Dynamics*. 36:285-306.
- Durrant-Whyte, H. and Henderson, T.C. 2008. “Multisensor Data Fusion.” *Springer Handbook of Robotics*. 585-610.
- Eusgeld, I., Nan, C., Dietz, S. 2011. “System-of-systems Approach for Interdependent Critical Infrastructures.” *Reliability Engineering and System Safety*, 96(6), 679-686.
- Fan, C., Jiang, Y., and Mostfavi, A. 2020. “Social Sensing in Disaster City Digital Twin: Integrated Textual-Visual-Geo Framework for Situational Awareness during Built Environment Disruptions.” *Journal of Management in Engineering*. 36(3).
- Fan, C., Wu, F., and Mostafavi, A. 2020. “A Hybrid Machine Learning Pipeline for Automated Mapping of Events and Locations from Social Media in Disasters.” *IEEE Access*. 8:10478-10490.
- Federal Emergency Management Agency (FEMA). 2016. “Louisiana Severe Storms and Flooding (DR-4277).” *Disasters*. <<https://www.fema.gov/disaster/4277-0>> (Sept. 20, 2017).
- Federal Emergency Management Agency (FEMA). 2017. “The Disaster Declaration Process.” <<http://www.fema.gov/declaration-process>> (Sept. 20, 2017).
- Federal Emergency Management Agency (FEMA). 2017. “Georgia Hurricane Irma (DR-4338).” <<https://www.fema.gov/disaster/4338>> (October 2017).
- Federal Emergency Management Agency (FEMA). 2018. “2017 Hurricane season FEMA after action report.”

- Genge, B., Siaterlis, C., and Hohenadel, M., 2012. "Impact of network infrastructure parameters to the effectiveness of cyber attacks against industrial control systems." *International Journal of Computers, Communications, and Control*. 7(4):674-687.
- Giudicianni, C., Di Nardo, A., Di Natale, M., Greco, R., Santonastaso, G.F., and Scala, A. 2018. "Topological Taxonomy of Water Distribution Networks." *Water*, 10(4), 444.
- Groves, D.G., Kuhn, K., Fischbach, J., Johnson, D.R., and Syme, J. 2016. "Analysis to Support Louisiana's Flood Risk and Resilience Program and Application to the National Disaster Resilience Competition." *Rand Corporation 2016: Project Report*.
- Guidotti, R., Chmielewski, H., Unnikrishnan, V., Gardoni, P., McAllister, T. and van de Lindt, J., 2016. "Modeling the resilience of critical infrastructure: The role of network dependencies." *Sustainable and resilient infrastructure*, 1(3-4), pp.153-168.
- Hackmann, G., Guo, W., Yan, G., Sun, Z., Lu, C., and Dyke, S. 2014. "Cyber-physical codesign of distributed structural health monitoring with wireless sensor networks." *IEEE Transactions on Parallel and Distributed Systems*, 25(1), 63-72.
- Hall, M., Frank, E., Holmes, G., Pfahringer, B. 2009. "The WEKA Data Mining Software: An Update." *ACM SIGKDD Explorations Newsletter*, 11(1), 10-18.
- Hirabayashi, Y., Mahendran, R., Koirala, S., Konoshima, L., Yamazaki, D., Watanabe, S., Kim, H., and Kanae, S. 2013. "Global Flood Risk under Climate Change." *Nature Climate Change*, 3, 816–21.
- Imran, M., Elbassuoni, S., Castillo, C., Diaz, F., and Meier, P. 2013. "Extracting Information Nuggets from Disaster-Related Messages in Social Media." *Proceedings of the 10th International ISCRAM Conference*.
- Ireson, N. 2009. "Local Community Situational Awareness During an Emergency." 3rd *IEEE International Conference on Digital Ecosystems and Technologies*. 49-54.
- Jensen, F. V., and Nielsen, T. D., 2007. *Bayesian Networks and Decision Graphs*, 2nd ed., New York: Springer.

- Johansen, C., Horney, J., and Tien, I., 2017. "Metrics for Evaluating and Improving Community Resilience," *ASCE Journal of Infrastructure Systems*, 23(2).
- Johansen, C. and Tien, I. 2018. "Probabilistic Multi-Scale Modeling of Interdependencies between Critical Infrastructure Systems for Resilience," *Sustainable and Resilient Infrastructure*, 3(1), 1-15.
- Jongman, B., Wagemaker, J., Romero, B.R., and De Perez, E.C. 2015. "Early Flood Detection for Rapid Humanitarian Response: Harnessing Near Real-time Satellite and Twitter Signals." *ISPRS International Journal Geo-Inf.* 4(4): 2246-2266.
- Kankanamge, N., Yigitcanlar, T., Goonetilleke, A., and Kamruzzaman, M. 2019. "Can Volunteer Crowdsourcing Reduce Disaster Risk? A Systematic Review of the Literature." *International Journal of Disaster Risk Reduction*. 36:101097.
- Khaleghi, B., A. Khamis, F. O. Karray, and S. N. Razavi. 2013. "Multisensor data fusion: A review of the state-of-the-art." *Inf. Fusion* 14 (1): 28–44. <https://doi.org/10.1016/j.inffus.2011.08.001>.
- Klotz, M., Wurm, M., Zhu, X., Taubenbock, H. 2017. "Digital Deserts on the Ground and from Space." *Joint IEEE in Urban Remote Sensing Event*, 1-4.
- Krishnamurthy, V., Kwasinski, A., and Duenas-Osorio, L. 2016. "Comparison of power and telecommunications dependencies and interdependencies in the 2011 Tohoku and 2010 Maule earthquakes." *ASCE Journal of Infrastructure Systems*. 22(3).
- Kumar, S., Morstatter, F., Liu, H. 2014. "Twitter Data Analytics." *New York: Springer*, 1041-4347.
- Leetaru, K. H., S. Wang, G. Cao, A. Padmanabhan, and E. Shook. 2013. "Mapping the global Twitter heartbeat: The geography of Twitter." *First Monday* 18 (5): 1–12. <https://doi.org/10.5210/fm.v18i5.4366>.
- Lekkas, E., Andreadakis, E., Alexoudi, V., Kapourani, E., and Kostaki, I. 2012. "The Mw=9.0 Tohoku Japan earthquake (March 11, 2011) tsunami impact on structures and infrastructure." *World Conference on Earthquake Engineering*.

- LSU AgCenter and LaDOTD (Louisiana State University Agricultural Center and Louisiana Department of Transportation and Development). 2017. "LA Floodmaps." Accessed March 3, 2017. <http://maps.lsuagcenter.com/floodmaps/>.
- Lynch, J.P. and Loh, K.J. 2006. "A summary review of wireless sensors and sensor networks for structural health monitoring." *The Shock and Vibration Digest*. 30(91).
- Martin, Y., Cutter, S.L., Li, Z., Emrich, C.T., and Mitchell, J.T. 2020. "Using Geotagged Tweets to Track Population Movements to and from Puerto Rico after Hurricane Maria." *Population and Environment*, 1-24.
- Miorandi, D., S. Sicari, F. De Pellegrini, and I. Chlamtac. 2012. "Internet of things: Vision, applications and research challenges." *Ad Hoc Networks* 10 (7): 1497–1516. <https://doi.org/10.1016/j.adhoc.2012.02.016>.
- Mone, G. 2015. "The New Smart Cities." *Communications ACM*. 58(7):20-21.
- Morss, R.E., Wilhelmi, O., Downton, M.W., and Grunfest, E. 2005. "Flood risk, uncertainty, and scientific information for decision making: Lessons from an interdisciplinary project." *Bulletin of American Meteorological Society*. 86(11): 1593-1602.
- Musaev, A., D. Wang, and C. Pu. 2014. "LITMUS: Landslide detection by integrating multiple sources." In Proc., 11th Int. Conf. on Information Systems for Crisis Response and Management, 677–686. State College, PA: University Park.
- Musaev, A., Wang, D., and Pu, C. 2015. "LITMUS: A Multi-Service Composition System for Landslide Detection." *IEEE Transactions on Services Computing*, 8(5), 715-726.
- Neirotti, P., De Marco, A., Cagliano, A.C., and Mangano, G. 2014. "Current trends in smart city initiatives: some stylised facts." *Cities: The International Journal of Urban Policy and Planning*. 38:25-36.
- Open Street Map. 2018. Map data copyrighted OpenStreetMap contributors and available from <https://www.openstreetmap.org>
- Ouyang, Min. 2014. "Review on modeling and simulation of interdependent critical infrastructure systems." *Reliability engineering & System safety* 121: 43-60.

- Ouyang, M., Duenas-Osorio, L., and Min, X., 2012. "A three-stage resilience analysis framework for urban infrastructure systems." *Structural Safety*. 36-37:23-31.
- Pakzad, S. N., Fenves, G. L., Kim, S., and Culler, D. E. 2008. "Design and implementation of scalable wireless sensor network for structural monitoring." *Journal of infrastructure systems*, 14(1), 89-101.
- Pant, R., Barker, K. and Zobel, C.W., 2014. "Static and dynamic metrics of economic resilience for interdependent infrastructure and industry sectors." *Reliability Engineering & System Safety*, 125, pp.92-102.
- Panteli, M. and Pierluigi, M., 2017. "Modeling and evaluating the resilience of critical electrical power infrastructure to extreme weather events." *IEEE Systems Journal*. 11(3):1733-1742.
- Pereira, J., Pasquali, A., Saleiro, P., and Rossetti, R. 2017. "Transportation in social media: an automatic classifier for travel-related tweets." *EPIA 2017: Progress in Artificial Intelligence*. 355-366.
- Poljansek, K., Bono, F., and Gutierrez, E., 2012. "Seismic risk assessment of interdependent critical infrastructure systems: The case of European gas and electricity networks." *Earthquake Engineering and Structural Dynamics*. 41:61-79.
- Rawat, P., Singh, K.D., Chaouchi, H., Bonnin, J.M. 2014. "Wireless Sensor Networks: A Survey on Recent Developments and Potential Synergies." *Journal of Supercomputing*, 68(1), 1-48.
- Reed, D.A., Kapur, K.C. and Christie, R.D., 2009. "Methodology for assessing the resilience of networked infrastructure." *IEEE Systems Journal*, 3(2), pp.174-180.
- Rinaldi, S.M. 2004. "Modeling and Simulating Critical Infrastructures and Their Interdependencies." *Proceedings of the 37th Annual Hawaii International Conference on System Sciences*, IEEE.
- Saini, A. and Tien, I. "Impacts of climate change on the assessment of long-term structural reliability." *ASCE-ASME Journal of Risk and Uncertainty in Engineering Systems*. 3(3).

- Sakaki, T., Okazaki, M., and Matsuo, Y. 2010. "Earthquake Shakes Twitter Users: Real-Time Event Detection by Social Sensors." *Proceedings of the 19th International Conference on World Wide Web*, 851-60.
- Smith, L. 2005. "Louisiana directory of cities, towns, and villages." Louisiana: Sims Memorial Library – Southeastern Louisiana University.
- Smith, L., Liang, Q., James, P., and Lin, W. 2015. "Assessing the utility of social media as a data source for flood risk management using a real-time modelling framework." *Journal of Flood Risk Management*.
- Stieglitz, S., Mirbabaie, M., Ross, B., Neuberger, C. 2018. "Social Media Analytics – Challenges in Topic Discovery, Data Collection, and Data Preparation. *International Journal of Information Management*, 39: 156-168.
- Tien, I., Musaev, A., Benas, D., Ghadi, A., Goodman, S., and Pu, C. 2016. "Detection of Damage and Failure Events of Critical Public Infrastructure Using Social Sensor Big Data." *Proceedings of the International Conference on Internet of Things and Big Data (IoTBD 2016)*, 435-40.
- Tsai, J. and Chatterjee, A. 2018. "Pothole Detection and Classification Using 3D Technology and Watershed Method." *ASCE Journal of Computing in Civil Engineering*.
- Twumasi-Boakye, R. and Sobanjo, J. 2018. "Resilience of regional transportation networks subjected to hazard-induced bridge damages." *ASCE Journal of Transportation Engineering, Part A: Systems*. 144(10).
- United States Census Bureau. 2010. Georgia Census Block Maps.
- United States Department of Energy. 2019. "What is the smart grid?" Office of Electricity Delivery and Energy Reliability. https://www.smartgrid.gov/the_smart_grid/smart_grid.html.
- United States Geological Survey (USGS). 2017. "National water information system (NWIS) data available on the World Wide Web (Water Data for the Nation)."
- University of Exeter. 2019. Center for Water Systems: Benchmarks. emps.exeter.ac.uk/engineering/research/cws/resources/benchmarks/.

- Van Der Wiel, K., S. B. Kapnick, G. Jan Van Oldenborgh, K. Whan, S. Philip, G. A. Vecchi, R. K. Singh, J. Arrighi, and H. Cullen. 2017. "Rapid attribution of the August 2016 flood-inducing extreme precipitation in south Louisiana to climate change." *Hydrol. Earth Syst. Sci.* 21 (2): 897–921. <https://doi.org/10.5194/hess-21-897-2017>.
- Vieweg, S., Hughes, A., Starbird, K., and Palen, L. 2010. "Microblogging during two natural hazard events." In *Proceedings of the 28th Int'l Conference on Human Factors in Computing Systems*. 1079-1088. New York: ACM.
- Vieweg, S., Palen, L., Liu, S.B., Hughes, A.L., and Sutton, J. 2008. "Collective Intelligence in Disaster: An Examination of the Phenomenon in the Aftermath of the 2007 Virginia Tech Shootings." *Proceedings of the 5th International Conference on Information Systems for Crisis Response and Management, ISCRAM 2008*.
- Wang, Q. and Taylor, J.E. 2015. "Process Map for Urban-Human Mobility and Civil Infrastructure Data Collection Using Geosocial Networking Platforms." *ASCE Journal of Computing in Civil Engineering*.
- Watson, K.M., Storm, J.B., Breaker, B.K., and Rose, C.E. 2017. "Characterization of peak streamflows and flood inundation of selected areas in Louisiana from the August 2016 Flood." *U.S. Geological Survey Scientific Investigations Report No. 2017-5005*. Reston, VA: U.S. Geological Survey.
- White House. 2013. "Presidential Policy Directive – Critical Infrastructure Security and Resilience." <https://obamawhitehouse.archives.gov/the-press-office/2013/02/12/presidential-policy-directive-critical-infrastructure-security-and-resil>
- Yin, J., Lampert, A., Cameron, M., Robinson, B., and Power, R. 2012. "Using Social Media to Enhance Emergency Situation Awareness." *International Joint Conference on Artificial Intelligence*, 4234-39
- Yuan, F. and Liu, R. 2018. "Integration of Social Media and Unmanned Aerial Vehicles (UAVs) for Rapid Damage Assessment in Hurricane Matthew." In *Construction Research Congress 2018* (pp. 513-523).
- Yuan, F. and Liu, R. 2020. "Mining Social Media Data for Rapid Damage Assessment During Hurricane Matthew: Feasibility Study." *ASCE Journal of Computing in Civil Engineering*.

Zhang, X., Miller-Hooks, E., and Denny, K., 2015. "Assessing the role of network topology in transportation network resilience." *Journal of Transport Geography*, 46:35-45.

Zhang, C., Fan, C., Yao, W., Hu, X., and Mostafavi, A. 2019. "Social Media for Intelligent Public Information and Warning in Disasters: An Interdisciplinary Review." *International Journal of Information Management*. 49:190-207.

Zhou, Z., Gong, J., and Guo, M. 2015. "Image-based 3D Reconstruction for Posthurricane Residential Building Damage Assessment." *ASCE Journal of Computing in Civil Engineering*.