STUDY OF NETWORK-SERVICE DISRUPTIONS USING HETEROGENEOUS DATA AND STATISTICAL LEARNING

A Thesis
Presented to
The Academic Faculty

by

Supaporn Erjongmanee

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in the
School of Electrical and Computer Engineering

Georgia Institute of Technology
May 2011

STUDY OF NETWORK-SERVICE DISRUPTIONS USING HETEROGENEOUS DATA AND STATISTICAL LEARNING

Approved by:

Chuanyi Ji , Advisor
School of Electrical and Computer Engineering
Georgia Institute of Technology

John Copeland
School of Electrical and Computer Engineering
Georgia Institute of Technology

Biing Hwang (Fred) Juang
School of Electrical and Computer Engineering
Georgia Institute of Technology

Russell Clark
College of Computing
Georgia Institute of Technology

George Riley
School of Electrical and Computer Engineering
Georgia Institute of Technology

Date Approved: 18 January 2011

To Mom and Dad,

The true wind beneath my wings

ACKNOWLEDGEMENTS

This thesis is a collaboration of support that I have received throughout my study. First, I would like to express my gratitude to my advisor, Dr. Chuanyi Ji, for her guidance, support, and encouragement. I greatly appreciate her devotion and patience.

I would like to thank my thesis committee members, Dr. Fred Juang, Dr. George Riley, Dr. John Copeland, Dr. Russell Clark, and Dr. Magnus Egerstedt, for their valuable comments. I would like to thank Dr. Ian Ferguson for his guidance and encouragement.

In addition, I would like to thank Paul Saldarriaga from MaxMind for GeoIP data, Dr. Lixia Zhang and Lucas Wang from UCLA for studied prefixes, Jere Stokely and Neale Hightower for information on network restoration, Robert Berg from National Hurricane Center for hurricane knowledge, Lans Rothfusz from National Weather Service for NCDC data, Sarah Swanson from Public Utility Commission of Texas for power outage reports, Dr. Gary May for help with seeking root causes, the network administrators from Houston Advanced Research Center, University of Texas Medical Branch at Galveston, Internet America, and NASA for providing their root causes, Yun Wang for collaboration on storm correlation, David Green and Richard Davies for developing graphic tools, Guang Cheng for data processing, Pat Dixon and Cordai Farrar for administrative assistance.

I also would like to thank Grandma Nell Cole, Dawn Alford, Jean Alford, and Johnny Alford, for their love, support, and encouragement. I also would like to thank Patrice Harduar Gopo and Pakorn Kanchanawong for being with me from start to finish.

Finally, I would like to thank my family. I would like to thank my sisters, Suphawadee Erjongmanee and Suchada Erjongmanee, for their motivation, support, and love. Last, I would like to deeply thank my parents, Sunee Erjongmanee and Adul Erjongmanee, for their unconditional love and patience, complete understanding, and endless support. Without my parents, this work will not be possible, and I dedicate this thesis to them.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

SUMMARY

Communications comprises a key infrastructure that supports every aspect of our daily lives. In the past ten years, large-scale disturbances have proven to cause extensive damage to the communications infrastructure and require millions of dollars to repair the damage. Considerable attentions in the prior work have focused on assessing network damages after natural or man-made disasters. However, network-disruption responses, i.e., how the disruptions occur depending on social organizations, weather, and power resources, have been studied little.

The objective of this research is to study network-service disruptions caused by large-scale disturbances with respect to (1) temporal and logical network, and (2) external factors such as weather and power resource, using real and publicly available heterogeneous data that are comprised of network measurements, user inputs, organizations, geographic locations, weather, and power outage reports.

In this study, network-service disruptions are studied at the subnet level. A subnet is a collection of connected computer devices generally owned by an organization. An unreachability of a subnet occurs if the Internet traffic can no longer route to this subnet.

Network-service disruptions caused by Hurricanes Katrina in 2005 and Ike in 2008 are used as the case studies. First, the identification of subnet unreachability is developed by applying unsupervised- and semi-supervised learning to large-scale network measurements and user inputs.

The network-disruption responses are studied with respect to temporal and logical network dependence. It is found that temporal dependence also illustrates the characteristics of logical dependence. Temporally dependent subnets became unreachable within organization, cross organization, and cross autonomous system. The comparison of subnet unreachability between Ike and normal operations illustrates that subnet unreachability due

to Hurricane Ike is indeed anomalous.

In addition, subnet unreachability is analyzed with respect to the storm characteristics. The storm path and coverage are reconstructed from the storm data, and the times and probabilities when the storm coverage overlapped subnet regions are computed. As a result, it is found that subnet unreachability and the storm are weakly correlated.

The weak correlation between subnet unreachability and the storm provides the motivation to search for what exactly caused subnets to become unreachable. We contacted organizations who own unreachable subnets to learn about their actual root causes. Finding root causes of disrupted networks from the subnet owners is challenging since such information is proprietary to organizations. Six out of seven organizations reportedly experienced network disruptions due to power outages or the lack of power generators.

Using power outage data obtained from the Public Utility Commissioner of Texas, the dependence of subnet unreachability on power outages is studied. The network data is aggregated to the same scale as the power outage data. The observations and correlation illustrate that subnet unreachability and power outages are strongly correlated.

The information sharing potentially can be used to improve the state of the art towards studying network-service disruptions. We explore more sharing information resources in weather, power, Internet service providers, daily lives, and emergencies. The best-fit data needed for the dependence study of network disruptions caused by large-scale disturbances are also presented.

This contribution of this thesis is the empirical study of network-service disruptions caused by large-scale disturbances using real and publicly available heterogeneous data and statistical learning. We incorporate network-, weather-, and power-related data into the analysis of network-service disruption caused by large-scale disturbances with respect to temporal and logical network, and external factors such as weather and power resources.

CHAPTER I

INTRODUCTION AND RESEARCH OBJECTIVES

Communications comprises a key infrastructure that supports every aspect of our daily lives. Consequently, the networking infrastructure needs to always be resilient and available. Since 2001, intensive and large-scale disturbances have proven to cause large-scale damage on the networking infrastructure. This poses a significant challenge as the networking infrastructure is coupled with social entities such as organizations and is further exposed to external factors such as weather and power resources. To enhance network resilience, we must gain a better understanding on interactions between networks, organizations, weather, and power resources.

1.1 Motivation

Large-scale disturbances push the communications infrastructure to the extreme and expose the interactions between networks, organizations, weather, and power resources that are not observable in day-to-day operations. Several examples of large-scale disturbances are presented below.

1. **September 11 attack**

   In 2001, the September 11 attack caused network outages resulting from fiber-optic cable damage and power outages [28]. Since some transatlantic inter-connections were routed through the infrastructure in New York City, the impact of network outages spread as far as Europe and Africa.

2. **Northeastern America blackout**

   In 2003, the loss of power took place during the blackout in the northeastern United States and Canada [33]. Electricity as well as network connectivity, were restored within four days. There was no significant damage reported on computer software or hardware.

3. **Hurricanes**

   In 2005, Hurricane Katrina caused large-scale disruptions in telecommunication networks. After Katrina, three million telephone lines were out of service. More than 1,000 wireless sites and 38 emergency 9-1-1 call centers went down [63]. Network connectivity was critical, but it was either unavailable or unstable according to disaster responders [63, 78].

   In 2008, Hurricane Ike reportedly caused 168 networks in Texas to experience service disruptions [18]. Suddenlink, the major Internet Service Provider (ISP) in the disaster area, reported the outages of their networks were caused by power outages and fallen trees [23].

4. **Earthquakes**

   In 2006, the Taiwan earthquake broke seven out of nine submarine cables that routed telecommunications services throughout Asia and caused communication losses in at least 14 countries. The impact spread from Taiwan to more distant countries, e.g., India and Pakistan [11, 19].

   In January 2010, the Haiti earthquake caused severe damage to undersea fiber-optic cables [5, 52]. The Chile earthquake in February 2010 badly damaged fixed-line and mobile telecommunications, and most of the Internet connectivity was disrupted [14]

5. **Wildfires**

   In 2007, California wildfires damaged communications infrastructure and caused broadband, telephone, and the Internet outages in local areas [1].

The above examples provide motivation to study network-service disruptions caused by large-scale disturbances. Considerable attentions in the prior work have focused on identifying network statuses after natural or man-made disasters [18, 19, 31, 33, 42, 83]. However, the network-disruption responses, i.e., how network disruptions occur depending on social organizations, weather, and power resources, have been little studied.

Network responses to disruptions caused by large-scale disturbances have mostly not been characterized due to many technical challenges. First, it remains open what types of data can reveal the complex interactions among networks, organizations, weather, and power resources. The second technical challenge is how to identify and untangle dependent variables to expose the network-disruption responses. The final technical challenge is the presence of social issues. Exposing information about network disruptions is often prohibited by organizations who own networks due to privacy, security, and business competition. In fact, it is commonly accepted as impossible for organizations to share information on their network disruptions. This is evident from the fact that most available reports of the previous network disruptions come from third parties [18, 19, 31, 33, 42, 83], and the available data to study the network disruptions has limited access. Here, our focus is to provide real and publicly available detailed study of network-service disruptions caused by large-scale disturbances.

1.2 Research Objectives

The objective of this research is to analyze and to understand the responses of network-service disruptions caused by large-scale disturbances with respect to (1) temporal and logical network, and (2) external factors such as weather and power resources, using real and publicly available heterogeneous data that consists of network measurements, user inputs, organizations, geographic locations, weather, and power outage reports.

Specifically, we search for answers to the following questions:

1. How do networks respond to large-scale disturbances with respect to temporal and logical networks and external factors?

2. What types of heterogeneous data that is publicly available can help discover the network responses to large-scale disturbances?

3. What approaches can be used to learn the network responses from the heterogeneous data?

4. What are the limitations of existing data, and what is needed to improve the study of network-service disruptions caused by large-scale disturbances?

In this research, network-service disruptions caused by Hurricanes Katrina in 2005 and Ike in 2008 are used as case studies. Our focus is also to have this study as a template to study network disruptions for future large-scale disturbances.

1.3 Thesis Outline

The thesis outline, including data sources, approaches, and results, are summarized below.

1.3.1 Development of Network-Disruption Identification

The research starts with the identification of network-service disruptions. In this work, the network-service disruptions are studied at the subnet level. A subnet is a collection of connected computer devices generally owned by an organization. The Internet routes network traffic from one subnet to the others. An unreachability of a subnet occurs if the network traffic can no longer route to this subnet.

Subnet unreachability caused by Hurricane Katrina in August 2005 is used as an example in this study. The 1,009 subnets in three states affected by Hurricane Katrina, Louisiana, Mississippi, and Alabama, are selected, and the Border Gateway Protocol (BGP) routing messages [72] are used as network measurements to identify unreachable subnets.

The machine learning approaches of clustering and feature extraction are applied to BGP measurements to reduce the dimensionality. As a result, the spatial dimensionality is reduced by 81%, and the temporal space is extracted to two features. Next, the representative subnet of each cluster is selected to have its unreachable status identified. Here, the user inputs, i.e., the reports by the Internet users on network outages, are obtained, and semi-supervised learning is applied to the extracted features of BGP measurements and user inputs to derive the classifier to infer statuses of subnets.

Two thresholds are obtained separately for each of the two different features and applied to the features of the representative subnets to determine subnet statuses. After validation, it is found that 25% of the studied subnets are identified as unreachable, and 43% of the

subnets exhibit behavior different from normal operations. There are respectively 50.79% and 72.83% of subnets that became unreachable before landfall and had unreachability durations longer than one month.

1.3.2 Understanding Temporal and Logical Characteristics of Network Disruptions

In this part of the research, the characteristics of subnet unreachability with respect to temporal dependence and logical characteristics are analyzed. Subnet unreachability due to Hurricane Ike in 2008 is used as the next case study. 3,601 subnets in Texas are selected, and BGP measurements between and prior to Hurricane Ike are collected. More data sources on organizations, Autonomous Systems (ASes), and ISPs of subnets are added in this study. The initial times–the times when subnets first became unreachable and carry particular information relating to the hurricane–and the unreachability durations when network disruptions occurred are also included.

A hypothesis test shows that the unique initial times of subnet unreachability occurred independently. Then, it is found that temporally dependent subnets became unreachable within organization, cross organization, and cross AS. Thus, temporal dependence also illustrates the characteristics of logical dependence.

Beside the analysis of temporal and logical dependence, subnet unreachability is compared with the pre-Ike period to prove that subnet unreachability after Hurricane Ike is indeed anomalous.

1.3.3 Analysis of Network Disruptions on Weather Dependence

Studying subnet unreachability caused by Hurricane Ike raises the next question: when did subnet unreachability occur with respect to the storm?

In this part of the research, two data sources are added: geo-locations of subnets and the storm data. The geo-locations of all IP-addresses for each subnet are obtained from GeoIP database [49]. GeoIP reports that approximately 79% of their geo-locations provided are accurate within 25 miles from the true location. Thus, this uncertainty in geo-location measurements is incooperated by using disks with 25-mile radius as subnet regions. The

hurricane path and coverage are constructed from the storm center, storm speed, and wind radii at hurricane force winds. Next, the hitting times, i.e., the times when the hurricane coverage first overlapped the subnet regions, are computed and correlated with the initial times of subnet unreachability. As a result, it is found that subnet unreachability is weakly correlated with the storm.

In addition to the correlation of storm data, the subnet unreachability caused by Hurricane Ike are also compared to subnet unreachability due to Hurricane Katrina.

1.3.4 Searching for Ground Truth

The weak correlation between subnet unreachability and the storm provides the motivation to search for the ground truth of subnet unreachability. A ground truth is an actual root cause of network-service disruption. Generally, the root causes of disrupted networks are speculated to be physical damage, power outages, and evacuation. Since it is found that subnet unreachability depends on organizations, to discover the ground truth, the very local information sources are contacted: organization and ISPs who own subnets.

Finding the ground truth from the subnet owners is challenging since such information is proprietary to organizations and ISPs. From more than 40 organizations contacted, only four responses were received. Besides the root causes, the information obtained also includes the restoration processes. In addition to these four organizations, the root causes for three additional organizations obtained online are included.

The findings illustrate that the common root cause among six out of seven organizations is related to power resources. Specifically, three organizations experienced power outages. Two organizations experienced power-backup failures, and the other one shut down their network due to the lack of power generators. This suggests dependence of network-service disruptions on power outages.

1.3.5 Preliminary Investigation of Network Disruptions on Power-Resource Dependence

The ground truth search shows that network disruptions caused by Hurricane Ike are related to power resources. Therefore, this provides motivation to study the dependence between

network disruptions and power resources.

Another data set is introduced, namely, the power outage data. Since power infrastructure is related to national security, in general public availability of power data is extremely rare. In this study, the power outage data is requested from the Public Utilities Commissioner of Texas (PUC). This PUC data consists of the daily reports on the number of customers without electricity from 63 counties between September 13-20, 2008.

The PUC data are available at spatial and temporal scales of county and day, whereas the spatial and temporal scales of network data are 25-mile radius disks and seconds respectively. Hence, the network data is aggregated to have the same scale as the PUC power data. The observations and correlation show that subnet unreachability and power outages are strongly correlated.

1.3.6 Investigation of Information Sharing

Information sharing networks provide the most refined spatial and temporal scales, i.e., where and when exactly network disruptions occur and what are the root causes. We explore existing information sharing networks used in many communities.

Shared information on weather data can be obtained from National Climate Data Center (NCDC) [51] and Community Collaborative Rain, Hail and Snow network (CoCoRaHS) [29]. More shared power data can be obtained from I-Grid, the online distributed power monitoring system, and some of I-Grid data is publicly available.

More information sharing networks used in other communities such as ISPs, daily lives, and emergencies are presented. In addition, we discuss the use of information sharing for network disruptions and power outages.

In this research, one major challenge is the lack of publicly available data. Thus, the best-fit data for the dependence study of network-service disruptions on large-scale disturbances such as weather and power resources is discussed. Here, the goal is to raise awareness on the effectiveness of information sharing to improve the study of network disruptions caused by large-scale disturbances.

1.3.7 Conclusion and Future Research Directions

The contribution of this thesis includes the empirical study of network-service disruptions to large-scale disturbances using real and publicly available heterogeneous data and statistical learning. We incorporate network-, weather-, and power-related data into the analysis of network-service disruption caused by large-scale disturbances, and we analyze the network responses to disruptions with respect to (1) temporal and logical network, and (2) external factors such as weather and power resources.

We introduce the organization variables in this network study and discover that the logical dependence of network disruptions lies within organizations and ASes. By including the weather data, we find that subnet unreachability and storm are weakly correlated. Last, our search for the ground truth reveal that network disruptions are dependent on power resources. Using the publicly available and aggregated power outage data, our study finds that subnet unreachability and power outages are strongly correlated.

Future research directions include a detailed study of the dependence between network disruptions, power resources, and weather.

CHAPTER II

DEVELOPMENT OF NETWORK-DISRUPTION IDENTIFICATION

The first objective of this research is to identify and diagnose large-scale network-service disruptions using heterogeneous information sources and the application of machine learning. Although most prior work has focused on diagnosing either sporadic network failures in day-to-day network operations or on one-time global network failures, few work has focused on network-service disruptions caused by large-scale disturbances.

The network-service disruptions caused by Hurricane Katrina are the motivation for this study. Hurricane Katrina caused large-scale physical damage including the Internet infrastructure in the disaster area. After landfall, network connectivity in the disaster area is critical for communications among disaster responders. However, the connectivity is either unavailable or unstable.

Quality of service and reliability are what Internet customers are looking for from Internet Service Providers (ISPs). Hence, when network failures occur, ISPs typically keep the record of failures confidential. Most reports of previous widely affected network failures were from either customers or third parties [18, 19, 31, 33, 42, 83]. Moreover, when large-scale network failures occur, different ISPs do not exchange network failure data among one another because of business competition. This also hinders the recovery of the Internet connectivity.

Our goal is to diagnose network-service disruptions caused by large-scale disturbances in order to obtain a diagnosis that is available to anyone and is not restricted to any particular ISP. Disaster responders and ISPs can use this diagnostic result to understand the details of network damage and seek for available network resources. This information is also vital for re-establishing the network connections and for repairing the network infrastructure.

Chapter 2 is organized as follows. Section 2.1 provides background of the Internet and Hurricane Katrina. Section 2.2 introduces BGP data and user inputs used in this study.

Section 2.3 presents related work, and Section 2.4 provides problem formulation. Section 2.5 illustrates the use of unsupervised learning, i.e., clustering and feature extraction to BGP measurements, and Section 2.6 presents the semi-supervised learning to obtain thresholds to determine subnet statuses from the extracted features of BGP data and user inputs. Sections 2.7 and 2.8 respectively provide the results and the remarks from the ISP. Section 2.9 summarizes Chapter 2.

2.1 Background

Before proceeding to the network-disruption identification, background information on the Internet and Hurricane Katrina is provided below.

2.1.1 Border Gateway Protocol

The Internet consists of interconnected autonomous systems (AS). Each AS is served by at least one ISP and is composed of one or several subnets identified by prefixes (network addresses). Figure 1 shows an example of AS network where X is an AS, and X $\in$ {A,B,...,G}.

The routing protocol among ASes is the Border Gateway Protocol (BGP) [72]. To route traffic from one AS to a specific subnet at another AS, a BGP peer router at each AS collects streams of routing messages from BGP peer routers of its neighbor ASes. Figure 1 shows an example where x is a BGP peer router of AS X, X $\in$ {A,B,...,G}, and AS E has two prefixes p and q. These messages are called BGP update messages.

There are two types of BGP update messages: BGP withdrawals and BGP announcements. An unreachability of a subnet occurs if the Internet traffic can no longer route to this subnet. When a subnet becomes unreachable, BGP peer routers withdraw routes to the subnet by sending multiple BGP withdrawals to other peer routers [72]. When the subnet becomes reachable, multiple BGP announcements are sent among BGP peer routers to establish new routes. Thus, BGP update messages have been used to remotely infer subnet unreachability [42, 46, 87].

Besides subnet unreachability, BGP announcements and withdrawals can be used to update routing information. Examples of routing information include a change of routes or an update of routing policies. Hence, a burst of multiple BGP withdrawals followed

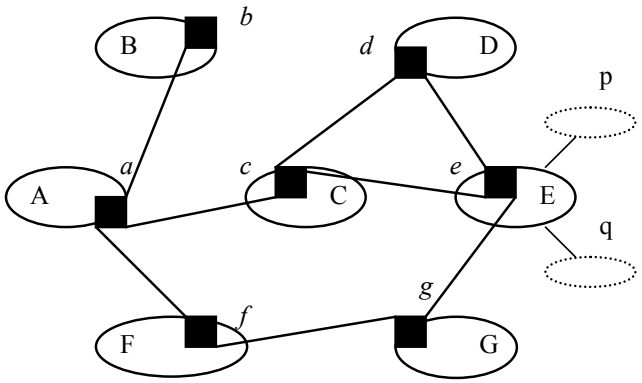


Figure 1: Example of the Internet, where X is an AS, and X $\in$ {A,B,...,G}. x is a BGP peer router of AS X, and p and q are two prefixes of subnet E.

by new BGP announcements is a symptom rather than a one-to-one mapping of network disruption [46, 87].

2.1.2 Hurricane Katrina

2005 saw more hurricanes than in any previous years since 1969. There were seven major hurricanes of category three and higher [13], and the most severe hurricane was Hurricane Katrina which also became the most costly hurricane in U.S. history.

Hurricane Katrina made landfall in the Gulf Coast on August 29, 2005, at approximately 6:00 a.m. and flooded Louisiana, Mississippi, and Alabama[1]. Katrina caused considerable large-scale disruption in telecommunications [63]. Network connectivity in the disaster area, critically needed by disaster-responder organizations such as hospitals and government agencies, became either unavailable or unstable [63, 78].

Despite reports of Katrina's impact on telecommunications, there were only a few public reports showing the impact of Katrina on the Internet communications [32, 79]. Since subnets are directly connected to the responder organizations after disasters, further study is needed on detailed network-service disruptions at the subnet level.

[1]The reported time here is Central Daylight Time (CDT), i.e., the local time in Louisiana, Mississippi, and Alabama.

2.2 Heterogeneous Data

In this work, we study network-service disruptions caused by Hurricane Katrina at the subnet level. Two types of measurements are used in this work. First, the network measurements are obtained from BGP routing messages. Second, the data on user inputs are collected from user reports on network outages.

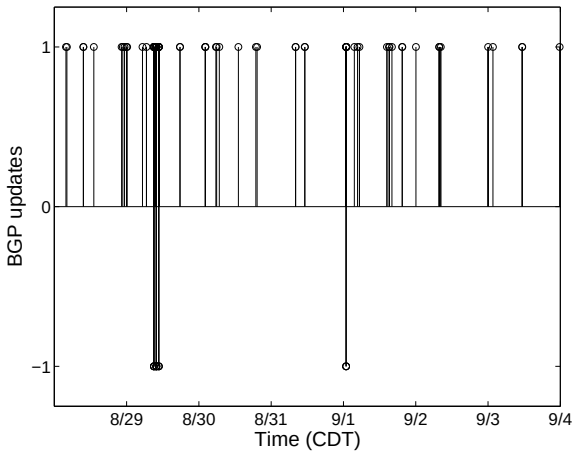
2.2.1 Network Measurements

BGP update messages can provide remote monitoring of network-service disruptions when local measurements are not directly available due to evacuation or limited accessibility to a disaster area.

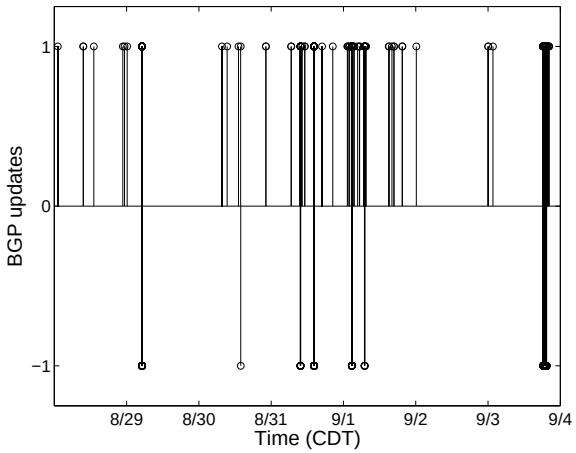
Geographic locations are pertinent for selecting subnets in a disaster area to study. In this study, the disaster area includes three states largely affected by Katrina: Louisiana (LA), Mississippi (MS), and Alabama (AL). The geographic locations of subnets are obtained from the Whois database [85], and we select 1,009 subnets from 48 ASes in the disaster area. This results in 1,009-dimensional time-series of BGP measurements. Figures 2(a), 2(b), 2(c) illustrate three examples of time-series; each exhibits distinct temporal characteristics.

The study duration of the Katrina interval is between August 28 and September 4, 2005, from the mandatory evacuation to the first week after the landfall. We collaborate with one ISP in the disaster area, who informed us that assessing network damage was mostly done within the first week after the landfall. In addition, for comparison with normal operations, BGP update messages belonging to the same set of subnets between August 1-28, 2005 are included in the study. This time duration is referred to as the pre-Katrina interval. With 1,009 subnets and an eight-day duration, the BGP measurements are considered to be large-scale both spatially and temporally.

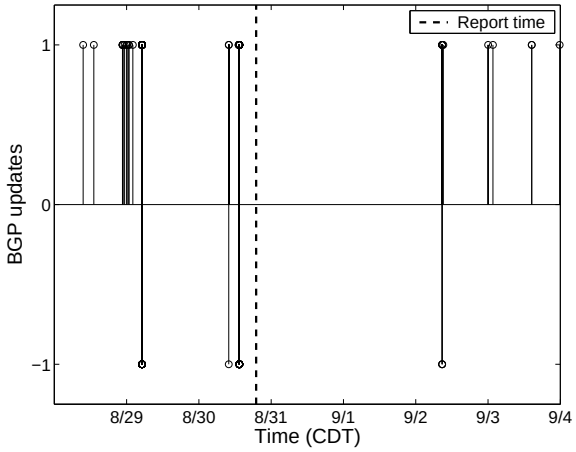
The 1,009 subnets are partitioned into eight subnet sets based on eight different local ISPs and their customer ASes. The partition of subnets imitates how subnets are serviced and maintained separately by different ISPs. The ISPs are local service providers who are business competitors; thus, they do not usually exchange their network-service data. Table



(a) First time-series example with unknown status



(b) Second time-series example with unknown status



(c) User input time-series

Figure 2: Examples of two time-series measurements with unknown statuses and user-input time-series. (1 = BGP announcement, -1 = BGP withdrawal.)

Table 1: List of subnet sets (LA = Louisiana, MS = Mississippi, AL = Alabama)

Subnet set	1	2	3	4	5	6	7	8
Number of subnets	166	27	49	115	26	180	232	214
Geographic location	LA	LA	LA	LA	LA	LA	MS	AL

1 shows the number of subnets in each subnet set and their geographic locations. Note that the last three subnet sets are obtained from our collaborating ISP.

As described in [46, 87], a burst of BGP update messages is a symptom rather than a one-to-one mapping of subnet unreachability. As a result, it is insufficient to use BGP measurements alone to infer unreachability of all subnets.

2.2.2 User Inputs

User inputs are the reports that "this network is down." Therefore, they provide valuable and mostly accurate information on statuses of networks. However, user inputs can be much delayed from the exact time when an unreachability occurs. For this study, we collect 37 user inputs located in the heavily damaged disaster area from two sources.

The first 28 user inputs are obtained from the online message on the NANOG mailing list posted by Todd Underwood from Renesys Corporation [79]. An example entry of this report is "Loyola University LA 70118 (141.164.0.0/16)." This means that by the time that this message was posted, the subnet 141.164.0.0/16 belonging to Loyola University and located in Louisiana with zip-code 70118 was unreachable. This online message does not provide complete information of user inputs such as initial times and durations of subnet unreachability.

The other nine user inputs are reports from customers of our collaborating ISP. Customers usually send a report based on either their perception of an outage, e.g., unsuccessful attempt to connect to the network for a humanly tolerable time, or physical evidence, e.g., power or hardware failures. These types of user inputs usually are proprietary to ISPs. In this work, we are provided with nine user inputs that contain the unreachable subnets of the customers and the time when the reports were received.

Using only these 37 user inputs may not be sufficient for inferring statuses for the remaining subnets. Hence, BGP measurements and user inputs complement each other in

inference of network-service disruptions.

2.3 Related Work

Machine learning has previously been used with BGP data focusing on day-to-day networks operations. Feamster *et al.* inferred a BGP topology by applying the single-linkage agglomerative clustering to BGP update messages [8]. Chang *et al.* identified the cause of path changes by using the single-linkage and the complete-linkage clustering algorithms to cluster ASPATHs [26]. Xu *et al.* proposed the use of principal component analysis (PCA) to BGP updates to infer different BGP events[87] while Zhang *et al.* detected BGP anomalies by applying wavelet transformation to BGP updates belonging to an individual subnet and to many subnets together at one time [90]. Although machine learning has been applied to BGP data in the prior work, these applications do not focus on network disruptions caused by natural disasters.

User data has been used in supervised learning to infer root causes of network failures using probabilistic models [9, 56, 60]. These studies used data from day-to-day operations. Moreover, they relied on complete knowledge of network status and complete underlying inference models. However, natural disasters are rare events. Thus, knowledge of network disruptions caused by natural disasters is incomplete, and no underlying model of network disruptions is available.

Semi-supervised learning is a cross-learning between using labeled (supervised) and unlabeled (unsupervised) data [27]. A seminal work [22] studied the information contents of labeled and unlabeled measurements and showed that both unlabeled and labeled data can be used for classification. Thus, this prior work has established the fundamental use of unlabeled and labeled measurements in pattern classification.

Three major types of semi-supervised learning algorithms are described in [27], i.e., generative models, transductive support vector machine, and graph-based models. Generative models use labeled and unlabeled data to learn the mixture models [67]. Transductive support vector machine uses the low density separation between unlabeled and labeled data to determine the classification boundaries [36, 55]. Graph-based models build the graph with

labeled and unlabeled data as nodes and distance measures among data as edges [10, 16]. Semi-supervised learning has been used in many applications. The most commonly used application is in text classification [17, 55, 67, 88]. The examples of other applications are remote sensing [75] and image processing [10, 61]. To our knowledge, there has not been any prior work that applies semi-supervised learning in the study of network disruptions.

2.4 Problem Formulation

The problem of inferring network-service disruptions using heterogeneous data can be formulated as follows.

Consider an underlying network with n nodes, where a node corresponds to a subnet. Let $Z_i(t)$ be a binary state of a subnet i for $1 \leq i \leq n$:

$$Z_i(t) = \begin{cases} 1 & \text{if node } i \text{ is outage (unreachable),} \\ -1 & \text{if node } i \text{ is normal (reachable),} \end{cases} \tag{1}$$

where $1 \leq i \leq n$, and $t \in [0, T]$ is a time duration of interest. Network-service disruption is defined as unreachability of subnets[2]. The state of a network is a collection of all n node states: $Z(t) = \{Z_i(t)\}_{i=1}^{n}$, $t \in [0, T]$, and considered to be unknown.

Let $X(t) \in R^n$ be an n-dimensional random vector, and this can be viewed as "response variables." Let $x(t)$ be a sample of $X(t)$, for $t \in [0, T]$. A sample $x(t)$ corresponds to BGP measurements and is assumed to be obtained when the underlying network state $Z(t)$ is unknown. Therefore, $x(t)$ is called an unlabeled sample [39].

A set D of m unlabeled samples is assumed to be available for m nodes. Therefore, there is 1,009-dimensional unlabeled data as shown in Section 2.2. Because the size of unlabeled data is large, the unlabeled data itself is insufficient to determine the underlying network state $Z(t)$ unless it is empowered by discriminative information [91].

User inputs provide discriminative information. A set D_l of k user inputs, where k can be small, i.e., $0 \leq k \ll m$, are assumed to be available for a fraction of subnets. The simplest form of a user input is a report of an unreachable subnet from a responder, i.e.,

[2]We use unreachability, outage, and network disruption interchangeably. We also use subnet and prefix interchangeably.

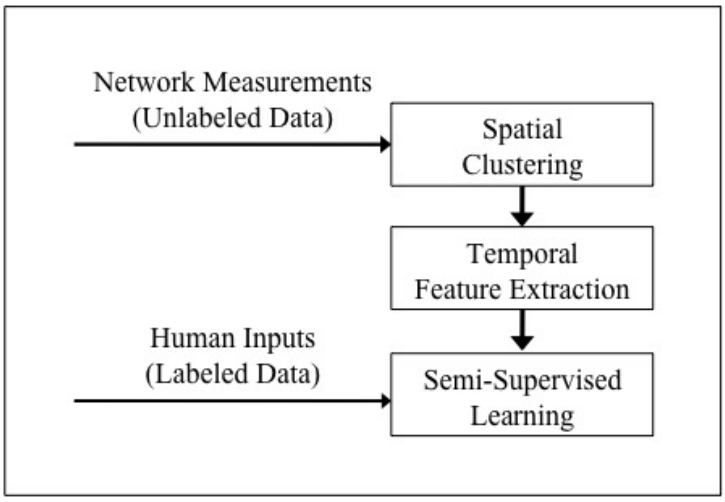


Figure 3: Approaches for inferring network-service disruptions.

"1" for state $Z_i(t)$ of a subnet i at time t, where $1 \leq i \leq n$. Let t_i be the time that subnet i becomes unreachable. An assumption is made that user reports unreachability of a subnet correctly[3] but with a delay, i.e., $t_{report} > t_i$.

In this work, 1,009 unlabeled BGP measurements and 24 user inputs are used for training. The other 13 user inputs are used for validation. Hence, $k = 24$, $m = n - k = 985$.

Problem: Given a set D of unlabeled network measurements, and a set D_l of user inputs, infer $Z(t)$ for $t \in [0, T]$.

The inference, i.e., dichotomies between outage and normal states of subnets can be learned from network measurements and user inputs. Our approaches are summarized in Figure 3 and outlined below.

- We first apply unsupervised learning to network measurements to perform clustering and feature extraction. Clustering is used to group subnets with similar temporal characteristics. Thus, clustering reduces the spatial dimension of the network time-series. We then extract temporal features from time-series in $[0, T]$ to a low-dimensional feature space and use these features as unlabeled data.

- We apply a semi-supervised learning algorithm. We convert 24 user inputs reported as unreachable subnets to labels "1." Then, we obtain 24 labeled temporal features

[3]Note that it is a natural assumption that user only reports when a network is experiencing outage.

in the low-dimensional feature space. We also collect a set of temporal features that correspond to a normal network operations and label them as "-1." Next, a large set of unlabeled data and a small set of labeled data are combined and used to infer the statuses of subnets.

- We validate the inference result of the subnet statuses.
- We provide an initial understanding of network-service disruptions caused by Hurricane Katrina

2.5 Unsupervised Learning

We now apply unsupervised learning to extract features from 1,009 time-series measurements belonging to the 1,009 subnets. The first step is to cluster these time-series to reduce the spatial dimension. The second step is to extract temporal features from the time-series.

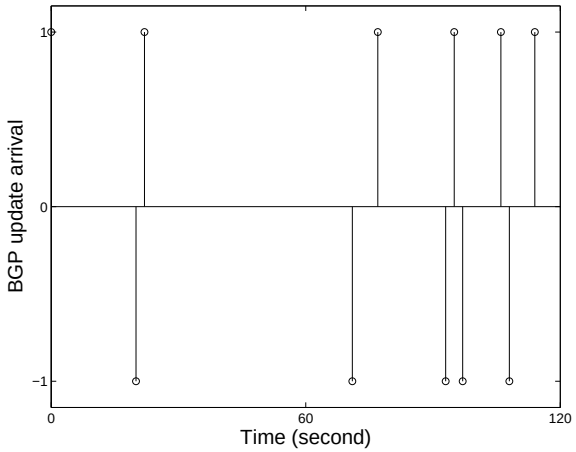
2.5.1 Spatial Clustering

Intuitively, subnets in the same disaster area may have experienced correlated network-service disruption due to impact from the same disaster. Therefore, we analyze the spatial correlation by first reducing the spatial dimensionality of time-series. The goal is to group similar time-series together.

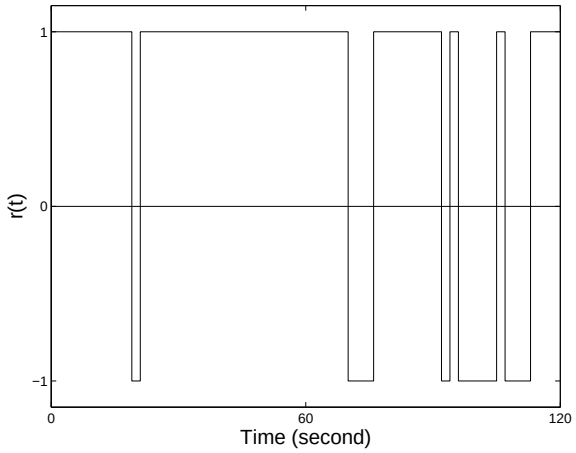
2.5.1.1 Clustering Algorithm

To measure the similarity of the time-series, we convert a discrete time-series of BGP update messages to a continuous waveform such that: when a BGP announcement of subnet i arrives at time t, the continuous waveform $r_i(t) = 1$; otherwise, for a BGP withdrawal, $r_i(t) = -1$, where $1 \leq i \leq n$, and n is number of subnets. Consider time t between two consecutive arrivals of BGP updates at times t_1 and t_2: $r_i(t) = r_i(t_1)$, where $t_1 \leq t < t_2$. For subnet i without BGP update arrival, $r_i(t) = 1$ for all $t \in [0, T]$. Figures 4(a) and 4(b) show an example of BGP updates and its corresponding $r(t)$.

The similarity between $r_i(t)$ and $r_j(t)$ of subnets i and j is measured by the average distance between the two waveforms $d(r_i(t), r_j(t))$, where $d(r_i(t), r_j(t)) = \int_{t=0}^{T} |r_i(t) - r_j(t)| dt$



(a) Example of BGP discreet time-series



(b) Example of corresponding $r(t)$

Figure 4: Example of BGP discreet time-series and corresponding continuous waveform $r(t)$. (1 = BGP announcement, and -1 = BGP withdrawal.)

for $1 \leq i, j \leq n$. The set of similarity measures, $L = \{d(r_i(t), r_j(t))\}$, where $1 \leq i, j \leq n$, is used as inputs for clustering.

We choose the average-linkage hierarchical clustering algorithm [57] which guarantees convergence. The pseudocode of this algorithm is presented as follows:

Let $L = \{d(r_i(t), r_j(t))\}$, where $1 \leq i, j \leq n$. Sort L ascendingly.

While L is not empty,

1. Remove the smallest similarity measure from L.

2. If both subnets are not yet clustered with any subnet, then these two subnets form a new cluster, or

3. Else if one subnet in a pair currently belongs to one existing cluster, and the other subnet is not, both subnets are included in the same existing cluster, or

4. Else if both subnets are in the same cluster, do nothing, or

5. Otherwise, if both subnets are in two different clusters, then merge these two clusters together.

6. Go back to while loop.

2.5.1.2 Clustering Threshold

A clustering threshold $\hat{T}$ is an important parameter that determines the number and compactness of clusters. To measure the compactness of each cluster, we use the Davies-Bouldin index [35].

The Davies-Bouldin index is a relative measurement between intra- and inter-cluster distances. Let $\underline{\mathbf{C}} = \left[\begin{array}{cccc} \mathbf{C_1} & \mathbf{C_2} & ... & \mathbf{C}_q \end{array} \right]$, where $\mathbf{C}_v$ is a set of subnets in v cluster, and q is a number of clusters. Let $\Delta(C_v, C_w)$be the average of $d(r_i(t), r_j(t))$ between clusters C_v and C_w, where $r_i(t) \in C_v$, $r_j(t) \in C_w$, and $1 \leq v, w \leq q$. Also, let $\delta(C_v)$be the average $d(r_i(t), r_j(t))$ inside cluster C_v where $r_i(t), r_j(t) \in C_v$. The Davies-Bouldin index is

$$DB(\mathbf{C}) = \frac{1}{q} \sum_{v=1}^{q} \max_{v \neq w} \left\{ \frac{\delta(C_v) + \delta(C_w)}{\Delta(C_v, C_w)} \right\}$$

The Davies-Bouldin index measures the intra-cluster distance ($\delta(C_v)$, $\delta(C_w)$) against the inter-cluster distance ($\Delta(C_v, C_w)$). A small Davies-Bouldin index identifies compact and well-separated clusters.

We select $\hat{T}$ using the time-series of BGP measurements belonging to1,009 subnets, collected prior to Katrina. The premise is that a range of clustering threshold values obtained from day-to-day network operation could be used for spatial clustering of measurements taken from the disaster. We first cluster the pre-Katrina time-series using different values of $\hat{T}$ and then measure the corresponding Davies-Bouldin indices. Figure 5 shows the Davies-Bouldin indices from clustering the pre-Katrina and the Katrina time-series measurements belonging to one of eight subnet sets presented in Table 1. Among all eight

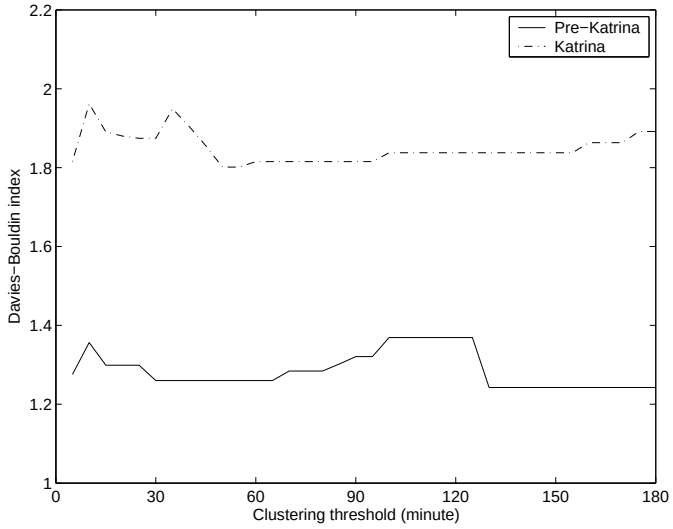


Figure 5: Clustering threshold.

subnet sets, the suggested values of clustering threshold $\hat{T}$ are between approximately 45-90 minutes. The compactness of clusters after applying this suggested value of clustering thresholds is discussed in the following section.

2.5.1.3 Clustering Katrina Time-Series

The average-linkage hierarchical clustering algorithm and the suggested values of $\hat{T}$ are applied to the Katrina time-series measurements belonging to each subnet set. The algorithm running time for these subnet sets ranges from 20 seconds to 3 minutes on a Linux workstation with Pentium 4, 3.00 GHz CPU, 512MB memory, and 64GB hard drive.

Clustering spatially reduces 1,009 time-series to 191 clusters (81% reduction). The reduction percentage of eight subnet sets are presented in Table 2. The large number of subnets demonstrates the large reduction percentage.

How are the three examples of time-series in Figure 2 clustered? They are clustered into three different clusters, where each cluster respectively contains seven, one, and two subnets.

We compute correlation coefficients of time-series in the same cluster. The results show that each cluster contains subnets with correlation coefficients between 0.9959-1.000.

Table 3 presents the example of two subnets from the same cluster and illustrates that subnets from the same cluster have a largely similar pattern of BGP updates. Furthermore,

Table 2: Percentage of subnet reduction.

Subnet set	1	2	3	4	5	6	7	8
Number of subnets	166	27	49	115	26	180	232	214
Percentage of reduction	84	48	67	82	65	77	82	91

Table 3: Examples of time-series patterns and geographic locations belonging to two subnets in the same cluster.

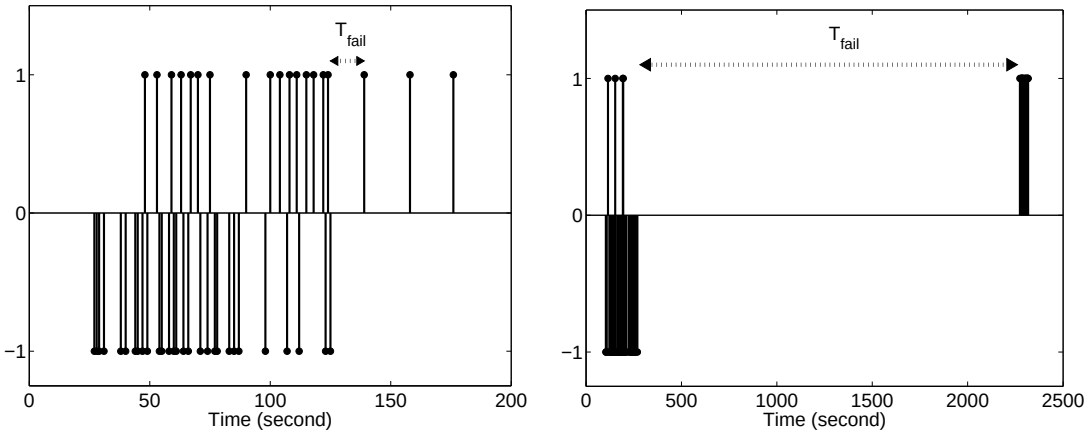
Subnet	AS	Geographic location	Initial time t when $r(t) = -1$	Duration of of $r(t) = -1$
1	X	Hammond, LA	8/30 6:53:42 p.m.	2 hours 53 minutes
			9/3 11:39:09 p.m.	17 minutes
			9/4 12:25:10 a.m.	10 minutes
2	X	Hammond, LA	8/30 6:53:42 p.m.	2 hours 53 minutes
			9/3 11:39:09 p.m.	17 minutes
			9/4 12:10:24 a.m.	10 minutes
			9/4 12:25:10 a.m.	10 minutes

ASes and geographic locations belonging to these subnets are similar. Among 191 clusters, 84.15% of clusters contain subnets from the same AS, and 88.30% of clusters have subnets located in the same geo-location. Hence, AS and geo-location are their defining features for clusters.

2.5.2 Temporal-Spatial Feature Extraction

Because the resulting clusters have correlation coefficients close to one, we can randomly choose one representative subnet per cluster and use these 191 representative subnets to extract temporal features of the time-series.

As described in [46, 87], a burst of multiple BGP withdrawals followed by a BGP announcement is a symptom of subnet unreachability. This symptom has two characteristics. The first characteristic is a burst of BGP withdrawals that is represented by the number of withdrawals sent in a specific time-duration. The second characteristic is the length of an unreachable duration between the last BGP withdrawal of a burst and the first new announcement after a burst. This duration can be used to infer whether a subnet is unreachable. Accordingly, a burst of BGP withdrawals followed by a new announcement and a succeeding unreachable duration form a BGP-burst pattern.



(a) Example of BGP withdrawal burst with the shortly-following BGP announcements.

(b) Example of BGP withdrawal burst with the late-following BGP announcements.

Figure 6: Examples of BGP withdrawal bursts with the subsequent BGP announcements. (1 = BGP announcement, and -1 = BGP withdrawal).

The inference of network statuses from a BGP-burst pattern has been studied for day-to-day network operations [21, 43, 46, 90]. For instance, a BGP-burst pattern with a short succeeding unreachable duration, as shown in Figure 6(a) can be caused by a temporary network-service disruption, e.g., a change of routes or routing policies. In this case, a subnet quickly becomes available after disruption. On the other hand, a BGP-burst pattern with a long succeeding unreachable duration, as shown in Figure 6(b) is usually caused by hardware or software failures [58]. However, network-service disruption caused by a large-scale disaster is a rare event. Thus, it is not clear how many BGP withdrawals should be considered as a burst and how long a succeeding duration is to be considered as unreachable due to a disaster. We formally define two features of the BGP-burst pattern.

Definition: Burst ratio S and unreachable duration T_{fail}

Let v be a time-duration where a burst of BGP withdrawals is located. Let n_v be the number of BGP withdrawals sent by BGP peer routers within a v time-duration, and n_p be the number of BGP peer routers that could reach this subnet prior to the Katrina interval. Note that a BGP peer router can send more than one BGP withdrawal during a disruption.

A burst ratio is defined as $S = \frac{n_v}{n_p}$. S represents the percentage of BGP withdrawals from BGP peer routers. An unreachable duration T_{fail} is defined as the time period between the last BGP withdrawal of a burst in a v-duration and the first new BGP announcement after

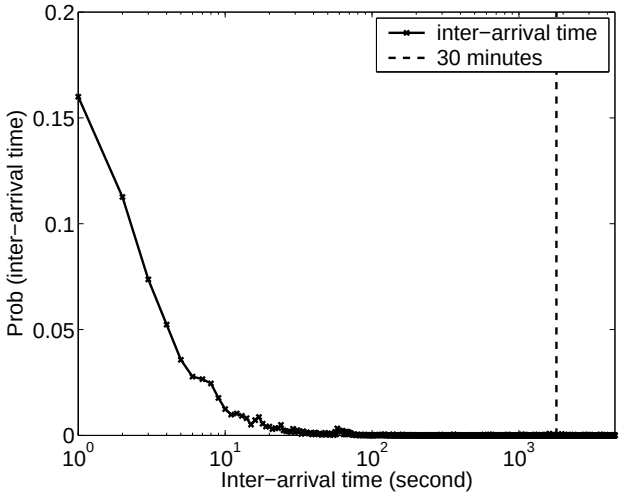


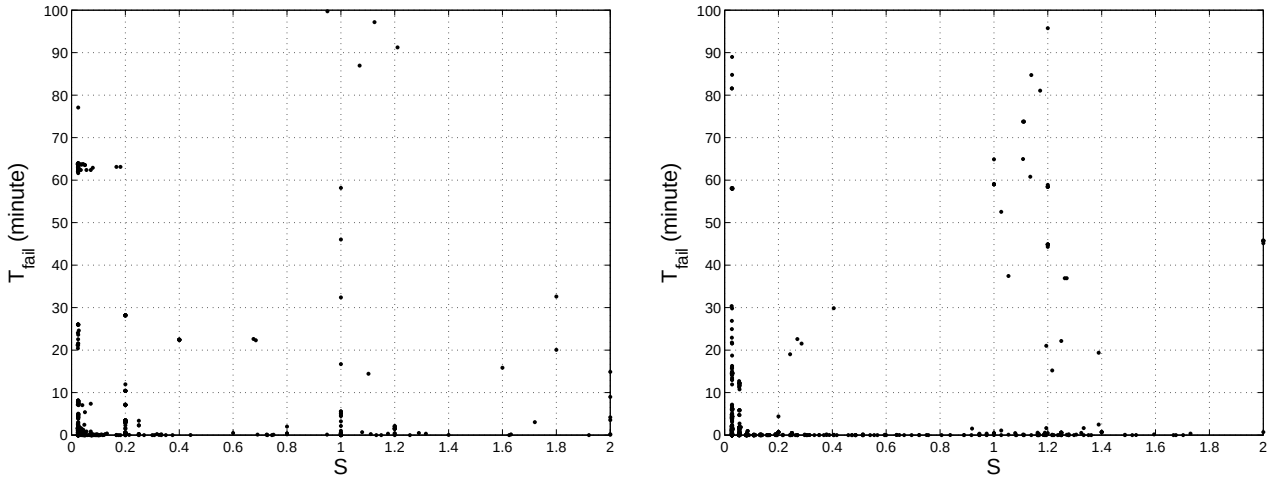
Figure 7: Empirical distribution of BGP withdrawal inter-arrival times.

a burst. Therefore, S is the spatial variable specifying how many BGP peer routers fail to reach a subnet, and T_{fail} is the temporal variable indicating the duration that a subnet remains unreachable.

The parameter v is a time window threshold such that if the inter-arrival time between two BGP withdrawals is larger than v minutes, these two withdrawals are not considered to be in the same burst. It is reported that, in day-to-day network operation, a burst generally lasts for 3 minutes [59] but can be up to 15 minutes [58]. However, there is no prior result on bursts caused by a natural disaster. Figure 7 illustrates the empirical distribution of BGP withdrawal inter-arrival times. Based on this information, we select a sufficiently large $v = 30$ minutes not to partition a burst. Note that a large v as a time window may include more than one burst. This demonstrates a disadvantage of using a fixed-size time window to locate a burst. In Section 3.2, we illustrate how to select the value of v directly from the data.

Statistics of S and T_{fail}

We extract 217 (S, T_{fail}) features from the time-series measurements of 191 representative subnets from the pre-Katrina and the Katrina intervals. Note that some subnets may have more than one (S, T_{fail}) feature while some subnets do not have any (S, T_{fail}) features.



(a) Pre-Katrina interval

(b) Katrina interval

Figure 8: Scatter plots of S and T_{fail}. Scatter plots only show values of T_{fail} up to 100 minutes.

The scatter plots of S and T_{fail} from the pre-Katrina and the Katrina intervals are shown in Figures 8(a) and 8(b) respectively. Trivially, there are more features with large T_{fail} values in the Katrina interval when compared to the pre-Katrina interval.

The conversion from BGP-burst patterns from 191 time-series to two-dimensional (S, T_{fail}) feature space illustrates that temporal feature extraction reduces the temporal dimensionality of the time-series. These (S, T_{fail}) features are used as unlabeled data.

How do we infer which BGP-burst pattern corresponds to subnet unreachability using this unlabeled data? Figure 2 demonstrates that two examples of time-series with unknown statuses (unlabeled) have different BGP-burst patterns from user-input time-series (labeled). Therefore, we cannot directly compare one-by-one BGP-burst patterns. However, we are able to use (S, T_{fail}) features of user inputs to help derive the classifier for subnet unreachability.

2.6 Semi-Supervised Learning

Next, we study what additional information user inputs provide, and how to jointly use such information with unlabeled (S, T_{fail}) features for inference of subnet statuses.

2.6.1 Labeling

As described in Section 2.2, a person can report the unreachability of a subnet, and this report (considered as a user input) can be regarded as a direct observation of a subnet status. However, there is usually a delay between when the subnet become unreachable and when the user report occurs. An open question is: how can the the delayed user inputs be used to identify BGP-burst patterns and obtain labeled (S, T_{fail}) features?

During a user input delays, there can be more than one BGP-burst pattern in a time-series of subnets prior to the report time. For example, there are three BGP-burst patterns before the report time in Figure 2(c). Among our 24 user inputs, 11 of them have more than one BGP-burst pattern before a report time. Therefore, the process to correlate delayed user inputs with BGP-burst patterns may be complex. For simplicity, we assume the person is prompt in reporting a network outage. Thus, we select a BGP-burst pattern immediately preceding a user input. With 24 user inputs, we have 24 (S, T_{fail}) features that are labeled as "1" (outage).

Using the pre-Katrina data, a portion of (S, T_{fail}) features are labeled as "-1" (normal). Figure 9 illustrates the empirical probability distribution of S and T_{fail} from the pre-Katrina interval. Small values, $S < 0.1$ and $T_{fail} < 3$ minutes, occurred with a large probability. This means that only a small (10%) percentage of the BGP peer routers send out BGP withdrawals while the rest of the BGP peer routers can still reach this subnet. A small T_{fail} corresponds to subnets that become reachable quickly after a burst. Hence, subnets with $S < 0.1$ and $T_{fail} < 3$ minutes are considered to be reachable during the pre-Katrina interval. These subnets correspond to the majority of (S,T_{fail}) features in the pre-Katrina interval as shown in Figure 9. We extract 460 (S,T_{fail}) features from the pre-Katrina time-series and label these features as "-1" (normal). 300 normal labels are selected for training and the other 160 are for validation. Figure 10 presents (S, T_{fail}) features belonging to normal and outage labels[4].

In summary, we have 217 unlabeled features, $\{(S_i, T_{fail_i})\}_{i=1}^{217}$, 24 features labeled as

[4]Fewer normal labels can be used since the dimension of (S, T_{fail}) feature space is two, and normal labels are well-clustered.

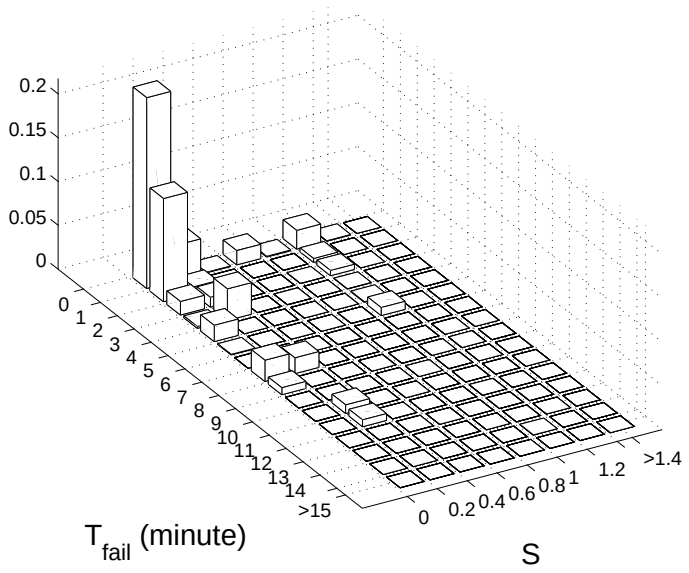


Figure 9: Empirical probability distribution of S and T_{fail} from pre-Katrina interval.

outage $\{(S_i, T_{fail_i}), 1\}_{i=1}^{24}$, 300 features labeled as normal, $\{(S_i, T_{fail_i}), -1\}_{i=1}^{300}$.

2.6.2 Learning

Labeled and unlabeled data are jointly used next in semi-supervised learning. The key idea of semi-supervised learning is that a small number of labeled data helps classify unlabeled data [22]. There are three major algorithms in semi-supervised learning: generative models, transductive support vector machine, and graph-based methods. Because generative models and graph-based methods require probabilistic models, we select the transductive support vector machine (TSVM) from Joachims [55]. This TSVM is applied to both labeled and unlabeled data.

The goal is to obtain a (S, T_{fail}) classifier in order to determine the statuses of subnets (normal/outage). To avoid over-fitting, we simply choose to apply TSVM to S and to T_{fail} separately. The result is that two one-dimensional linear classifiers (one for S and the other for T_{fail}) are combined together as one two-dimensional classifier; this classifier is used to infer statuses of all subnets.

Let (x_i, y_i) be labeled data and x_j^* be unlabeled data where x_i or x_j^* is a generic variable

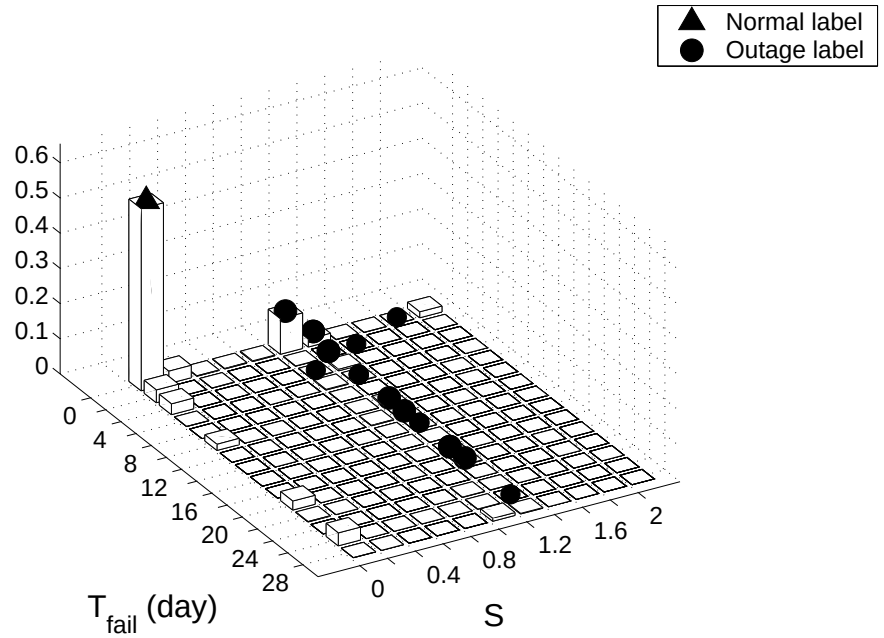


Figure 10: Empirical probability distribution of S and T_{fail} from Katrina interval, along with S and T_{fail} of normal and outage labels.

that corresponds to S or to T_{fail}, $1 \leq i \leq k$, and $1 \leq j \leq m$. y_i is a class label of x_i that is in Section 2.6.1. Let y_j^* be an unknown class label of x_j^* to be determined by a classifier. y_i, $y_j^* \in \{1, -1\}$. Let ξ_i be a slack variable for x_i and ξ_j^* be a slack variable for x_j^*, used in the support vector machine. The use of slack variables allows penalties for misclassified samples (see [20] for details).

Let w be the weight and b be the bias of a linear classifier to obtain from minimizing

$$\frac{||w||^2}{2} + C\sum_{i=1}^{k}\xi_i + C_-^*\sum_{j:y_j^*=-1}\xi_j^* + C_+^*\sum_{j:y_j^*=+1}\xi_j^* \tag{2}$$

subject to

$$y_i(w\cdot x_i + b) \geq 1 - \xi_i, \tag{3}$$

$$y_j^*(w\cdot x_j^* + b) \geq 1 - \xi_j^*, \tag{4}$$

$$\xi_i \geq 0, \tag{5}$$

$$\xi_j^* \geq 0, \tag{6}$$

where $\frac{2}{||w||}$ is the margin width of a classifier, and $\sum_{i=1}^{k} \xi_i$ and $\sum_{j=1}^{m} \xi_j^*$ are bounds of classification error. C, C_-^*, and C_+^* are trade-off parameters between margin width and classification error (see [55] for details).

The obtained outputs are w and b. $\frac{-b}{w}$ is the threshold of S or T_{fail}, used to determine class labels $\{y_j^*\}_{j=1}^{m}$.

2.6.3 Experimental Setting and Results

Since unlabeled data is abundant, we separate the unlabeled features into 10 different subsets. Ten different classifiers are trained, and each training uses (a) one subset of 21 (or 22) unlabeled features, (b) all 24 outage-labeled features, and (c) one subset of 30 randomly chosen normal-labeled features.

Parameters used in the TSVM algorithm are initialized such that t $C = 10^{-1}$, $C^* = 10^{-1}$, and $num_+ = 0.5k'$, where k' is the number of unlabeled features in each training set. These parameters affect the convergence (see [55] for details on a choice of parameters).

To choose C, C^*, and num_+, we experiment with different values of C, C^*, and num_+. The different values of num_+ do not change the resulting classifier. However, we find that $C \geq 10^{-1}$ and $C^* \geq 10^{-4}$ are the appropriate ranges for S classifier. If smaller values of C or C^* are used, then the classifier assigns all features to only one class. For T_{fail} classifier, $C \geq 10^{-4}$ and $C^* \geq 10^{-3}$.

The resulting classifiers are nearly the same when different values in the range are chosen for parameters; however, smaller values converge faster. To keep the balance between unlabeled or labeled data at the beginning of learning, we choose the same value for both C and C^*. Thus, $C = 10^{-1}$ and $C^* = 10^{-1}$ are the best choices for common parameters for training both S and T_{fail} classifiers.

Let S^* and T_{fail}^* be the thresholds that if any subnet has a feature $S > S^*$ and $T_{fail} > T_{fail}^*$, this subnet is inferred as unreachable as a result of Hurricane Katrina. Ten thresholds of S resulting from training 10 different classifiers are averaged to yield S^*. We follow the same process to find the value of T_{fail}^*. The semi-supervised learning results in $S^* = 0.6153$ and $T_{fail}^* = 1$ hour 38 minutes.

Using these thresholds S^* and T^*_{fail}, we are able to infer the subnet with the time-series in Figure 2(a) as normal and the subnet with the time-series in Figure 2(b) as unreachable.

2.6.4 Validation

Usually, little information is available for the underlying statuses of subnets during a disaster. A key challenge when inferring unreachable subnets caused by a natural disaster is how to validate the inference results. We consider two aspects of the validation process: (a) the consistency of the learning methods on training and testing data, and (b) the correctness of the inference of the subnet statuses.

First, we validate the "consistency" of unsupervised learning and semi-supervised learning to this application. We use a test set that consists of the remaining 13 user inputs and 160 normal labels that are not used in semi-supervised learning to validate the learned S^* and T^*_{fail}. The result shows that the features belonging to these 13 user inputs have $S > S^*$ and $T_{fail} > T^*_{fail}$ and thus are inferred as unreachable. The inferred unreachable statuses of these user inputs are consistent with the reports that these subnets were unreachable. Furthermore, the testing normal labels are also correctly classified. Thus, the values of S^* and T^*_{fail} perform "consistently" for both training and testing data.

Second, we validate the "correctness" of our inference by examining the connectivity of the subnets one-by-one. This can be done using two methods. The first is to examine BGP routing tables. A BGP routing table is a table where a BGP peer router keeps a list of reachable subnets. The second method is to explicitly examine the contents of every BGP update message and keep track of which BGP peer routers stay connected with a subnet.

The state of a subnet can be obtained by checking the existence of such subnet in a BGP routing table at a specific time. However, RouteViews provides the BGP routing tables at two-hour intervals. This time-scale is not refined enough for checking unreachability that lasted shorter than two hours.

Besides examining the BGP routing tables, the contents of the BGP updates can be explicitly examined in the following manner. Let $P^t_i = \{p^j_i\}^{|P|}_{j=1}$ be a set of BGP peer routers, where p^j_i is a BGP peer router j that connects to a subnet i at time t, $|P|$ is the

size of P_i^t, and $t \in [0, T]$. If a BGP peer router p_i^j sends a BGP withdrawal regarding to a subnet i at time t, then $P_i^t = P_i^t - p_i^j$. If a BGP peer router p_i^j sends a BGP announcement for a subnet i at time t, then $P_i^t = P_i^t \cup p_i^j$. When P_i^t is null, a subnet i is unreachable at time t.

The validation shows that subnets that are inferred as unreachable do not exist in the BGP routing tables during the network-service disruptions. In addition, by examining the contents of the BGP updates belonging to an inferred unreachable subnet i, we find that its P_i^t is null when t is the time that the last BGP withdrawal of a burst arrives.

Note that while explicit examination of the contents of the BGP updates can be used to directly indicate the reachable status of a subnet, this requires us to check every BGP update for every subnet. If machine learning were not applied, this means we would have to examine all 173,947 BGP updates that belong to 1,009 subnets by brute force. In contrast, using the inference approaches in this work, only 217 BGP-burst patterns from 191 representative subnets are considered. This clearly demonstrates how learning approaches significantly reduce the cost of inferring large-scale network-service disruptions.

2.7 Inferring Network Disruption

The resulting thresholds are then used to infer network disruption caused by Katrina for the other 985 subnets.

2.7.1 Inferred Subnet Statuses

The decision boundaries of $S = S^*$ and $T_{fail} = T_{fail}^*$ partition the feature space into the following two main regions as shown in Figure 11:

- Outage region where $S > S^*$ and $T_{fail} > T_{fail}^*$ (upper right region in Figure 11). This region contains S and T_{fail} belonging to the inferred unreachable subnets.
- Normal region where either $S \leq S^*$ or $T_{fail} \leq T_{fail}^*$. This region contains S and T_{fail} belonging to the inferred reachable subnets.

The results from the normal region can be further studied to gain more knowledge on network responses to Katrina. In the normal region, there are two sub-regions marked as

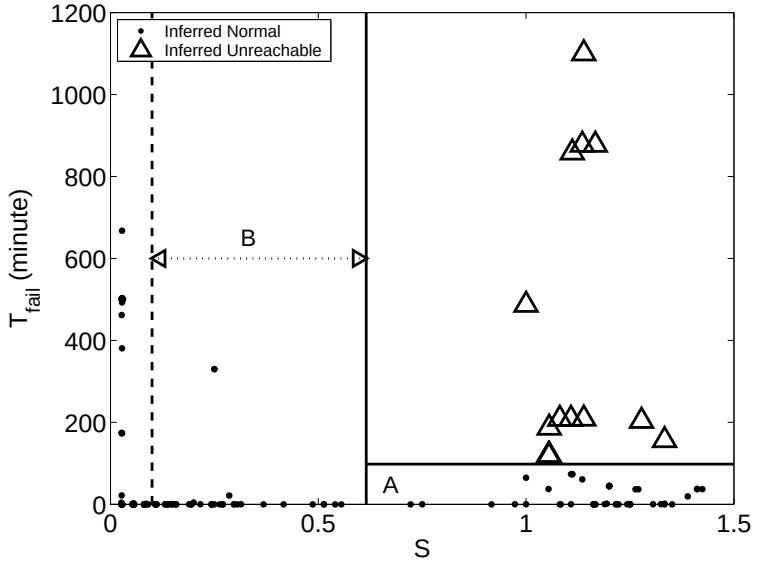


Figure 11: Scatter plot of inferred S and T_{fail}. (Solid vertical line: $S = S^*$, dashed vertical line: $S = 0.1$, and horizontal line: $T_{fail} = T^*_{fail}$.) Plot only shows values of T_{fail} up to 1200 minutes.

A and B in Figure 11. These two sub-regions contain the features that are inferred as normal but show the interesting characteristics of network resilience and responses due to Hurricane Katrina. Note that, we exclude the region where $S < 0.1$ and $T_{fail} < 3$ minutes because these inferred normal subnets did not respond differently from the pre-Katrina normal subnets.

Region A is located where $S > S^*$, $T \leq T^*_{fail}$, and $T >$3 minutes. The features in this region correspond to the subnets where after Katrina, multiple BGP peer routers responded with bursty BGP withdrawals. However, these subnets only experienced brief T_{fail} and resumed reachability soon after. The empirical probability distribution of T_{fail} presented in Figure 12 shows that there are significantly more T_{fail}'s with moderate values, 35-55 minutes, in the Katrina interval while T_{fail} of such values were scarce during the pre-Katrina interval. This shows that Katrina caused the network to respond differently from day-to-day network operation.

Region B is located where $0.1 \leq S \leq S^*$. The features in this region correspond to the subnets where only a small number of the BGP peer routers responded to Katrina. Examining the statistics of S, we find that there are more S with values between $[0.1, 0.5]$ in the Katrina interval than the pre-Katrina interval. We also study the corresponding

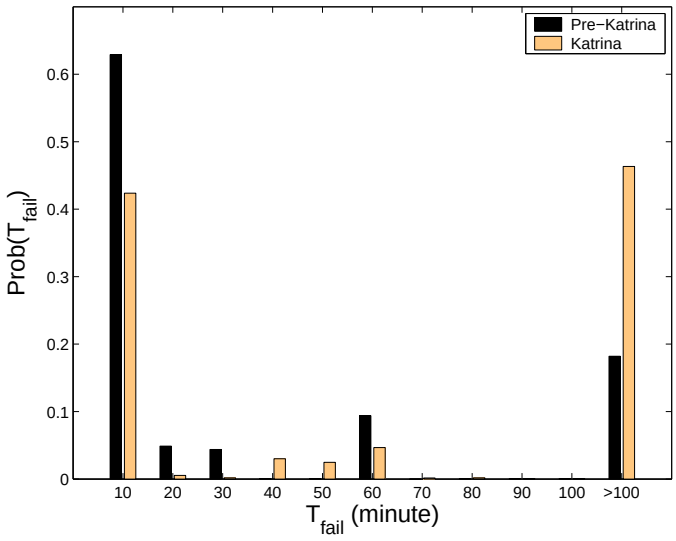


Figure 12: Empirical probability distribution of T_{fail} from the pre-Katrina and the Katrina intervals.

Table 4: Percentage of subnets in four regions.

Region	Percentage of subnets
Normal	75
Outage	25
A	12
B	30

subnets and find that these subnets maintained their reachability statuses, but there may have been small parts of the Internet that were affected by Katrina.

We quantify the percentage of subnets in these four regions as shown in Table 4. The results show that 25% of subnets are inferred as outages. There are 42% of subnets corresponding to both regions A and B. The number of subnets that maintained reachability or responded with brief disruption duration provide signs of network resilience under stress.

2.7.2 Spatial-Temporal Damage Maps

We obtain the spatial damage map presented in Figure 13. The spatial map shows network-service disruptions at different levels based on the average unreachability durations of the inferred unreachable subnets in each geographic location. The worst network-service disruptions occurred near the coast of Louisiana. Nonetheless, our results show that not all

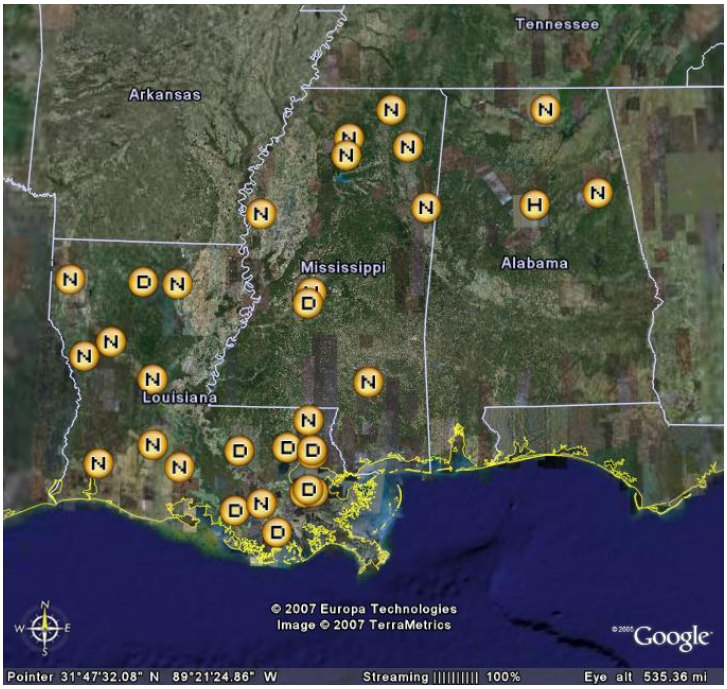


Figure 13: Degree of impact of network-service disruptions. (N): $T_{fail} < T^*_{fail}$, (H): $T^*_{fail} < T_{fail} < 24$ hours, and (D): $T_{fail} \geq 24$ hours.

subnets in the entire disaster area suffered from network-service disruptions. This suggests that available network resources in the area could have been utilized if this information was shared among disaster responders.

We use T_{fail} to identify initial times and durations of network disruptions. The temporal map in Table 5 and Figure 14(a) show that 49.21% of network disruption occurred after the landfall of Katrina while only 5.12% occurred on August 28, 2005 (the mandatory evacuation day). Substantial network disruption (45.67%) also occurred during the six hours before the landfall of the hurricane.

For the unreachable subnets that occurred during six hours before the landfall, our collaborative ISP expected that these subnets likely were intentionally withdrawn by network operators rather than disrupted by Katrina. On the other hand, there was the report of network connectivity loss because of the lack of power supply on August 29, 2005 at 3:00 a.m. [84]. Thus, it is inconclusive what exactly caused subnet unreachability during the six hours prior to the landfall. In Chapter 5, we seek for the exact causes why some subnets become unreachable before the landfall of the hurricane.

Table 5: Percentage of unreachable subnets with different initial times.

Initial time	Before 8/29/05 12:00 a.m.	Between 8/29/05 12:00-6:00 a.m.	Between 8/29/05 6:00 a.m.-11:59 p.m.	Between 8/30/05-8/31/05	Between 9/1/05-9/4/05
Percentage of unreachable subnets	5.12	45.67	37.01	6.69	5.51

Table 6: Percentage of unreachable subnets with different unreachability durations.

Unreachability duration	Less than 1 day	1-3 days	3-7 days	1-2 weeks	2-4 weeks	Longer than 4 weeks
Percentage of unreachable subnets	7.09	6.30	2.76	3.54	7.48	72.83

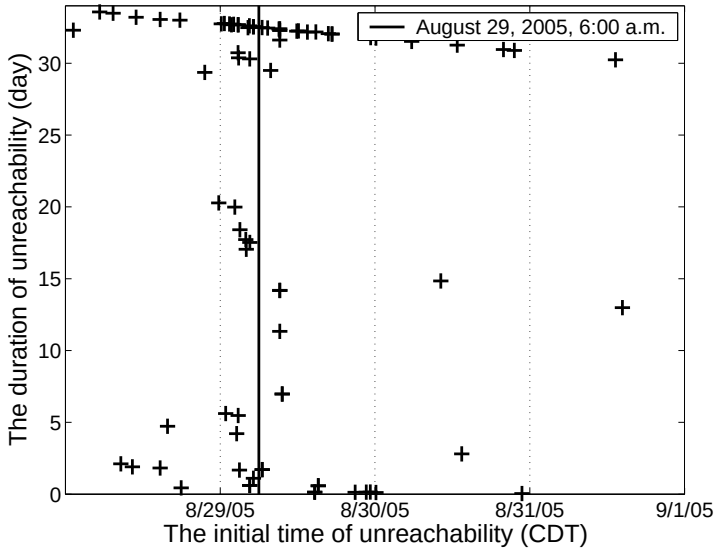
Figure 14(b) shows the initial times and the durations of unreachable subnets located in different cities that were critically damaged by Katrina. These cities are near the coast of Louisiana. The results show that unreachable subnets located in the same city did not necessarily occur at the same time or last with the same approximate duration.

The percentage of unreachable subnets with different unreachable durations is shown in Table 6. Approximately 73% of subnet unreachability lasted longer than four weeks. This illustrates the severity of extreme damage to networks caused by Hurricane Katrina.

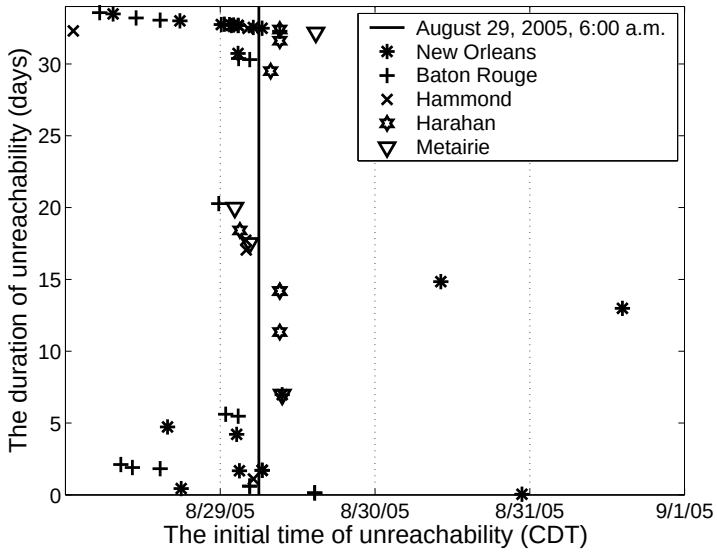
2.8 Remarks from ISP

From the discussion with our collaborating ISP, networks in the critically damaged cities were certainly disrupted. However, network disruptions inside the same city randomly occurred and did not necessarily become unreachable close in time.

The causes of some networks were disrupted were hardware failure and power outages. The occurrences of network disruptions depended locally on where users placed their network equipment or their choices of power backup resources. Our collaborating ISP maintained their network connectivity at all time despite the location of its office in New Orleans because their equipment was above the flood level.



(a) All inferred unreachable subnets. ("+" = an inferred unreachability.)



(b) Inferred unreachable subnets in different coastal cities of Louisiana.

Figure 14: Initial times and durations of inferred unreachable subnets.

Our collaborating ISP reported that the number of disrupted networks did not incrementally decrease during the recovery process. Instead, network disruptions gradually occurred. Some network disruptions were not caused by Katrina during the landfall but by the recovery process. There were networks that maintained connectivity after Katrina; however, the networks were disrupted later because the recovery workers unavoidedly or accidentally damaged transmission cables.

From our view, unreachability durations of disrupted networks were independent from initial times or geo-locations. However, the ISP informed us that the unreachability durations in fact depended on the the service level agreement between the ISP and customers. In general, ISPs serve customers ranked by priority. For example, networks that belonged to emergency services, e.g., hospitals, polices, communication services, had the highest priority. Other priorities included number of affected customers, accessibility to damaged locations, etc.

2.9 Summary

This work introduces machine learning to a new networking application, that is, an inference of large-scale network-service disruptions caused by a natural disaster. The uniqueness of this network application includes (a) large-scale network measurements that exhibit spatially- and temporally-correlated bursty behavior triggered by a disaster event, and (b) a small number of real user inputs from disaster responders.

We find that machine learning has played a vital role in understanding large-scale and complex network measurements that can be divided into two aspects. First, clustering reduces the spatial dimension of network measurements by 81%, and temporal feature extraction reduces the temporal dimension down to two informative features. Second is that semi-supervised learning uses the large number of network measurements and the small number of user inputs to derive a classifier to infer the Katrina network-service disruptions.

The results show that 25% of subnets are inferred as unreachable. A large fraction (42%) of subnets are found to be either maintained or to resume reachability briefly after Katrina. There are 12% of subnets that experienced a moderate unreachability durations

between 35-55 minutes.

The use of remote network monitoring demonstrates the feasibility and the advantage of inferring network disruption when the disaster area is physically inaccessible. Moreover, because ISPs do not generally disclose information on unreachability of their networks, this application can be used by anyone to infer network disruptions.

Limitations and open issues:

There are limitations to our study.

1. BGP updates may not provide a complete view of subnet connectivity. It is possible that a subnet could maintain connectivity even though a BGP peer router of Route-Views cannot reach this subnet. Thus, local measurements can be included. The approach presented in this work can be extended with more local measurements.

2. Bursts of BGP updates are currently located using the fixed threshold of 30 minutes based on the prior work [58, 59]. In the following part of this research, we improve the burst detection by using a threshold learned from the BGP update directly.

3. Although Whois [85] is a commonly-used database, it is also considered to be out of date. In the subsequent part of this research, we address the accuracy of Whois database and also introduce another database to use for geo-locations of subnets.

The uniqueness of this inference problem has also challenged the existing machine learning approaches. The first challenge is in working with large-scale of data. Although the simple hierarchical clustering algorithm is used, this clustering algorithm is not scalable to large data sets because the running time is $O(N^2)$, where N is the number of measurements. That is why in Section 2.5.1.3, we partition subnets into subnet sets, and apply the clustering algorithm to each set separately. This has helped scale down the number of clustering inputs and thus has reduced the computation cost. The other challenge is that semi-supervised learning has not been used for network data, and that user inputs are not in a ready form of labels.

CHAPTER III

UNDERSTANDING TEMPORAL AND LOGICAL CHARACTERISTICS OF NETWORK DISRUPTIONS

We continue our study of network-service disruptions caused by large-scale disturbances using another case study, i.e., Hurricane Ike in 2008. After Hurricane Ike in 2008, 168 subnets in Texas were reportedly disrupted [18]. However, there was little study on how networks-disruptions responded to the disturbances.

In this Chapter 3, we focus on studying network-service disruptions caused by large-scale disturbances with respect to temporal and logical network. We describe the real data on networks and organizations and analyze network behaviors due to disruptions in logical networks and organizations. Compared to Hurricane Katrina, we are able to select larger data sets of subnets in the disaster area.

Chapter 3 is organized as follows. Sections 3.1 and 3.2 provide background and identification of unreachable subnets. Section 3.3 presents temporal independence of subnet unreachability and hypothesis test. Section 3.4 studies temporal and logical dependence of unreachable subnets, whereas Section 3.5 illustrates the hierarchy of logical network. Section 3.6 compares subnet unreachability under normal operations and those caused by Hurricane Ike. Section 3.7 reports societal impact, and Section 3.8 summarizes Chapter 3.

3.1 Background: Hurricane Ike

Hurricane Ike is the strongest hurricane in 2008. It made landfall at Galveston, Texas on 7:10 a.m. September 13, 2008[1] [12]. Once Ike passed the Gulf Coast, the hurricane caused strong wind, heavy rain, and floods in Texas[12, 48]. Later, Ike weakened to a tropical storm at 1:00 p.m. September 13 and exited the state of Texas by 2:00 a.m. September 14.

The high level of impact from Hurricane Ike has been reported for the following key

[1]The reported time here is Central Daylight Time (CDT), i.e., the local time in Texas.

components of our nation's critical infrastructure:

- Communication networks: 168 networks in Texas experienced service disruptions at the Internet scale [18].
- Organizations: In preparation for Hurricane Ike, the mandatory evacuation order was announced on September 10–three days prior to the landfall. Many organizations, e.g., hospital, responded to the evacuation order [80].
- Power networks: As reported in [25], Ike caused power failures in the disaster area. More than 2 millions of customers experienced power failures.

3.2 Heterogeneous Data

We begin by identifying network-service disruptions using: (a) subnet unreachability, in order to characterized network disruptions, and (b) organization identities of the networks, in order to understand how different social entities responded to the disaster.

First, subnets in Texas are selected using GeoIP City database from Maxmind [49]. Maxmind provides geo-locations of IP addresses in terms of latitude and longitude, and reports that approximately 79% of the geo-locations are "correctly resolved within 25 miles from a true location" [49]. Maxmind updates GeoIP database monthly. Since GeoIP database is updated more frequently, and the accuracy is known, in this part of the research, we replace Whois database [85] with GeoIP as our geo-location data source.

We randomly sample two IP addresses for each prefix from GeoIP. If geo-locations corresponding to two sampled IP addresses are located in the state of Texas, then this subnet is considered to be in the state. As a result, 3,601 subnets in Texas are allocated.

3.2.1 BGP Measurements

We collect BGP measurements to identify unreachable subnets. The BGP update messages belonging to 3,601 subnets are collected from Route Views [74] during Hurricane Ike, i.e., between September 10-20 (from the announcement of evacuation to one week after a landfall of Hurricane Ike). We also collect BGP updates prior to Ike, i.e., between August 1-September 9, as a baseline for normal operations. The collected update messages result in

3,601-dimensional time series of 51 days and amount to 96 megabytes of data.

Identification of unreachable subnets

Unreachable subnets are identified using the collected BGP measurements. As bursts of BGP withdrawals and new announcements appear when a subnet becomes unreachable [72], our first step is to identify bursts of BGP updates. Secondly, we examine, within each burst, when all the prior logical connectivities to the subnet are withdrawn from BGP peer routers.

Burst detection: We detect bursts of multiple BGP update messages since these BGP update bursts represent symptom of subnet unreachability. Instead of analyzing all BGP updates in the time series, we only need to focus on BGP updates inside the bursts to determine the statuses of subnets. Hence, burst allocation also reduces temporal dimensionality.

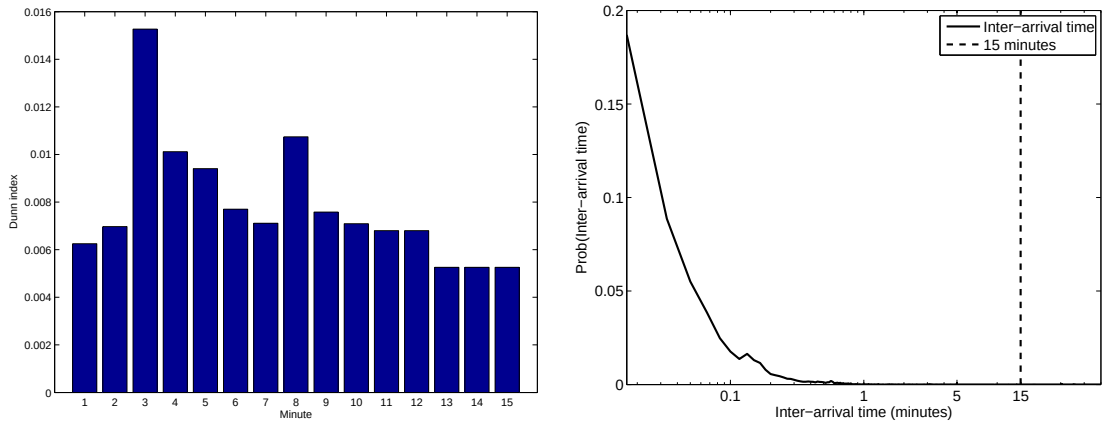
The burst detection is improved from the Katrina study in Section 2.5.2. In Section 2.5.2, bursts of BGP updates are detected using a fixed threshold from prior works [58, 59]. Here, bursts are detected using threshold directly learned from the BGP update messages belonging to subnets.

To detect bursts of BGP updates, we cluster BGP updates using a threshold τ such that if any two BGP updates arrive within τ minutes, these two BGP updates are considered to be in the same burst. To select the threshold τ, we perform the following:

- Using different values of τ, we cluster BGP updates so that two BGP updates with their inter-arrival time less than τ are included in the same burst. Labovitz *et. al.* reported that BGP update bursts last between 3-15 minutes [58, 59]; thus, the values of τ are varied between this range.

- Compute Dunn index [40] to measure compactness of BGP update bursts. The Dunn index $D = \frac{d_{min}}{d_{max}}$, where d_{min} and d_{max} are, respectively, the minimum difference of inter-arrival times among different bursts, and the maximum difference of inter-arrival times inside the bursts. The larger Dunn index, the more compact BGP update bursts are.

- Select the minimum value of τ that provides a consistent Dunn index. The minimum



(a) Dunn index for different values of τ.

(b) Inter-arrival times

Figure 15: Dunn index and inter-arrival times of BGP update messages.

value is preferred because the larger value of τ, the more BGP updates outside the underlying bursts would be incorrectly included in the detected bursts. We look for the consistent allocation of BGP update bursts despite the increasing values of τ. Figure 15(a) shows the values of Dunn index for different values of τ. The Dunn index decreases from 0.0068 at $\tau = 12$ minutes to 0.0053 at $\tau = 13$ minutes, and the Dunn index becomes largely stable after 13 minutes. Hence, since the minimum τ that results in consistent Dunn index is preferred, we select the value of τ to be 13 minutes.

Thus, in this work, if any two BGP updates arrive within 13 minutes, we consider them to be in the same BGP update burst.

Figure 15(b) shows the inter-arrival times of BGP update messages, and illustrates that probability of inter-arrival time between two BGP updates larger than 13 minutes is very small. Hence, the clustering threshold of 13 minutes is not likely to separate one underlying BGP update burst into two detected bursts.

Unreachabilitiy inference and validation: To identify whether a burst is an unreachable burst, and a subnet becomes unreachable, we collect BGP peer routers that have connectivities to a subnet prior to the burst. We then track which BGP peer router sends BGP withdrawal. If all BGP peer routers with prior connectivitys send BGP withdrawals, this

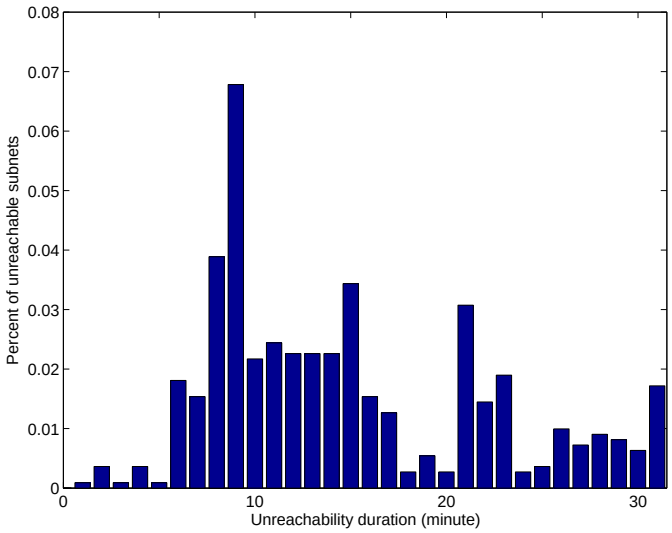


Figure 16: Empirical distribution of unreachability durations between 8/1/08 - 9/9/08.

subnet is identified as unreachable.

Note that BGP withdrawals can be lost. From the observation of this data set, the majority of BGP peer routers send withdrawals, and 0.23% of bursts have numbers of missing withdrawals between 1-3. In this case, to determine whether a burst corresponds to an unreachable burst, we use an unreachability duration. The unreachable duration is the time duration between the instance of unreachability and the next BGP announcement. Figure 16 presents the unreachability durations collected from normal operations, i.e., pre-Ike interval. The unreachability durations during the pre-Ike interval occurred at a distinguishable level starting at 6.02 minutes. (The remaining 55.15% (not shown in Figure 16) of unreachable subnets from the pre-Ike interval had their unreachability durations longer than 30 minutes.) Thus, if a subnet has a burst with between 1-3 missing BGP withdrawals and an unreachability duration longer than 6 minutes, we identify this subnet as unreachable.

The time instance when the last BGP peer router withdraws the connectivity is the time when the subnet becomes unreachable. The unreachable duration is the period between this instance and the next BGP new-route announcement.

Figure 17 presents the number of unreachable subnets between August 1-September 20. Note that during the initially chosen Ike interval (September 10-September 20), 282 unreachable subnets intensely occurred between 12:00 a.m. September 12 and 12:00 p.m.

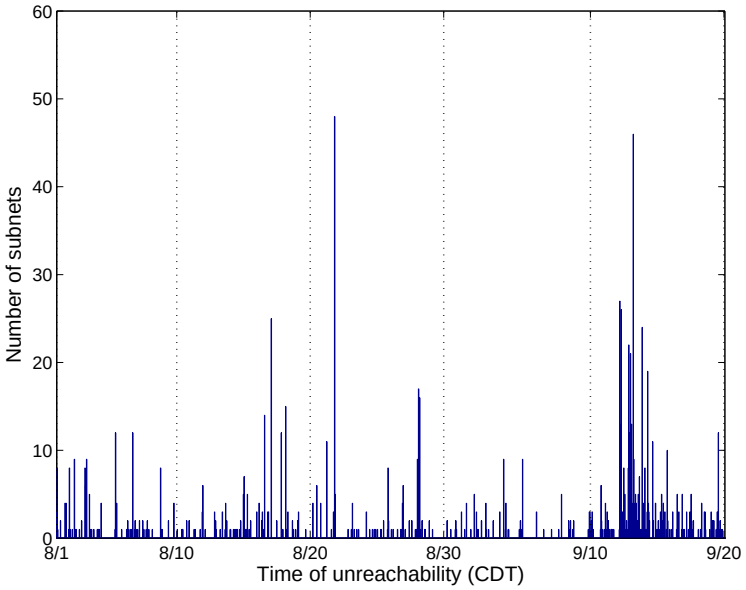


Figure 17: Number of unreachable subnets between 8/1/08 - 9/20/08.

September 14. After 12:00 p.m. September 14, subnet unreachability randomly occurred. Hence, these 282 unreachable subnets in the 2.5-day duration are considered as related to Hurricane Ike. We shall validate this assumption in Section 3.6 by comparing this with subnet unreachability during normal operations or pre-Ike interval.

3.2.2 Organizations, ASes, and ISPs

Organization identities of the networks provide understanding how social entities responded to the disaster. In network operations, organizations own subnets, form ASes, and are served by ISPs. We categorize ISPs into tier-1, major U.S., and local ISPs. Organizations who own unreachable subnets are identified from Whois database [85] whereas ASes and ISPs are obtained from ASPATHs. The 282 unreachable subnets are found to correspond to 107 organizations, 84 ASes, and 36 ISPs. As an example of subnet, prefix 66.140.22.0/24 is owned by the Texas Medical Center, assigned to AS30107, and has AT&T as its ISP.

Combining the above variables, we have a four-tuple of prefix (subnet address), organization, AS, and ISP as a logical representation of a subnet. In addition, the time instance (initial time) when a subnet first becomes unreachable and unreachability duration contain particularly valuable information relating to network disruption caused by the hurricane.

Hence, the union of (1) prefixes, (2) organizations, (3) ASes, (4) ISPs, (5) initial times, and (6) unreachability durations form the heterogenous network data for this study.

3.3 Temporal Independence

We then study temporal dependence of subnet unreachability, i.e., how subnets became unreachable either independently or in groups. First, we analyze whether subnets became unreachable (statistically) dependent in time.

3.3.1 Test of Independence

From the 282 initial time epochs of unreachable subnets, we observe that subnets became unreachable in two time scales. One time scale is large, i.e., subnets became unreachable tens of minutes or hours apart from the other subnets. The other is small time scale, i.e., subnets became unreachable in groups within less than three minutes.

We group unreachable subnets that became unreachable in a small time scale of less than three minutes. Each group is represented by a unique initial time when the last subnet in the group became unreachable. For example, prefixes 209.33.57.0/24 and 64.72.54.0/23 became unreachable at 3:04 a.m. and 3:05 a.m. September 13. Thus, these two prefixes are in the same group with a unique initial time 3:05 a.m. The 282 initial time epochs result in 109 groups, each with a unique initial time as displayed at the bottom of Figure 18.

The unique initial times occur in the following four intervals with different intensities:

1. Gradually from 9/12, 4:00 a.m. to 9/12, 9:00 p.m.,
2. Intensely between 9/12, 9:00 p.m. and 9/13, 2:00 a.m.,
3. Decreasingly between 9/13, 2:00 a.m. and 9/13, 3:00 p.m.,
4. Sparsely between 9/13, 3:00 p.m. and 9/14, 9:00 a.m.

The sample averages of the inter-unreachable times for these four intervals are estimated as 34, 9, 26, and 72 minutes respectively. As these initial time epochs are sufficiently apart, we pose a hypothesis (H_0) that the unique initial times occur independently in the four disjoint intervals according to a non-uniform Poisson point process, and uniformly inside

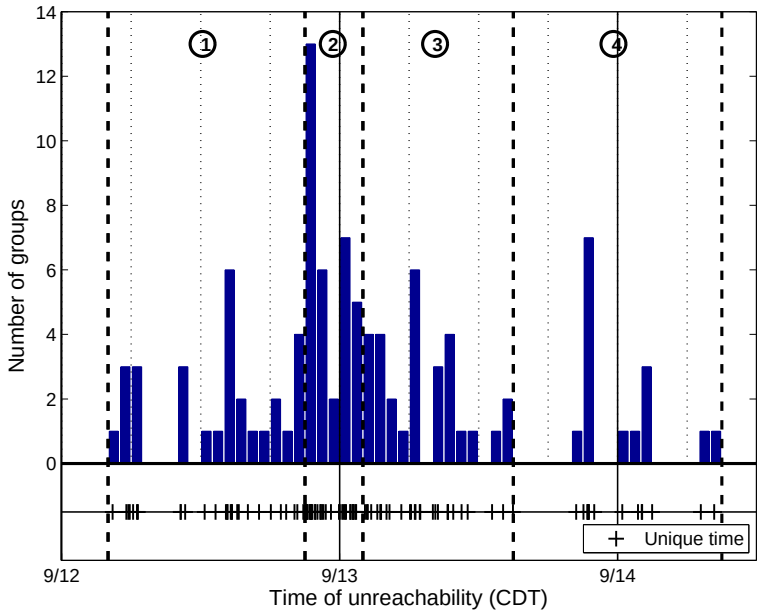


Figure 18: Unreachable subnet groups with unique initial times between 9/12/08, 12:00 a.m. and 9/14/08, 12:00 p.m. Intervals 1-4 have different average inter-unreachable times.

each interval. We construct a chi-square statistic within each interval by comparing the initial time epochs with their sample mean. The independent chi-square statistics are summed up from the four intervals, resulting in another chi-square statistic. Pearson's chi-square test can then be applied to test the hypothesis H_0. H_0, if accepted, shows that these subnets became unreachable independently in time.

The 109 unique initial times are used to test H_0, and the details of the hypothesis test are given below.

1. Separate the unique initial times into four disjoint intervals: T_1= [*9/12, 4:00 a.m.-9:00 p.m.*], T_2=[*9/12, 9:00 p.m.-9/13, 2:00 a.m.*], T_3=[*9/13, 2:00 a.m.-3:00 p.m.*], and T_4=[*9/13, 3:00 p.m.-9/14, 9:00 a.m.*].

2. Assume the unique initial times in T_k are from a Poisson point process with arrival rate λ_k, $1 \leq k \leq 4$. Compute the estimated arrival rate $\hat{\lambda}_k$ that is the number of unique initial time counts in T_k (N_k) divided by the duration T_k. Table 7 shows the values of N_k, T_k, and $\hat{\lambda}_k$.

3. Calculate the chi-square statistic $\chi^2_{d_k}$ with degree of freedom d_k for each interval T_k.

Table 7: Parameters of disjoint intervals. (unit of T_k and T_k/N_k are minute.)

k	T_k	N_k	T_k	T_k/N_k	$\hat{\lambda}_k$	$\chi^2_{d_k}$
1	9/12, 4:00 a.m. - 9/12, 9:00 p.m.	30	1020	34	0.0294	0.3379
2	9/12, 9:00 p.m. - 9/13, 2:00 a.m.	34	300	9	0.1133	0.2275
3	9/13, 2:00 a.m. - 9/13, 3:00 p.m.	30	780	26	0.0385	1.2068
4	9/13, 3:00 p.m. - 9/14, 9:00 a.m.	15	1080	72	0.0139	2.1412

3.1 Divide each interval T_k into uniform sub-intervals.

3.2 Count the number of sub-intervals with zero, one, and greater than one unique initial times. Let c_j be the number of unique initial times in each sub-interval, for $1 \leq j \leq 3$; hence, values of c_1, c_2, and c_3 are 0, 1, and > 1 respectively.

3.3 Assign the measurements in 3.2 to $O_{(k,c_j)}$ that is the observed number of sub-intervals in T_k with c_j number of unique initial times. Values of $O_{(k,c_j)}$ are presented in Table 8.

3.4 Use the estimated arrival rate $\hat{\lambda}_k$ to compute $E_{(k,c_j)}$ that is the expected number of sub-intervals in T_k with c_j number of unique initial times. Let v be the total number of sub-intervals in T_k. Thus, $E_{(k,c_1)} = ve^{-\hat{\lambda}_k}$, $E_{(k,c_2)} = v\hat{\lambda}_k e^{-\hat{\lambda}_k}$, and $E_{(k,c_3)} = v - E_{(k,c_1)} - E_{(k,c_2)}$. Values of $E_{(k,c_j)}$ are shown in Table 8.

3.5 Compute chi-square statistics:
$\chi^2_{d_k} = \sum_{j=1}^{3} \frac{(O_{(k,c_j)} - E_{(k,c_j)})^2}{E_{(k,c_j)}}$, for $1 \leq k \leq 4$. Table 8 shows values of $\chi^2_{d_k}$. As a result, $\chi^2_{d_1} = 0.34$, $\chi^2_{d_2} = 0.23$, $\chi^2_{d_3} = 1.21$, $\chi^2_{d_4} = 2.14$.

3.6 Obtain degree of freedom: $d_k = 3$ - (number of independent parameters fitted) - 1. Because one parameter $\hat{\lambda}_k$ is estimated, $d_k = 1$, for $1 \leq k \leq 4$.

4. Obtain the summation of the chi-square statistics, $\chi^2_m = \chi^2_{d_1} + \chi^2_{d_2} + \chi^2_{d_3} + \chi^2_{d_4} = 3.92$. Due to the independence assumption in H_0, χ^2_m is a chi-square statistic of degree $m = d_1 + d_2 + d_3 + d_4 = 4$ since the sum of independent chi-square random variables is also a chi-square random variable [73].

With a confidence level at 95%, a threshold value is obtained as $\chi^2_{0.05,4} = 9.49$, where $\Pr(\chi^2 < \chi^2_{0.05,4}) = 0.95$, and $\Pr()$ is a chi-square distribution. The sum of chi-square

Table 8: Chi-square statistics. $S_{(k,c_j)} = \frac{(O_{(k,c_j)} - E_{(k,c_j)})^2}{E_{(k,c_j)}}$, $1 \leq k \leq 4$, $1 \leq j \leq 3$.

		T_1			T_2		
j	c_j	$O_{(1,c_j)}$	$E_{(1,c_j)}$	$S_{(1,c_j)}$	$O_{(2,c_j)}$	$E_{(2,c_j)}$	$S_{(2,c_j)}$
1	0	21	19.1746	0.1738	12	11.8996	0.0008
2	1	13	14.0990	0.0857	11	12.1376	0.1066
3	> 1	6	6.7265	0.0785	10	8.9627	0.1200
$\chi^2_{d_k}$				0.3379			0.2275
		T_3			T_4		
j	c_j	$O_{(3,c_j)}$	$E_{(3,c_j)}$	$S_{(3,c_j)}$	$O_{(4,c_j)}$	$E_{(4,c_j)}$	$S_{(4,c_j)}$
1	0	19	18.0714	0.0477	27	23.7327	0.4498
2	1	11	13.9011	0.6054	6	9.8886	1.5292
3	> 1	9	7.0275	0.5536	3	2.3787	0.1623
$\chi^2_{d_k}$				1.2068			2.1412

statistics obtained from the non-uniform Poisson point process is $\chi^2_4 = 3.92 < \chi^2_{0.05,4}$ [73]. Thus, the hypothesis cannot be rejected. This suggests that these 109 groups of subnets became unreachable at distinct time epochs statistically independently.

In the last disjoint interval, there are only 15 unique initial times. Thus, we only have three sub-intervals with $\geq$1 unique initial time, where the expected number of sub-intervals with 0, 1, or $\geq$1 unique initial times is at least five. This may suggest uncertainty; however, we do not expect the uncertainty in this last interval to affect the non-rejection of the hypothesis.

3.3.2 Isolated Unreachability

There are 72 out of 109 groups (66.06%), each of which contains one subnet that became unreachable at a large time scale. Hence, these 72 subnets became unreachable statistically independently. The inter-unreachable times of these unreachable subnets vary from 9 to 72 minutes. These subnets belong to 57 organizations, 49 ASes, and 25 ISPs. The organizations and ASes range from small to large entities, e.g., hospitals (University of Texas Medical Branch at Galveston), government agencies (NASA, Texas Learning and Computation Center), local businesses (Houston Association of Realtors, RMI Physician Services), local ISPs (Awesomenet, Eastex Net), and major ISPs (Comcast, XO).

Table 9: Example of withdrawal bursts belonging to two subnets from the same organization and the same AS (Announce = BGP announcement, Withdraw = BGP withdrawal).

		Prefixes	
		66.212.124.0/24	66.212.127.0/24
Type	BGP Peer	Arrival time (9/12)	Arrival time (9/12)
Announce	203.181.248.168	9:08:03 p.m. *ASPath:* 7660 2516 1239 - 3356 6395 16852 6361	9:08:03 p.m. *ASPath:* 7660 2516 1239 - 3356 6395 16852 6361
Withdraw	203.62.252.186	9:08:05 p.m.	9:08:05 p.m.
Withdraw	209.123.12.51	9:08:30 p.m.	-
Withdraw	137.164.16.12	9:08:50 p.m.	9:08:50 p.m.
Withdraw	209.123.12.51	-	9:09:01 p.m.

3.4 Temporal and Logical Dependence

We now study temporal dependence of the remaining 37 groups of subnets that became unreachable at small time scale. We then examine whether the temporal dependence leads to the dependence in the logical space of organizations and ASes.

For the 37 groups, each contains two or more subnets that became unreachable in less than three minutes. The inter-unreachable times within a given group vary between 0-2.52 minutes, with an average of 8.29 seconds. We examine each group to determine whether subnets from the same group became unreachable dependently.

3.4.1 Within-Organization Dependence

We begin with unreachable subnets from the same organization and the same AS. Patterns of BGP withdrawal bursts provide basic information for understanding the dependence. Table 9 presents an example of two withdrawal bursts belonging to two subnets from the same group and the same organization. The patterns of these withdrawal bursts are identical, i.e., messages from the same BGP peers have the same updated ASPATHs; only the arrival times of the messages exhibit a delay of a few seconds. Thus, if subnets from the same group and the same organization exhibit a similar pattern of withdrawal bursts, these two subnets are dependent within an organization.

The similarity of two withdrawal bursts for two unreachable subnets in a group is measured by the correlation coefficient of inter-arrival times of BGP updates from the same peer

routers. If the correlation coefficient is close to one, two withdrawal bursts are considered similar.

Burst Pattern Correlation: To determine whether two BGP bursts have the same pattern, we compute the correlation coefficient using inter-arrival times of BGP updates inside the bursts.

Let r_k be a BGP peer router that has logical connectivity to subnets i and j respectively, where $1 \leq i, j \leq n$, $1 \leq k \leq m$, n is number of subnets, and m is number of BGP peer routers. Let T_{ik} and T_{jk} respectively be the inter-arrival times of BGP updates from BGP peer router r_k to subnets i and j. Since we consider the unreachable subnets that occurred within less than three minutes as a group, the correlation coefficients are also computed using the delay τ, $\tau \in [0, 3]$ minutes.

The correlation coefficient of the two burst patterns is

$c_{(i,j,\tau)} = \frac{1}{m}\sum_{k=1}^{m} \frac{E((T_{ik}+\tau)-\hat{E}[T_{ik}+\tau])(T_{jk}-\hat{E}[T_{jk}])}{\hat{\sigma}_{T_{ik}}\hat{\sigma}_{T_{jk}}}$,

where $\hat{E}[\cdot]$ and $\hat{\sigma}^2$ are the sample mean and the sample variance. If $c_{(i,j,\tau)}$ of subnets i and j is close to one, bursts of these two subnets exhibit the same pattern.

We find 24 within-organization dependent groups, each of which has a correlation coefficient between 0.9370-1.0000. Each group contains 2-10 unreachable subnets. These 24 groups correspond to 103 subnets, 19 organizations, 20 ASes, and 15 ISPs. There are 50% of within-organization dependent groups that consist of subnets from business sectors, e.g., Baker Hughes, Christus Health. The other 50% contain subnets belonging to local ISPs (e.g., Gower, Moore, Internet America) and major ISPs (e.g., Suddenlink).

The majority of subnets in each group have the same ISP (90.62%) and the same unreachability duration (83.61%). There are 19 out of 24 groups, each of which has 100% of subnets with the same unreachability duration. These 19 groups exemplify strongly dependent unreachable subnets within organizations that became and remained unreachable at the same time.

Our findings of within-organization dependence illustrate that organization is a logical variable whose subnets can become unreachable dependently.

3.4.2 Cross-Organization and Within AS Dependence

We now study groups that contain subnets from different organizations but the same AS. We analyze the temporal dependence from the patterns of their withdrawal bursts. As these subnets belong to the same AS, their bursts exhibit similar patterns to those of within-organization dependence. Thus, we use the same criterion based on correlation coefficients to determine the similarity of burst patterns.

We find six cross-organization dependent groups whose correlation coefficients of BGP bursts are between 0.9493-1.0000. Each group contains 2-23 subnets. There are 42 subnets from these six groups that belong to 14 organizations, six ASes, and six ISPs. Similar to within-organization dependence, the majority of the unreachable subnets in each group have the same ISP (100%), and the same unreachability duration (77.78%). Three out of six groups contain at least one subnet from an ISP, e.g., Windstream Communications. Hence, a characteristic of cross-organization dependence is that subnets from different organizations that became unreachable dependently can be ISP subnets.

Our findings of cross-organization dependence show that AS is another logical variable that characterizes subnets from different organizations to become unreachable dependently.

3.4.3 Cross-AS Dependence

We now move up to the AS-level and examine whether subnets from the same group but from different ASes became unreachable dependently.

Table 10 shows an example of withdrawal bursts belonging to two subnets from the same group but different ASes. The patterns of withdrawal bursts are less similar than those of within- or cross-organization dependence since subnets from different ASes can have different BGP peer routers. To measure the similarity between the two withdrawal bursts, we use only the inter-arrival times of BGP updates from the common BGP peer routers to compute the correlation coefficient.

Among seven cross-AS dependent groups, the correlation coefficients of BGP bursts vary between 0.8590-1.0000, with 0.9590 on average. Each group contains 2-40 subnets. These seven groups correspond to 65 subnets, 26 organizations, 17 ASes, and 15 ISPs. Unlike

Table 10: Example of withdrawal bursts belonging to two subnets from the same group but different ASes (Annouce = BGP annoucement, Withdraw = BGP withdrawal).

		Prefixes	
		216.230.224.0/20	64.134.11.0/24
Type	BGP Peer	Arrival time (9/13)	Arrival time (9/13)
Withdraw	154.11.11.113	5:18:10 a.m.	5:17:39 a.m.
Annouce	203.181.248.168	5:18:10 a.m. *ASPath:* 7660 2516 1239 - 3561 40156	5:17:39 a.m. *ASPath:* 7660 2516 1239 - 3356 14654
Withdraw	195.22.216.188	-	5:17:41 a.m.
Annouce	62.72.136.2	5:18:15 a.m. *ASPath:* 5413 702 701 - 3561 40156	5:17:45 a.m. *ASPath:* 5413 702 701 - 3356 14654
Annouce	194.85.4.55	5:18:17 a.m. *ASPath:* 3277 3216 702 - 701 3561 40156	-
Withdraw	164.128.32.11	5:18:27 a.m.	-
Withdraw	194.85.4.55	5:18:31 a.m.	5:17:47 a.m.
...	...	...	...
Withdraw	209.123.12.51	5:19:18 a.m.	5:19:18 a.m.

within- and cross-organization dependence, only 30.60% and 17.02% of subnets in cross-AS dependent groups have the same ISP, and the same unreachability duration, respectively.

Unreachable subnets from cross-AS dependent groups also exhibit characteristics related to ISPs. For example, there are five groups that have at least one subnet from organizations that are major ISPs, e.g., Suddenlink and Comcast. One of these seven groups exhibit additional ISP characteristic, where one AS is the ISP of the other, i.e., AS16687 is the ISP of AS12117.

Our findings suggest that cross-AS dependent subnets are from organizations that are ISPs.

3.5 Logical Network Hierarchy

We present the network dependence in a logical hierarchy using four-tuples of unreachable subnets: prefixes, organizations, ASes, and ISPs. Figure 19 illustrates the example of logical hierarchy with temporal dependent relationship. Figure 20 shows the logical hierarchy for all unreachable subnets where subnet, organization, AS, and ISP nodes are shown in order from the outside to inside. Figure 20 also highlights the dependent and independent unreachable

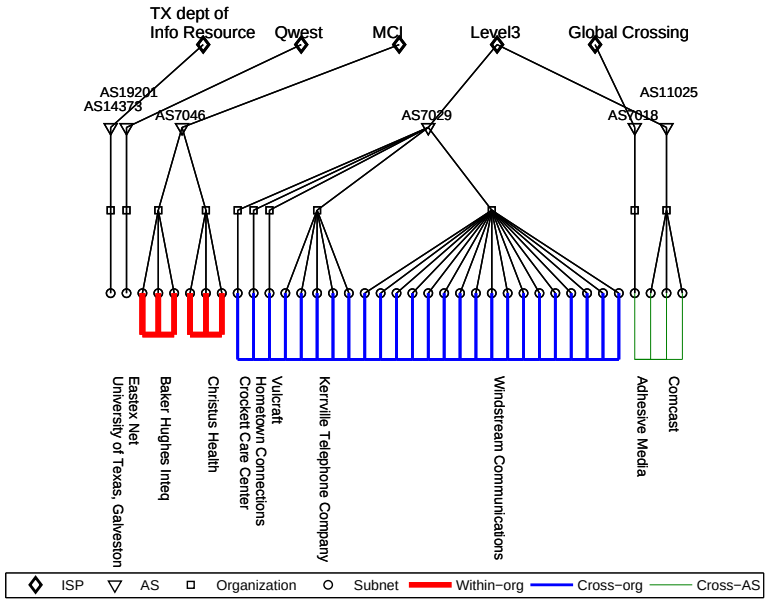


Figure 19: Example of logical network hierarchy.

subnets, as well as provides the following observations:

- At the subnet-level, 72 subnets became unreachable independently. 81.94% subnets are /24 subnets, and the organizations who own these subnets range from small entities to major ISPs. This shows that a large number of subnets that became unreachable independently are at the network edge, and organizations who own these networks could be of any size.

- At the organization-level, there are 16 organizations and 16 ASes that correspond to only within-organization dependent subnets. Hence, within-organization dependence has one-to-one relationship between organizations and ASes. Among these organizations, eight are local ISPs, and the remaining are business sectors.

- At the AS-level, there are five ASes and 12 organizations that contain only cross-organization dependent subnets. There are 12 ASes and 20 organizations that correspond to only cross-AS dependent subnets. Hence, for cross-organization and cross-AS dependence, one AS corresponds to many organizations. Furthermore, cross-organization and cross-AS dependent subnets are commonly from organizations that are ISPs.

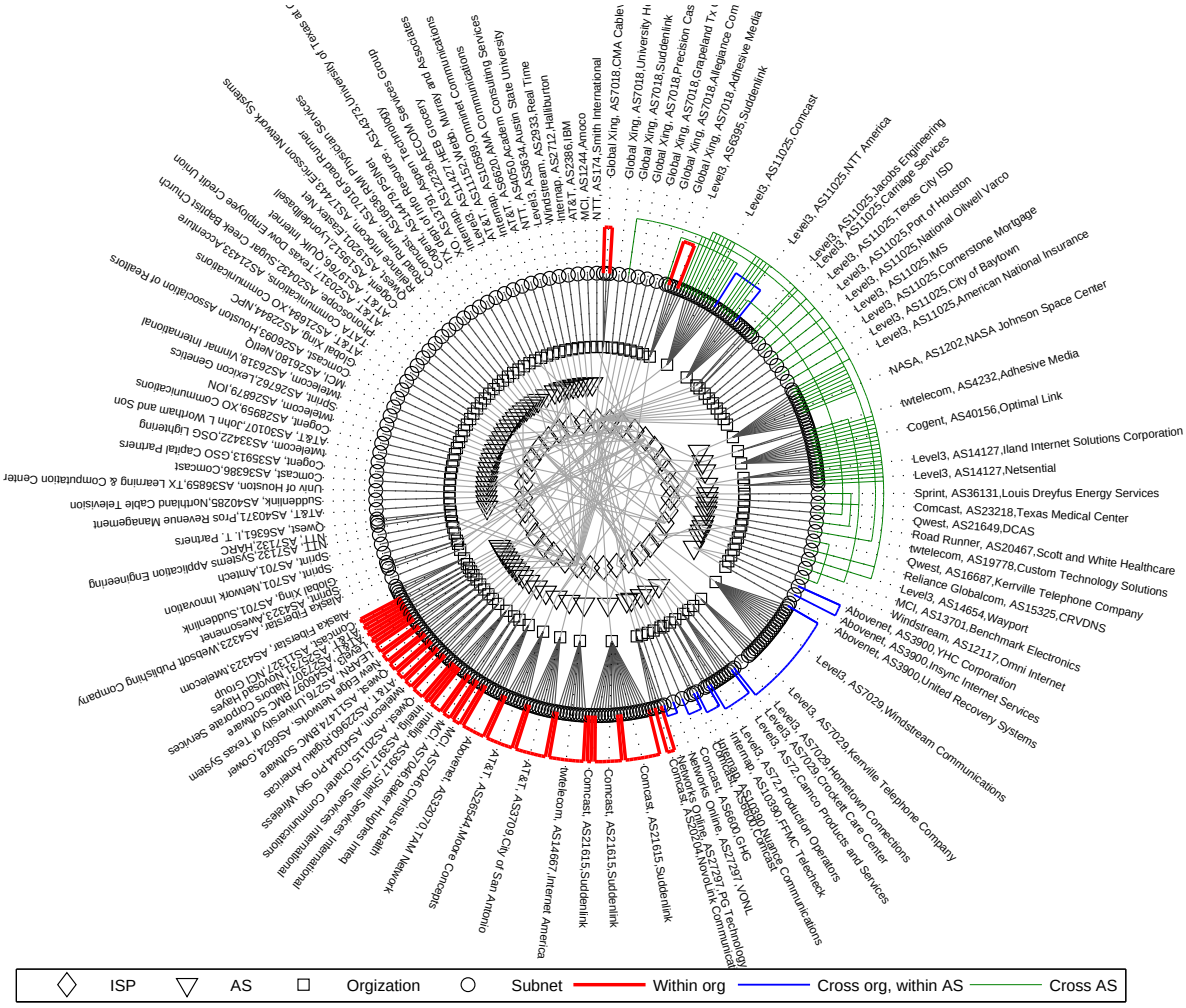


Figure 20: Logical network hierarchy: subnets, organizations, ASes, and ISPs (from outside to inside nodes). Each text label respectively contains ISP, AS, and organization.

- At the ISP-level, eight ISPs are tier-1 networks. Twenty one ISPs serve isolated, within-, and cross-organization dependent subnets all together but not cross-AS. Only one of these 21 ISPs is the tier-1 network. For the remaining 15 ISPs that provide services for cross-AS dependent subnets and others, seven of them are tier-1 networks and three are major networks, e.g., Comcast and Abovenet. This suggests that isolated, within-, and cross-organization dependence occurred more locally that cross-AS dependence.

Overall, there are 217 isolated, within-, and cross-organization dependent unreachable subnets (76.95%), compared to 65 cross-AS dependent subnets. Since the majority of subnet

Table 11: Comparison of subnet unreachability between the pre-Ike and Ike periods.

	Pre-Ike	Ike
Number of independent unreachable subnet groups	5.75	109
Number of unreachable subnets	39.00	282
Number of isolated subnets	8.44	72
Number of within-organization dependent groups	4.25	24
Number of cross-organization dependent groups	0.63	6
Number of cross-AS dependent groups	0.88	7
Average inter-unreachable time	4.33 hours	0.02 hours
Average unreachability duration	0.16 days	1.22 days

unreachability occurred at local organizations, local ISPs, and major ISPs but few at tier-1 networks, this suggests that Hurricane Ike impacted subnets at local network operations.

3.6 Comparison with Normal Operations

We compare 282 unreachable subnets relating to Hurricane Ike with subnet unreachability during the pre-Ike interval (August 1-September 9) when no major network disruptions were reported [62].

We separate 40-days of the pre-Ike BGP update messages into smaller disjoint data sets, each of which has the same duration of 2.5 days as the Ike period. This results in 16 pre-Ike data sets. The average number of unreachable subnets is obtained per data set, and then used as the baseline to compare with subnet unreachability during the Ike period.

Table 11 provides the comparison of subnet unreachability between the pre-Ike and Ike periods. Subnet unreachability occurred during Ike are 7.23 times greater in quantity, at 216.50 times higher rate, and with 7.63 times longer duration. Furthermore, the pre-Ike unreachability randomly occurred across Texas whereas subnet unreachability occurred non-uniformly during the Ike period. Fifty-nine unreachable subnets occurred inside the wind-radii of the hurricane (20-75 miles), and 156 subnets (53.61%) occurred within 125 miles from the hurricane storm path. Thus, our findings show that subnet unreachability during the Ike period between 12:00 a.m. September 12 and 12:00 p.m. September 14 is indeed anomalous.

3.7 Societal Impact

From the list of organizations obtained from Whois database [85], we categorize them as follows.

Healthcare and related institutions: There are two hospital institutions: University of Texas Medical Branch at Galveston, and Texas Medical Center. These two subnets resumed connectivity after 16 hours and 1.84 days after they became unreachable. Besides these hospitals, there are health insurance companies: Scott and White Healthcare, Christus Health, and RMIPSC. They became unreachable for 15.13 minutes, 3.48 hours, and 17.04 hours, respectively.

Education and university systems: Several networks that became unreachable during Hurricane Ike. These are: Texas City Independent School District, Austin State University, University of Texas System, Texas A&M Network, Texas Learning&Computation Center, and Houston Advanced Research Center. Texas City Independent School District and Austin State University respectively were disrupted for 0.63 and 5.40 hours. The remaining networks became unreachable from 22.62 hours up to seven days.

Government agencies: Several government agencies also had their subnets became unreachable. These include Port of Houston, City of San Antonio, and NASA. The Port of Houston and City of San Antonio subnets respectively were disrupted for 0.63 and 3.94 hours. NASA had many unreachable subnets with various unreachability durations.

Oil-related businesses: Hurricane Ike impacted on oil refineries and related businesses, including the following: National Oilwell Varco, Shell, Amoco, Baker Hugh Inteq, and Production Operators. The corresponding durations of these organizations are largely varied between 0.63 hours and 5.32 days.

Internet Service Providers: Many local ISPs that were disrupted during Hurricane Ike, for example, Eastex Net, Bravo, Gower, Optimal link, Real Time, Awesomenet, and CRVDNS. However, their unreachability durations are small, i.e., 0.45-4.86 hours, and only Awesomenet was disrupted for the longest period of 4.63 days.

There are also regional ISPs (serve few states in U.S.) whose subnets were disrupted, for example, Allegiance Communications, CMA Cablevision, and Northland Cable Television.

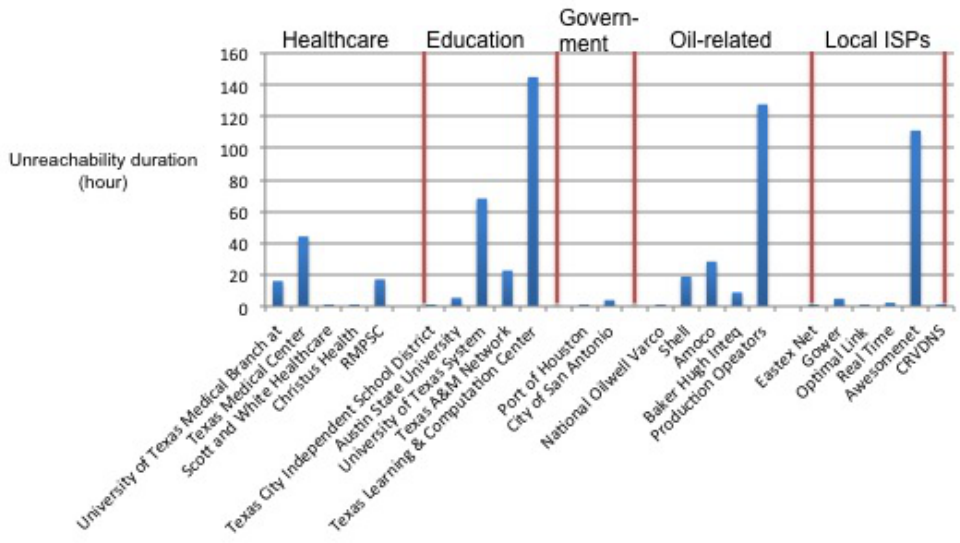


Figure 21: Unreachability durations of various organizations.

These subnets became unreachable between 5.58 hours-1.30 days.

We also observe that the major ISPs experienced disruptions after Hurricane Ike. These are Windstream, Suddenlink, and Comcast-Houston. Their unreachability durations are varied.

In the past discussion with our collaborating ISP from Hurricane Katrina (see Section 2.8), the network-service recovery is prioritized with communications service providers, government agencies, and healthcare institutions to receive high priority. Figure 21 present unreachability durations of organizations listed. Compared to other organizations, communications service providers, government agencies, and healthcare institutions resumed connectivity relatively early after Hurricane Ike. Although there are many local ISPs that became unreachable, they were disrupted for a short time and also resumed the connectivity relatively early. Education and university networks experienced longer unreachability than other organizations. Our findings illustrate that network-service recovery for these organizations were performed according to priority.

3.8 Summary

In this work, we identify unreachable subnets caused by Hurricane Ike in 2008 and analyze their temporal and logical dependence. It is found that subnets became unreachable at

large time scale and in groups, and subnets became dependently within-organization, cross-organization, and cross-AS. It is found that most organizations corresponding to within-organization dependence are business sectors and ISPs. This confirms the strong dependence at local organizations. Furthermore, it also shows that local organizations and ISPs are both vulnerable to the hurricane.

Besides the findings that the cross-organization dependent subnets are likely to be ISP subnets, it is also found that are Hurricane Ike caused multiple subnets entirely from one AS to become unreachable. It is found that cross-AS dependent unreachable subnets are from major ISPs whose service providers are tier-1 networks. The comparison of subnet unreachability between Hurricane Ike and normal operations shows that subnet unreachability relating to Ike is anomalous.

Limitations and open issues:

There are limitations to this part of our research.

1. As previously mentioned in Chapter 2, BGP update messages may not provide a complete view of subnet connectivity. Our results tell which subnets were unreachable from BGP routers where the unreachability corresponds to two scenarios. The first is that a subnet was disconnected from the rest of the network. The second is that the subnet maintained its connectivity to some parts of the network but could not be reached from BGP routers, possibly due to other disconnectivity in the network [21]. Therefore, BGP update messages can be used to infer reachability of subnets but do not provide sufficient information to delineate these cases. Nevertheless, a significant number of unreachable subnets were at the network edges or the end of ASPATH. This illustrates that these subnets were indeed unreachable.

2. Although GeoIP database [49] is selected to replace Whois database [85] as our geo-location source, Whois database is still used in this part of research to provide organization information. We perform the accuracy check of Whois information for 50 organizations and find that only 56% of organization information in Whois database is up to date. The accuracy check is presented in Appendix A.

Prior work in networking focused mostly on identification of network statuses after the hurricanes [18, 42]. The interactions between subnet unreachability and organizations have not been considered. In this work, we study how network disruptions occur depending on social organizations. Our work adds organization variables into this study to understand how different social entities respond to network disruptions caused by large-scale disturbances. In the next Chapter 4, we incorporate the storm data and the subnet geo-locations to analyze how network disruptions occurred depending on weather.

CHAPTER IV

ANALYSIS OF NETWORK DISRUPTIONS ON WEATHER DEPENDENCE

Large-scale natural disasters push the communication infrastructure to the extreme and expose the possibly counterintuitive interactions among networks, organizations, and weather that are not observable in day-to-day operations. One essential aspect is to understand whether and how external disturbances such as hurricanes impact on network disruptions. Prior work has not considered weather in the study of network disruptions; our work aims to understand whether and how weather plays in a role of network disruptions.

Network administrators respond differently once they hear of an impending hurricane. Thus, the first assumption one could make is that there is causality of natural disasters such as Hurricane Ike on subnet unreachability during these times of extreme weather since network administrators make abnormal decisions in response to the hurricane. For this study, we consider causality between subnet unreachability and the storm if a subnet becomes unreachable after it appears in the storm coverage and experiences the physical impact of the hurricane.

We incorporate various storm data with subnet geo-locations to understand how network disruptions occur relative events due to the hurricane. We encounter difficulties where geo-locations of subnets are either inaccurate or unavailable; thus, we introduce a probability measure to incorporate uncertainty of subnet geo-locations. We provide visual characterization between the initial times of subnet unreachability and the storm hitting times.

An interesting observation emerges from our study: some subnets became unreachable hours before areas where they were located appeared in the storm coverage. In addition, the initial times of subnet unreachability and the storm hitting times are shown to be weakly correlated. These findings imply that subnet unreachability might not directly be caused by physical impact of the hurricane. Moreover, there might exist hidden causes which are

not obtainable from the publicly available data. Therefore, these possible hidden causes lead us to search for ground truth of network disruptions in the next Chapter 5.

Chapter 4 is organized as follows. Section 4.1 introduces the storm and subnet geo-locations data used in this study. Section 4.2 characterizes subnet unreachability and storm interactions, and Section 4.3 computes the correlation coefficients between subnet unreachability and the storm. Section 4.4 provides the comparison of subnet unreachability between Hurricanes Katrina and Ike. Section 4.5 summarizes Chapter 4.

4.1 Heterogeneous Data

Two new data sources: storm data and geo-locations of subnets are added to this research to understand how networks are impacted by hurricane.

4.1.1 Storm Data

The storm data of Hurricane Ike are obtained from National Hurricane Center (NHC). NHC, a part of National Weather Service, monitors and analyzes the tropical storms in both North Atlantic and Pacific basins and issues warnings and forecasts to public.

From NHC, we collect the observation data of Hurricane Ike from the public and forecast advisories [7] and the best track data [6] from 10:00 a.m. September 12 to 4:00 a.m. September 14. Note that during this period, Ike weakened to a tropical storm at 1:00 p.m. September 13 and passed the state of Texas by 2:00 a.m. September 14. The time scale of the available storm data is in scale of hours during the Ike landfall, and in three hours before and after the landfall. In total, there are 27 available advisory reports. The collected storm data consist of latitude and longitude coordinates of storm center, storm speed, and wind radii of hurricane force winds (at 74 miles or more per hour).

We use the storm data to reconstruct the storm path and coverage as a reference to understand when subnets became unreachable. The storm path is a trajectory of the hurricane center, and the coverage is an area of the hurricane surface spanned by a wind radii of hurricane force wind. The constructed storm path and coverage are presented in Figure 22. The red disks illustrate storm coverage at different hours, and the height of the disks represents time, not actual spatial position of storm. Storm coverage moved toward inland

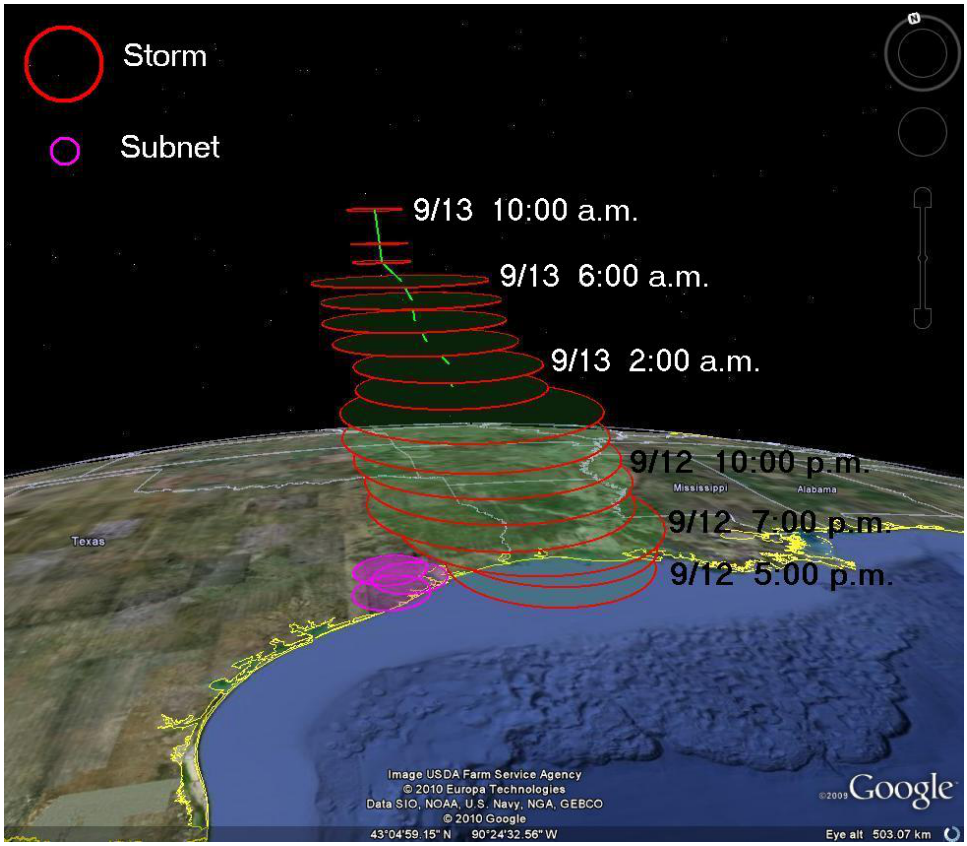


Figure 22: Reconstructed path and coverage of Hurricane Ike

as time progresses. The wind radii of hurricane force winds first increased until 10:00 p.m. September 13 reached its maximum, and decreased afterward.

Despite the storm data that is available hourly, the data can be interpolated into smaller scale at 15 minutes as presented in Appendix B. The interpolated data at 15-minute intervals are constrained to match the observation data of every hour. Note that we experiment that this constraint could be violated if an interpolation interval is too small, e.g., less than 10 minutes.

4.1.2 Geographic Locations of Subnets

The geo-locations of subnets are needed to relate the unreachability to the hurricane. Specifically, the geo-locations are obtained from GeoIP City from Maxmind [49]. Maxmind provides geo-locations of IP addresses in terms of latitude and longitude, and reports that approximately 79% of their provided geo-locations are "correctly resolved within 25 miles

from a true location" [49].

We obtain the geo-location of every IP-address inside each prefix (e.g., 256 locations for a /24 subnet) from GeoIP. Each geo-location then becomes the center of a 25-mile radius disk which incorporates the uncertainty of the true location. The geo-location of a subnet is presented as a region from the union of all geo-location disks belonging to all its IP-addresses. The actual (unknown) geo-location of the BGP router can be any where within the union region. The example of one subnet region consisting of three geo-location disks is shown in Figure 22.

4.2 Characterizing Network and Storm

We continue using subnet unreachability result from Chapter 3. Consider 282 unreachable subnet i with geo-location S_i, $1 \leq i \leq 282$. Let R_t be the hurricane coverage at time t which consists of a storm center and wind radii, for $t \in$ [12:00 a.m. September 12, 12:00 p.m. September 14]. As the storm moves across the region, the storm appears in a random part of S_i. Hence, this characterizes the likelihood for the actual subnet to be in the storm coverage. We define $p_{(i,t)} = \frac{|S_i \bigcap R_t|}{|S_i|}$ to be the coverage probability that subnet i appears in the hurricane coverage at time t, where $|\cdot|$ is the area of a region. Coverage probability indicates how likely for subnet to experience physical impact by the hurricane.

Figure 23 shows the color-coded values of coverage probability for 282 unreachable subnets, where each row corresponds to one subnet. The coverage probability for a subnet started to increase at 6:15 p.m. September 12. Then, the coverage probability increased until $p_{i,t}$ reached a maximum value, and then decreased as storm moved out of subnet region S_i. This spatially and temporally demonstrates how the hurricane passed through subnets. All subnets were out of the hurricane coverage by 9:00 p.m. September 13. There were 65 subnets that did not appear in the storm coverages despite their geo-locations in Texas.

We define the "hitting time" as the time the geo-location of a subnet first overlaps with the hurricane coverage. This corresponds to the time when probability $p_{(i,t)}$ first becomes positive. The empirical distribution of the hitting times is obtained by projecting these hitting time epochs to the top figure in Figure 23. The hitting times of 217 subnets

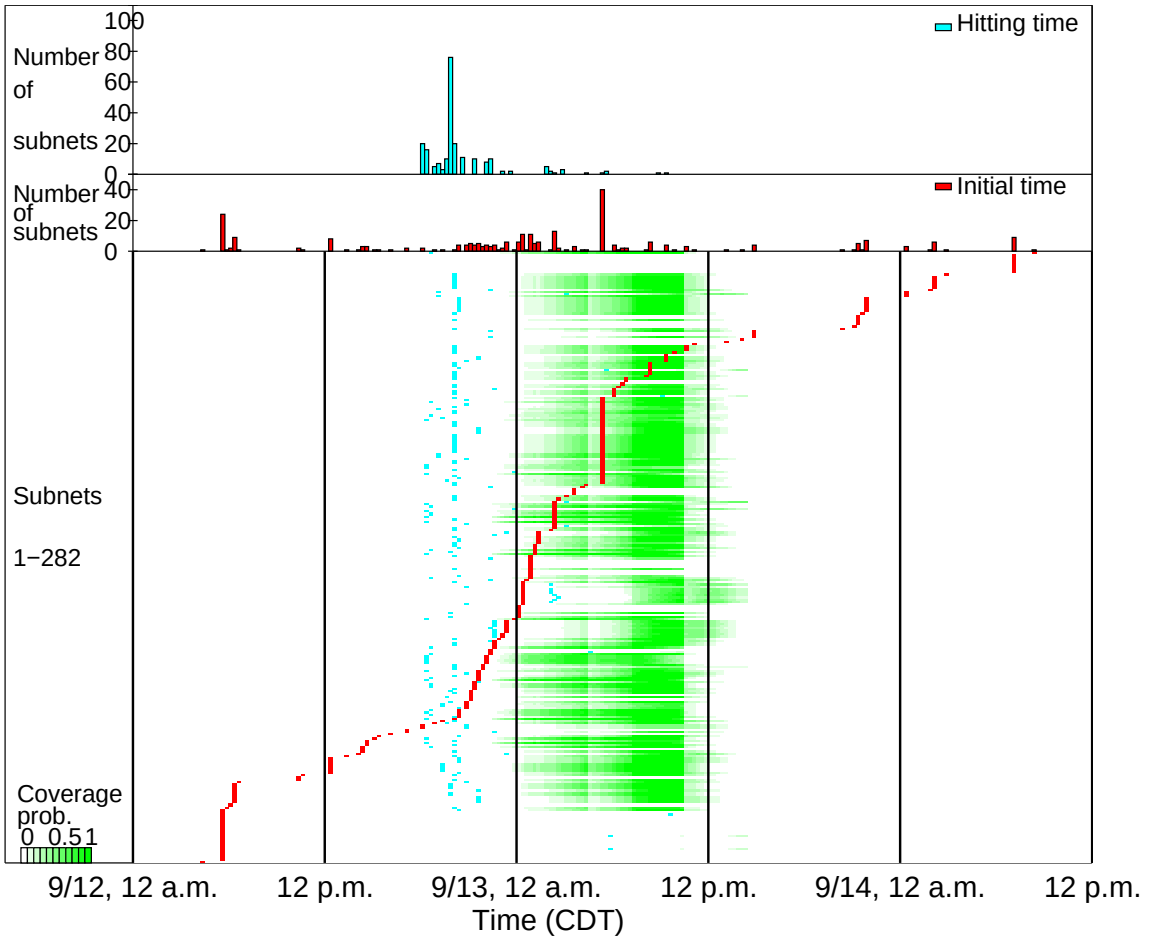


Figure 23: Distributions of initial times, hitting times, and coverage probability.

occurred between 6:15 p.m. September 12 and 9:30 p.m. September 13, and the majority of subnets first appeared in the storm coverage between 7:45-8:00 p.m. September 12.

The empirical distribution of the initial times when the subnets first became unreachable is shown in the middle of Figure 23 for comparison. The initial times span a longer duration than the hitting times. A large percentage of the subnets became unreachable between 5:15-5:30 a.m. September 13, approximately 9.50 hours after the storm coverage first appeared in the majority of subnet regions.

We consider the "non-causality" when a subnet became unreachable before the storm coverage first appeared in its region, and the subnet experienced the physical storm impact. Figure 23 shows that 117 subnets (41.49%) became unreachable non-causally, where 52 subnets had their initial times (in red) occurred earlier than their hitting times (in blue).

Note that 65 subnets never appeared in the storm coverage.

In contrast, 165 subnets (58.51%) became "causally" unreachable, i.e., after storm coverage appeared in their subnet regions as shown in Figure 23. This can also be observed from the delay between the two peaks of the hitting-time and the initial-time empirical distributions.

4.3 Network and Storm Correlation

We now correlate the initial time of unreachability with the hitting time for individual subnets. Let t_{hi} and t_{ui} be the hitting time and the initial time of subnet i respectively, for $1 \leq i \leq 282$. Let $I_i(t)$ and $I_{hi}(t)$ be two indicator functions for subnet i, where

$$I_i(t) = \begin{cases} 1 & \text{if } t = t_{ui}, \\ & \text{i.e., when subnet } i \text{ became initially unreachable;} \\ 0, & \text{otherwise.} \end{cases} \tag{7}$$

$$I_{hi}(t) = \begin{cases} 1 & \text{if } t \in [t_{hi} - T, t_{hi}], \\ & \text{i.e., the storm hit the subnet region } i \text{ when } t \text{ falls in this interval;} \\ 0, & \text{otherwise.} \end{cases} \tag{8}$$

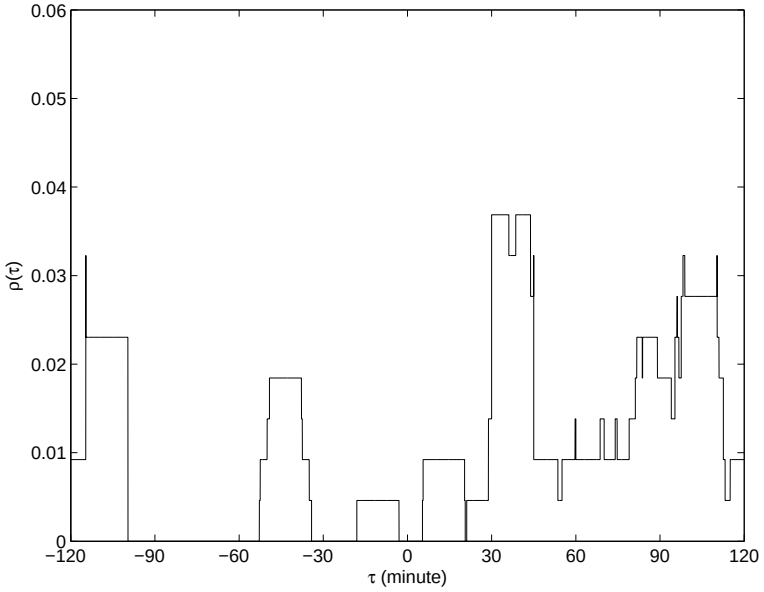
$T > 0$ is a parameter that takes into consideration of discrete storm coverage at the scale of 15 minutes. For example, $T = 15$ assumes that the actual hitting time finer than 15 minutes would be uniformly distributed in $[t_{hi} - 15, t_{hi}]$.

The sample correlation function between the hitting times and initial times is computed as:

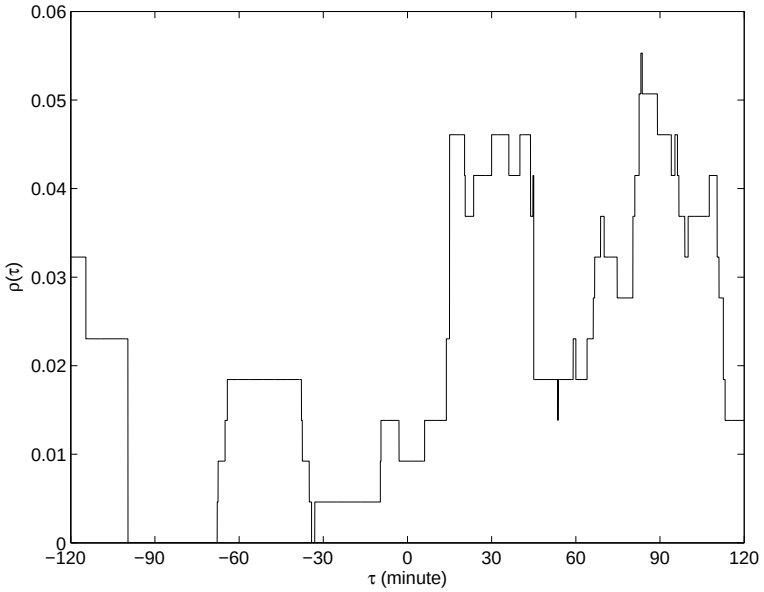
$$\rho(\tau) = \frac{1}{n} \sum_{i=1}^{n} I_i[t_{ui}] I_{hi}[t_{ui} - \tau], \tag{9}$$

where $\tau > 0$ is a delay variable. Intuitively, $\rho(\tau)$ shows the likelihood for a subnet to became unreachable τ minutes before being in the storm coverage.

Figure 24(a) presents $\rho(\tau)$ where τ varies between [−2,2] hours when $T = 15$ minutes. The maximum sample correlation coefficient is 0.0369 at τ=30.02 minutes. To illustrate the robustness of this result with respect to T, Figure 24(b) presents $\rho(\tau)$ for T=30 minutes.



(a) 15 minutes



(b) 30 minutes

Figure 24: Sample correlation coefficients for $T = 15$ and 30 minutes.

The maximum sample correlation coefficient is 0.0553, when τ=83.20 minutes and $T = 30$ minutes. To fairly interpret these results, we note that 25-mile radius used in constructing a subnet region bounds an actual subnet location, and thus results in the earliest hitting times for the subnet. Hence, the magnitude of maximum correlation should be accurate with probability 0.79, where 0.79 characterizes the likelihood of correctness of geo-locations within 25 miles [49].

The above results show that sample correlation coefficients are consistently small; thus, the hitting times and the initial times are weakly correlated. In addition, the positive values for τ when the correlations are maximum suggest the non-causality, i.e., subnets could become unreachable before being in the storm coverage.

All these findings suggest that the storm may not be the only direct cause of subnet unreachability, and there may exist hidden variables that are not inferable from the data. This motivates us to seek for the actual root causes of subnet unreachability in Section 25.

4.4 Katrina and Ike Comparison

From two case studies of subnet unreachability from Hurricanes Katrina and Ike, we compare the distributions of initial times and unreachability durations. Tables 12 and 13 respectively present the percentages of unreachable subnets with different initial times and unreachability durations from both hurricanes. The result shows that there exists substantial subnet unreachability prior to the landfall for both Katrina (50.79%) and Ike (54.24%). Larger subnet unreachability occurred much earlier than the landfall for Hurricane Ike (23.40%). This signaled the better response to the evacuation announcement for Ike than Katrina. For the unreachability durations, subnet unreachability caused by Katrina lasted longer than Ike. The majority (72.83%) of Katrina unreachability lasted longer than one month whereas the majority (66.31%) of Ike unreachability lasted less than one day. This illustrates the smaller degree of impact on network-service disruptions caused by Hurricane Ike.

4.5 Summary

In this part of the research, we analyze the dependence between subnet unreachability and the hurricane by incorporating storm data to subnet geo-locations. It is found that there

Table 12: Comparison between Katrina and Ike: percentage of unreachable subnets with different initial times.

Initial time	> 6 hours before landfall	0-6 hours before landfall	0-18 hours after landfall	18 hours - 2 days	3-7 days
Katrina	5.12	45.67	37.01	6.69	5.51
Ike	23.40	30.85	32.98	12.77	-

Table 13: Comparison between Katrina and Ike: percentage of unreachable subnets with different unreachability durations.

Unreachability duration	Less than 1 day	1-3 days	3-7 days	1-2 weeks	2-4 weeks	Longer than 4 weeks
Katrina	7.09	6.30	2.76	3.54	7.48	72.83
Ike	66.31	19.15	12.77	1.77	0.00	0.00

exist subnets that became unreachable hours before they appeared in the storm coverage. Moreover, the initial times and the hitting times of subnet unreachability are weakly correlated. Certain causes of subnet unreachability might not directly be physical impact from the hurricane and not observable from the publicly available data. This provides the motivation for us to search for the ground truth of disruptions by directly considering the network owners.

Limitations and open issues:

There are limitations to our weather-dependence study of network-service disruptions.

1. The time scale of the available storm data is in hours during the Ike landfall, and in three hours before and after the landfall. Compared to the time scale of subnet unreachability available in seconds, the time scale of the storm data is very crude. National Climatic Data Center (NCDC) provides another type of weather data, namely, the hourly surface data [51] (See details in Chapter 7). The surface data is local weather measurements collected at different NCDC stations across the U.S. However,

during Hurricane Ike, local measurements were not available because of evacuation. Thus, to have more storm data available in smaller time scale, we propose the interpolation of the storm data as presented in Appendix B.

2. The accuracy of geo-locations data is a major issue in this research. A subnet region with 25-mile radius is large, compared to the wind radii of Hurricane Ike between 20-75 miles. Thus, our computation of hitting times may not be accurate. Since the subnet region with 25-mile radius is much larger than the actual location, our computed hitting times provide the earliest possible hitting times for the subnets. Therefore, our computed hitting times are likely to occur earlier than the initial times of unreachable subnets. As a result, we expect to see more subnets that became unreachable after they appeared in the storm or more storm causality. Nonetheless, our small sample correlation coefficients demonstrate that subnet unreachability and the storm are weakly correlated.

In the next Chapter 5, we seek ground truth on what may have caused subnets to become unreachable non-causally.

CHAPTER V

SEARCHING FOR GROUND TRUTH

The weak correlation between subnet unreachability and the storm provides us with the motivation to search for the ground truth of subnet unreachability. A ground truth is an actual root cause of network-service disruption.

Ground truth information that may explain the network responses to disturbances is in general proprietary to organizations who own the networks. Therefore, prior studies of network-service disruptions caused by large-scale natural or man-made disasters focus on the identification of network statuses. Little or no ground truth of network disruptions has been reported [18, 19, 33, 42].

In 2003, the Committee on the Internet Under Crisis Conditions formed by groups of industry and researcher representatives reported that network disruptions after the September 11 attack were caused by damage to hardware directly to the day's events [28]. In 2005, the U.S. House of Representatives released the investigation on communications failures during Hurricane Katrina [78]. It was reported that physical impact, power outages, and inoperability of communications equipment were the causes of Katrina network disruptions. The conclusions of these reports indicate that the ground truth in each situation was made available because they were reported by either ISPs who owned their ground truth information [28], or an administrative institution who had authority to request the ground truth information from the organizations involved [78]. Nonetheless, the third-party reports of network disruptions generally report on the symptoms of disruptions and give conjectures on the root causes [18, 33, 42]. No validation of the conjectures are given, and the ground truth does not resurface as the fact.

This work is a result of active attempts at contacting the owners of unreachable subnets and learning about their root causes. In Chapter 3, it is found that subnet unreachability depends on organizations; thus, to discover the ground truth, we contacted the very local

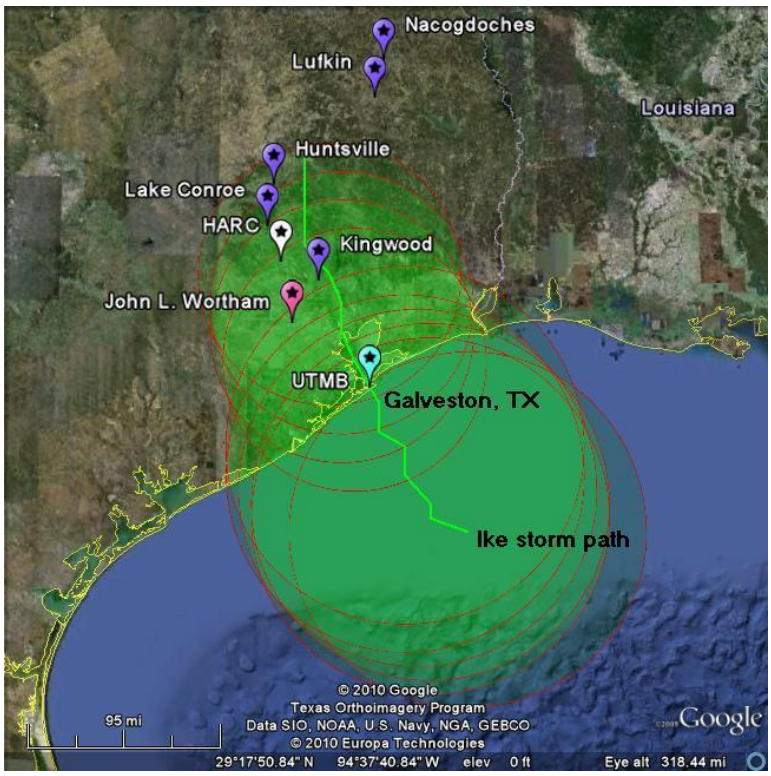


Figure 25: Known geo-locations of ground truth from HARC, UTMB, John L. Wortham, and Suddenlink.

information sources: organizations who own subnets. From more than 40 organizations contacted, only four responses were received. Beside the root causes, the information from the responses also includes the restoration processes.

We list the collected information of the ground truth from the network administrators of these four organizations. In addition, we include the information obtained online on root causes from three other organizations.

Chapter 5 is organized as follows. Section 5.1 reports ground truth information from seven organizations. Section 5.2 computes information bits provided by the ground truth, and Section 5.3 provides the summary.

5.1 Ground Truth Reports

Below, the collected information of the ground truth from the seven organizations is listed. Among these organizations, four are individual organizations, and the other three are ISPs. Figure 25 presents the known geo-locations of these organizations.

Houston Advanced Research Center (HARC) is located in the Woodlands, and its location is presented in Figure 25. Their network became unreachable on 2:10 p.m. September 12, approximately 12 hours prior to the Ike landfall at Galveston, and lasted for more than a week.

The network administrators shut down the network as a precaution due to insufficient spare power. In addition, HARC experienced power outages starting on 7:45 a.m. September 13. The power outages lasted until September 18, and their ISP experienced an unknown network connectivity problem on September 19. HARC was able to regain their Internet connections on 10:30 a.m. September 20. The administrators reported no physical damage caused by the hurricane.

University of Texas Medical Branch at Galveston (UTMB) is a hospital and located in Galveston where Hurricane Ike first made landfall. Its geo-location is presented in Figure 25. Their network became unreachable on 8:06 p.m. September 12 and lasted approximately for 16 hours.

The network administrators reported that there was staff staying on site during the hurricane; thus, communications was needed. They experienced power failures because of rising water, and some of their network equipment and power generators were on the ground level. Once the power was restored on the next day, their Internet connectivity operated normally.

Because the university used underground fiber cables, the network infrastructure itself did not receive any physical damage. The hospital website remained functioning throughout and after the storm because they used a remote site at another location in Dallas.

Internet America, Inc. is the ISP in the southwest of the U.S. We observe that all of their eight subnets became unreachable together three times: (1) 12:54 a.m. September 13, for 1.17 hours, (2) 2:10 a.m. September 13, for 5.73 hours, and (3) 8:54 a.m. September 13, for approximately 6 hours.

The network administrators confirmed the service disruptions and acknowledged that the root causes were power outages. Specifically, the subnets experienced power outages three times, which is consistent with our observation. Since this organization is an ISP,

they did not disclose the geo-locations of the root causes.

Lyndon B. Johnson Space Center of NASA is located at Houston and its vicinity. NASA also serves as the ISP to their affiliated contractors. Their networks experienced various service disruptions between 2:18 p.m. September 12 and 2:25 a.m. September 13, and the unreachability durations lasted from eight minutes up to more than six days.

The network administrators stated that some of their networks were unreachable due to power outages, and the electricity was quickly restored after the hurricane. Because of evacuation, the Center was closed between September 11-22. Therefore, their on-site data was not sufficient to confirm all cases of their network disruptions. NASA did not provide the geo-locations of the root causes.

In addition to the information from the administrators, the online news archive from this organization showed the physical damage to the Center caused by Hurricane Ike [34]. Hence, physical damage could be another root cause for the network outages.

In addition to these four organizations, we are able to obtain online information related to the root causes of three other organizations, as described below.

Texas Medical Center is a medical institution consisting of hospitals and universities. We observe that their network outages started from 1:08 a.m. September 13 and lasted for 1.35 days. The root cause of this organization was obtained from their power service provider: Centerpoint Energy [24]. Centerpoint Energy updated news report that they restored power for the Texas Medical Center before 8:00 p.m. September 13 . Because this organization consisted of several facilities, the geo-locations of the root causes are not presented here.

John L Wortham & Son, L.P. is an insurance company whose headquarters is located in Houston and shown in Figure 25. Their network became unreachable on 1:13 a.m. September 13 and lasted for about 3.29 days. This organization provided information online that their headquarters was physically inspected on September 13 and found that it was severely damaged. The company set up the trailers as their temporary workstations afterward, and they were back on operation on 10:30 a.m. September 15.

Suddenlink (former name: Cebridge Connections) is the ISP served in the midwest and southwest of the U.S. We observe that their networks became unreachable between 12:20 p.m. September 12 and 3:02 p.m. September 13, and lasted from 22 minutes up to eight days.

Suddenlink provided the post-Ike updates about their restoration processes in the following cities: Huntsville, Kingwood, Lake Conroe, Lufkin, and Nacadoches [23]. Figure 25 presents the locations of these cities. The root causes were loss of power and physical damage, e.g., fallen trees and damaged cables. The network restoration was completed on September 22.

They reported that their service restoration mainly depended on the power restoration. For example, in some areas, the network crew could not operate until the power restoration was completed. On the other hand, there were areas that the network infrastructure was completely repaired, but network services could not resume because the power was not yet restored. This illustrates that loss of power is the main root cause for this ISP.

Ground Truth Summary: Six out of seven organizations informed that their network outages were directly related to power outages. Specifically, three organizations reportedly experienced power outages, two organizations had power-backup failures, and the other shut down the network due to the lack of power generators. Hence, the findings of root causes illustrate the dependence between network disruptions and power outages. Next, we compare how much this ground truth information from these seven organizations contributes to the all ground truth needed for all 282 unreachable subnets.

5.2 Estimation of Contributed Information Bits

Assume that a root cause of each group is considered to be an independent variable, and each independent variable requires one bit of information to disclose a root cause. We estimate the information bits provided by the ground truth reports and compare to information bits needed for total root causes. First, each unreachable subnet from these ground truth reports is examined whether it became unreachable independently or within a group. Since there are many unreachable subnets from Suddenlink, and their ground truth report did not

specify the subnet addresses, Suddenlink is excluded from the information bit estimation. Unlike Suddenlink, Texas Medical Center and John L Wortham & Son each has only one unreachable subnet.

We assume subnets that became unreachable within the same group shared the same root cause, and each organization knew the ground truth of its root cause. From Section 3.3, there are 109 groups of unreachable subnets.

Out of 109 groups belonging to all unreachable subnets, there are 51 and 16 organizations that respectively contained only isolated and within-organization dependent subnets. Hence, these are 67 organizations that need to be contacted the ground truth separately. The remaining cross-organization and cross-AS organizations had unreachable subnets in the same groups as other organizations. From our assumption, only one organization per group is needed to contact for the root cause. Therefore, we select the organizations that share the most number of groups first since such organizations could provide root causes for many groups. As a result, Comcast is first selected since it belonged to four out of 13 cross-organization and cross-AS dependent groups. The next selected organization is Kerrville Telephone Company since it shared one cross-organization and one cross-AS groups with other organizations. For the remaining seven cross-organization and cross-AS dependent groups, only one organization from each of these seven groups needs to be contacted for the group root cause. Therefore, nine organizations (Comcast, Kerrville Telephone Company, and the remaining seven) need to contacted for the root causes of 13 cross-organization and cross-AS dependent groups. In total, there are 76 organizations to be contacted (51 organizations for isolated subnets, 16 organizations for within-organization subnets, and nine organizations for cross-organization and cross-AS subnets). Since one bit of information is needed per a root cause, we need to contact 76 organizations to obtain 109 bits for all root causes.

The six ground truth reports generate 10 bits of information on the root causes: HARC, UTMB, Internet America, Texas Medical Center, and John L Wortham & Son each provided one bit of root cause. NASA provided five bits of root causes (four bits for the isolated subnets, and one bit for the cross-AS group). This amounts to 9.17% known information

bits that the root causes were either power outages or the lack of power generators. The majority of root causes (90.83%) are yet to be found.

5.3 Summary

Considerable prior work including our own has focused on assessing damage to the communications infrastructure due to natural or man-made disasters [18, 33, 42]. Two prior works have discussed that network routing infrastructure was impacted by power outages. Brown reportedly "expected" that number of network outages would increase after Hurricane Ike because of power outages [18]; however, no validation was further obtained to support his statement. In November 2009, it was reported that large power outages in Brazil disrupted more than 150 networks, including some in Paraguay and Uruguay [31]. Similarly, this study conjectured that power outages affected networks on a large-scale. Nonetheless, no detailed information was available to validate this conjecture.

Our findings show the dependence between network disruptions and power outages. Six out of seven ground truth reports in Section 5.1 have explicitly shown that power outages and the lack of power generators are ones of the causes for network-service disruptions. Therefore, the availability and the resilience of communications infrastructure can be enhanced not only from within but also from the supporting infrastructure such as power.

Limitations and open issues:

Two important issues stand out after ground truth information is received.

1. The first issue is that how to acquire power outage information and what network and power variables to consider and observe the dependence between network disruptions and power outages. We will propose the study of this dependence in Chapter 6.

2. The second issue is that information sharing is crucial but rooted deeply in privacy and security. The information bits contributed by the ground truth organizations comprise only 9.17% of total root causes. However, the information content given by the seven organizations revealed the significant impact of power infrastructure on communication networks. Therefore, more information sharing across organizations

is needed in studying network disruptions caused by natural disasters. In Chapter 7, we discuss more on the information sharing and its challenges.

CHAPTER VI

PRELIMINARY INVESTIGATION OF NETWORK DISRUPTIONS ON POWER-RESOURCE DEPENDENCE

Our searching for ground truth show that there exists dependence between network-service disruptions and power infrastructures. A number of questions arise: What information is available to study the dependence between communication and power infrastructures? How do network disruptions depend on power outages?

Due to the critical importance of the electricity infrastructure to national security [37], power data is extremely difficult to obtain. We examine the dependence between network disruptions and power outages using the publicly available and aggregated power outage data from the Public Utilities Commission of Texas[1].

Chapter 6 is organized as follows. Section 6.1 introduces the power data used in this study. Section 6.2 characterizes subnet unreachability and power data, and reports the observations between network and power outages. Section 6.3 computes the correlation coefficient between subnet unreachability and power outages. Section 6.4 provides summary.

6.1 Power Data

We start the analysis of dependence between network-service disruptions and power outages with the data from the Public Utilities Commission of Texas (PUC). PUC provides the reports of number of customers without electricity from 63 counties following Hurricane Ike, between September 13-20, 2008.

After the storm, PUC collected the daily restoration updates from the power service providers in the impacted area and aggregated number of customers without electricity. For example on September 14, 65,809 customers experienced power outages in Galveston county. Figure 26 presents the number of customers without electricity on September

[1]In the U.S., each state has its own public utility commission to monitor and ensure consumers safe and reliable utility service, and enforce regulations to utility providers.

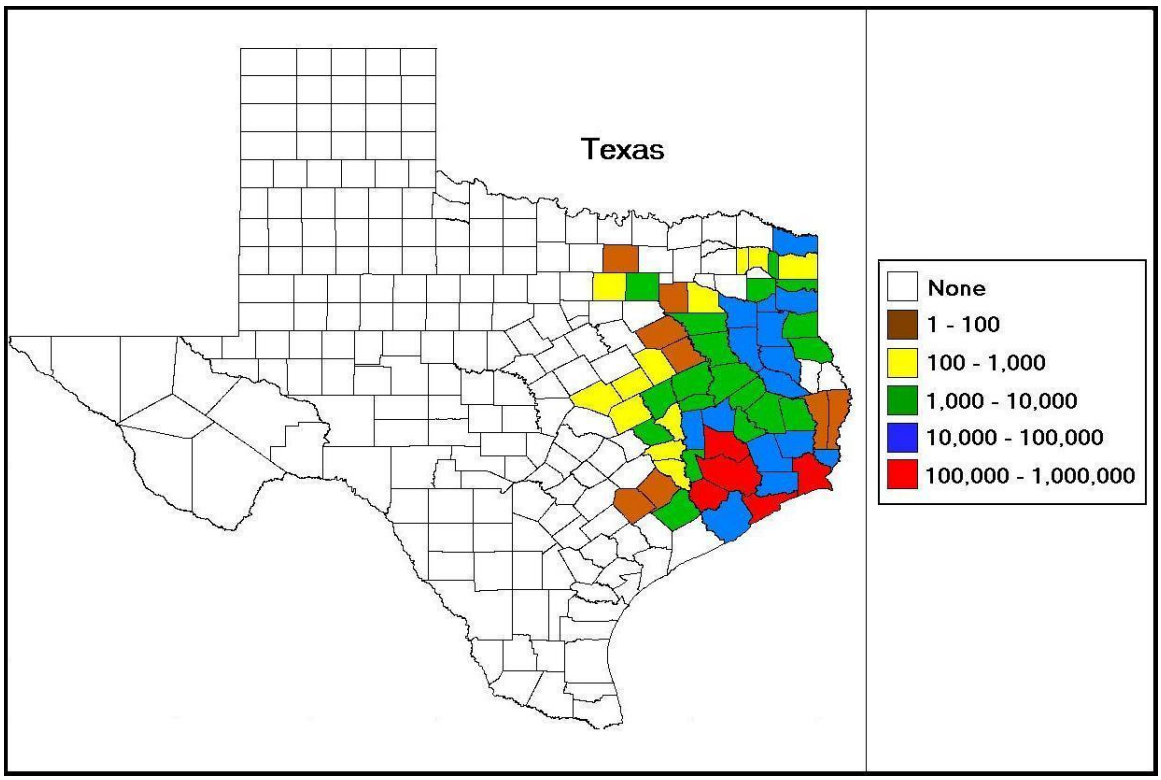


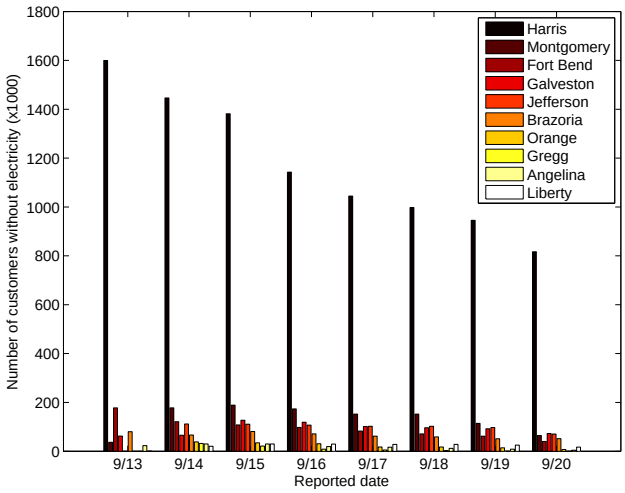
Figure 26: Number of customers without electricity reported by Public Utilities Commission of Texas on 9/14/08.

14. Figure 27(a) shows top 10 counties with the maximum number of customers without electricity between September 13-20, and Figure 27(c) displays the locations of these 10 counties.

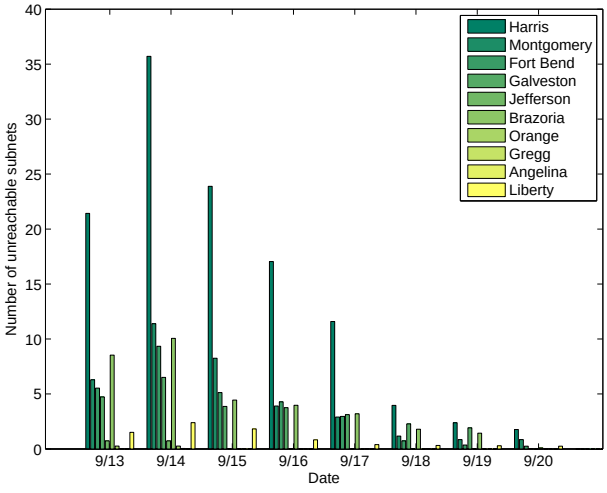
From Figure 27(a), Harris county had the maximum number of customers without electricity, i.e., about 1.6 millions on September 13. There are 39.68% of counties that had the maximum numbers of customers without power on September 14, where the number of customers without power increased from September 13 to 14, and decreased afterward. After a week, the power was not fully restored as eight counties still had more than 10,000 customers without electricity.

6.2 Characterizing Network and Power

The spatial and temporal scales of the PUC data are available at the levels of county and day respectively. On the other hand, subnet unreachability data is available at a spatial scale of 25-mile radius disks (from GeoIP) and time scale of seconds (from BGP updates).



(a) Numbers of customers without electricity.



(b) Numbers of unreachable subnets.

(c) Top 10 counties with the maximum number of customers without electricity.

Figure 27: Number of customers without electricity reported by Public Utilities Commission of Texas and unreachable subnets from the top 10 most impacted counties between 9/13/08-9/20/08.

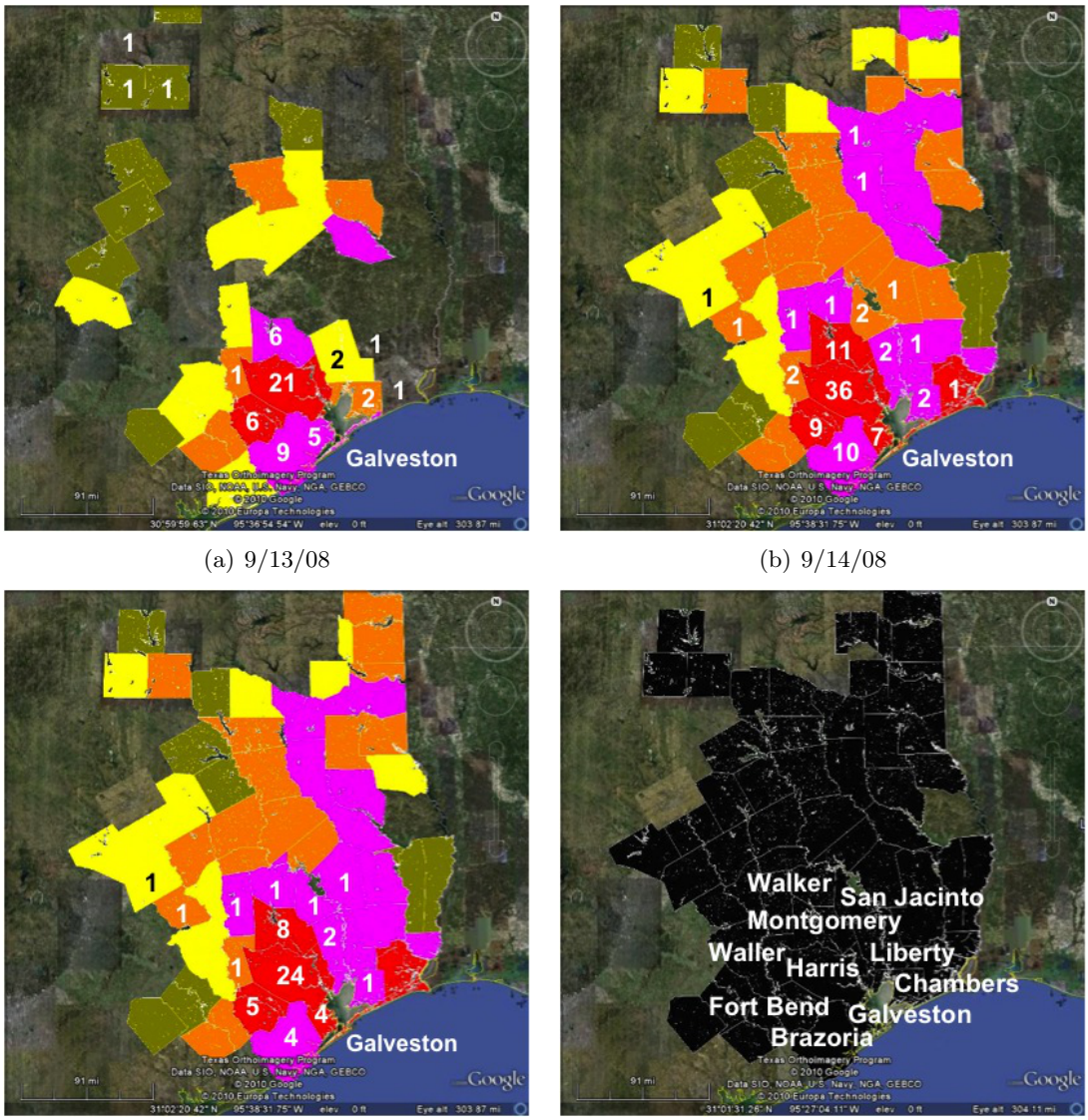
Hence, for a detailed study of dependence, the scale of the power outage data is clearly not comparable to the network data. Therefore, we use the PUC data to compare the observations between network and power outages and to analyze whether network disruptions occurred phenomenologically with the power outages. First, we aggregate the unreachable subnets to the same scale as the PUC data.

Spatial aggregation: In general, the sizes of most counties in Texas are smaller than the 25-mile radius subnet regions. Since one subnet region can cover several counties, we consider unreachable subnets in each county as fractions of subnet regions overlapping the county area. For example if 60% and 40% of a subnet region overlap counties A and B, then counties A and B respectively contain 0.6 and 0.4 unreachable subnets. As a result, 94.45 subnets are assigned to 44 counties from the PUC reports.

Temporal aggregation: Next, we aggregate subnet unreachability at the time scale of day. Ideally, the time for collecting unreachable subnet data should be the same as the power outage data. However, the collection times of power data are unknown. In this work, we simply aggregate subnet unreachability for each day by counting subnets that remained unreachable at 12:00 a.m. of that day. For example, to compare our subnet unreachability with the PUC report on September 13, we count number of subnets that remained unreachable on 12:00 a.m. September 13.

Figure 27(b) presents the number of unreachable subnets corresponding to the 10 counties in Figure 27(a). Similar to power outages, it is observed that Harris county had the maximum number of unreachable subnets. There are 77.27% of counties with the maximum number of unreachable subnets on September 14, where number of unreachable subnets increased from September 13 to 14, and decreased afterward. By September 20, two subnets had not resumed their connectivities.

Figures 28(a), 28(b), and 28(c) respectively present the number of unreachable subnets spatially displayed on the power outage maps based on the PUC data between September 13-15. The color assigned to each county corresponds to number of customers without electricity in that county on a specific day (For clear presentation, numbers of unreachable subnets are approximated to integers). Figure 28(d) displays the locations of top 10 counties



(a) 9/13/08

(b) 9/14/08

(c) 9/15/08

(d) Top 10 counties with the maximum unreachable subnets

Figure 28: Number of unreachable subnets displayed on PUC power outage maps at county level. For clear presentation, numbers of unreachable subnets are approximated to integers. (Number of customers without electricity: Red $> 100,000$, Pink $= 10,000-100,000$, Orange $= 1,000-10,000$, Yellow $= 100-1,000$, and Green $= 1-100$).

with the maximum number of unreachable subnets. Note that some of these 10 counties are different from the 10 counties with the maximum number of customers without electricity.

These Figures 28(a), 28(b), and 28(c) illustrate that counties with a large number of customers without electricity, i.e., Harris, Montgomery, Brazoria, and Galveston, also contain a large number of unreachable subnets. Furthermore, the majority of counties had the numbers of unreachable subnets increased from September 13 to 14, and decreased afterward.

Although we cannot use the PUC data to study the dependence between network disruptions and power outages in details, the observations from subnet unreachability and the PUC reports show that: (1) Harris county had the maximum numbers of unreachable subnets (34.05) and customers without electricity (1.6 millions on September 13). (2) Out of the eight days from the PUC reports, 77.27% and 39.68% of counties had maximum numbers of unreachable subnets and customers without electricity on September 14. Hence, we consider that network disruptions and power outages phenomenologically occurred with each other.

6.3 Network and Power Correlation

How are the numbers of unreachable subnets and customers without electricity correlated? Table 14 shows the number of customers without electricity in the top 10 counties with the maximum number of unreachable subnets.

We compute the correlation coefficient between the numbers of customers without electricity and unreachable subnets by $\rho(e,i) = \frac{1}{C}\sum_{c=1}^{C} \frac{E((N_{e,c}-\hat{E}[N_{e,c}])(N_{i,c}-\hat{E}[N_{i,c}])}{\hat{\sigma}_{N_{e,c}}\hat{\sigma}_{N_{i,c}}}$, where $N_{e,c}$ and $N_{i,c}$ respectively are the numbers of customers without electricity and unreachable subnets in county c, $1 \leq c \leq C$, and C is number of counties. $\hat{E}[\cdot]$ and $\hat{\sigma}^2$ are the sample mean and the sample variance. The resulting $\rho(e,i)$ is 0.9366. Hence, based on the PUC data, the numbers of customers without electricity and unreachable subnets are strongly correlated.

Using the PUC data, we are able to observe the phenomenological relation and report the correlation between network disruptions and power outages. To study the dependence between network and power outages in details, we need to seek power outage data with

Table 14: Numbers of customers without electricity in the top 10 counties with the maximum number of unreachable subnets.

County	Number of unreachable subnets	Number of customers without electricity
Harris	35.71	1,599,710
Montgomery	11.39	189,080
Brazoria	10.05	81,348
Fort Bend	9.34	177,792
Galveston	6.51	127,284
Liberty	2.39	30,107
Waller	2.09	6,352
Chambers	1.83	16,492
San Jacinto	1.58	13,832
Walker	1.30	24,096

more refined spatial and temporal scales.

6.4 Summary

In this part of research, we obtain the publicly available power data from the Public Utilities Commission of Texas. They are the reports of aggregated number of customers without electricity in the disaster area after Hurricane Ike. With the large temporal scale of power outage data, i.e., days, the dependence study between network disruptions and power outages cannot be performed in details. Furthermore, the spatial scale of both network and power data are large (25-mile radius disk and county). Thus, only the observations and the correlation between two aggregated data sources are reported.

Limitations and open issues: There are limitations to our power-resource dependence study of network-service disruptions caused by Hurricane Ike. Clearly, the limitations to study the dependence between network disruptions and power outages are the data itself. The network data is currently available at spatial scale of 25-mile radius disks, and the power data that is available in the scales of county and day. These scales are too large to consider two events are dependent or correlated. Power outage occurs at 10 miles from where the disrupted network is located, or power outage that occurs an hour after the network disruption are unlikely to be related. Hence, the network and power data at finer spatial and temporal scale is needed in our study.

This work illustrates the insufficiency of publicly available power data needed to study the dependence of network disruptions and power outages. This results in a spatial and temporal disparity between the available and the needed data for power and communication networks. An engineering issue here is how to obtain network and power measurements at a fine spatial and temporal scales. With the power infrastructure being important for national security, it is unlikely that more power data will be made publicly available by governments or power service providers. In the next Chapter 7, we discuss the information sharing that can potentially be used to improve the study of network-service disruptions caused by large-scale disturbances.

CHAPTER VII

INVESTIGATION OF INFORMATION SHARING

The nation's communications infrastructure is interdependent with organizations. Information on network disruptions has been proprietary to organizations and service providers. The social perception of network-service disruptions is negative for the organizations who own unreachable subnets; the users perceive network disruptions as a job poorly done by network administrators. Furthermore, privacy and confidentiality agreement allows little or no disclosure of information by service providers to the public. Thus, the information on disruptions is generally kept confidential and not exchanged among different organizations.

Aside from sharing network-disruption information, information sharing network has been used for data collection in many communities. In Section 7.1, existing information sharing networks are described. Section 7.2 discusses the use of information sharing for network disruptions and power outages, and Section 7.3 describes the ideal data for our study of the dependence between network-service disruptions and external factors such as weather and power resources. Section 7.4 provides the summary.

7.1 Existing Information Sharing Networks

Information sharing networks have been used to collect data in many communities. Below, we list existing information sharing networks in weather, power, and others.

7.1.1 Weather

Shared weather data can be obtained from National Climatic Data Center (NCDC) [65] and Community Collaborative Rain, Hail, and Storm network (CoCoRaHS) [29].

NCDC was established in 1950 as the Weather Records Center to archive the global and domestic weather data [65]. It currently serves as the World Data Center for Meteorology. In the U.S., the daily archived data comes from the National Weather Service, Military Services, Federal Aviation Administration, and the Coast Guard. With all the data acquired,

NCDC adds 228 gigabytes of new data to their records daily.

NCDC provides the global hourly surface data [51] that contains temperature, precipitation measurements (rain/snow/hail), wind and wave measurements, visibility, and automated and manual weather observations. NCDC collects data from all their stations across the U.S. Although NCDC is a government agency, the data is made available to public. Thus, NCDC can be considered as one of the information sharing data sources.

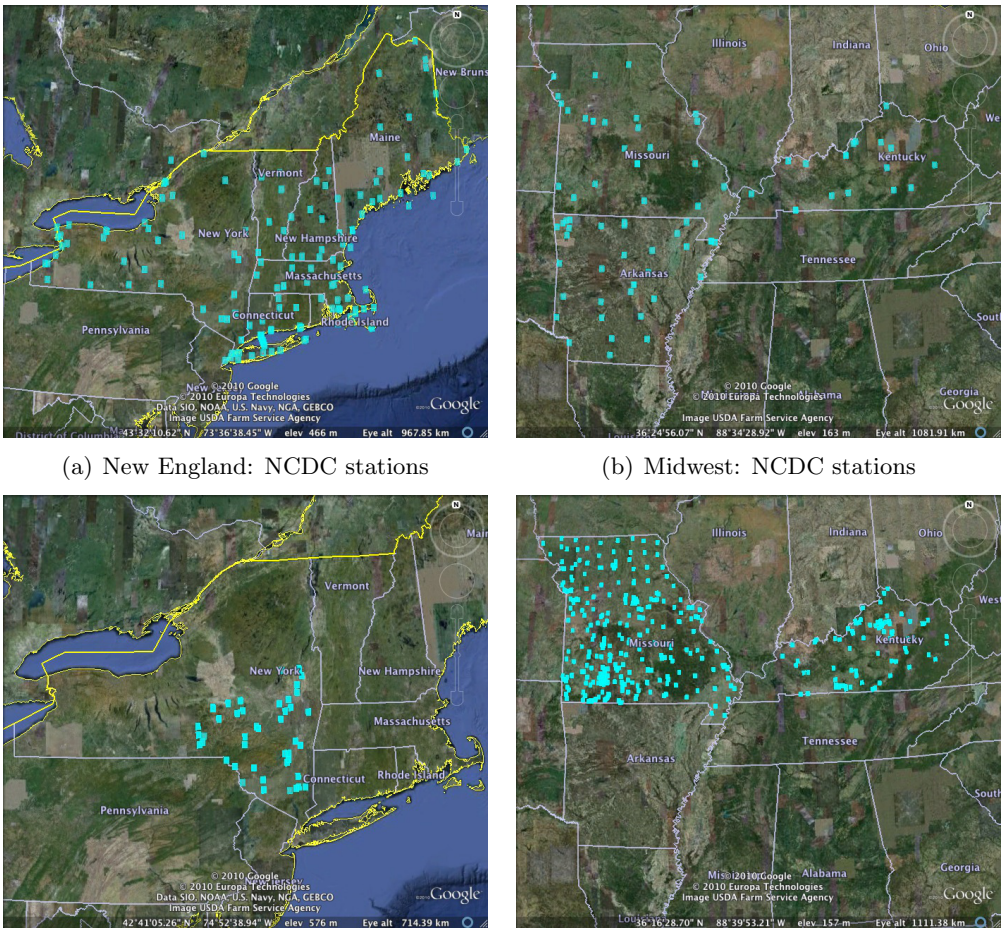
Here, we introduce two major natural disaster events that demonstrate the NCDC and CoCoRaHS data.

1. **The New England ice storm:** About 2-4 inches of freezing rainfall occurred during the night between December 11 and 12, 2008 [66]. More than one million of people lost power [2, 81], and Federal Emergency Management Agency (FEMA) issued the emergency declarations in four states: Maine, Massachusetts, New Hampshire, and New York [44]. Power was restored about 2-10 days afterward (with 5 days on average).

2. **The Midwest ice storm:** On January 27, 2009, approximately 1-6 inches of freezing rain occurred, and FEMA declared the emergency in Arkansas, Kentucky, and Missouri [45]. It was reported that more than one million people were without power [4, 81]. The restoration process took between 1-16 days, with an average of 6 days, to complete.

The NCDC data is spatially available in scale of latitude and longitude coordinates of NCDC stations. In general, NCDC stations are uniformly distributed in all the states as illustrated in Figures 29(a) and 29(b). Temporally, each NCDC station provides hourly reports of the measurements [51]. Combining NCDC measurements from stations over the area of interest results in a more refined timescale than hourly. This is demonstrated by NCDC temperature measurements shown in Figures 30(a) and 30(b).

CoCoRaHS, the second information sharing resource, is a non-profit and community-based organization whose participating volunteers collect the rain, hail, or storm measurements daily [29]. CoCoRaHS was founded in 1998 due to the Fort Collins flood in Colorado. While there was moderate rainfall in other parts of Fort Collins, the storm resulted in a



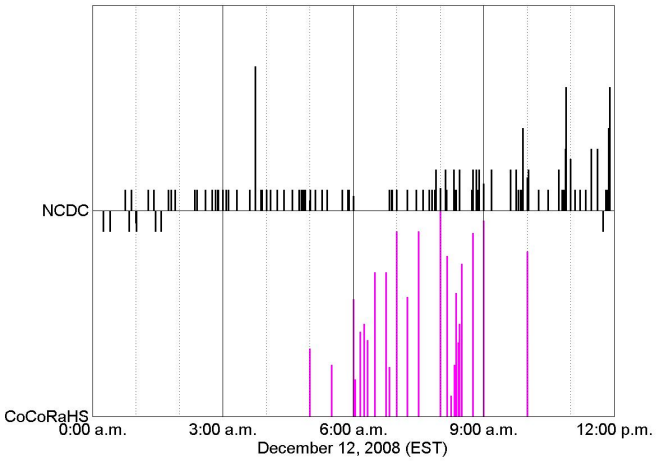
(a) New England: NCDC stations

(b) Midwest: NCDC stations

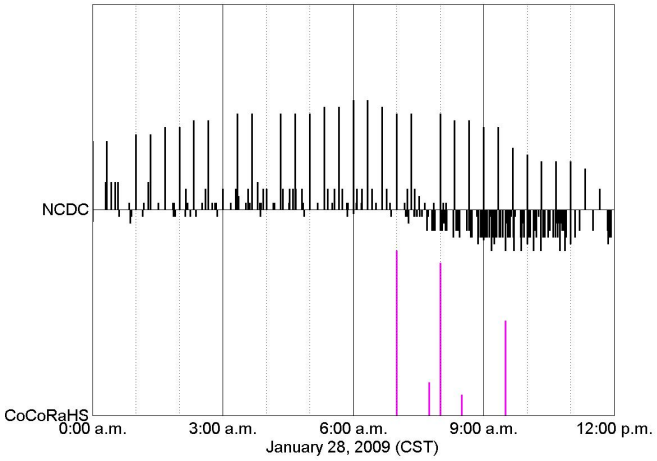
(c) New England: CoCoRaHS stations

(d) Midwest: CoCoRaHS stations

Figure 29: (a,b) NCDC and (c,d) CoCoRaHS stations in affected states of the New England and the Midwest ice storms.



(a) New England ice storm: New York state.



(b) Midwest ice storm: Kentucky state.

Figure 30: Timelines of NCDC temperature measurements and CoCoRaHS new snow measurements from (a) the New England, and (b) the Midwest ice storms.

Table 15: Number of NCDC and CoCoRaHS reports.

New England Ice Storm	Connecticut	Massa-chusetts	Maine	New Hampshire	New York	Vermont
Number of NCDC reports	513	2386	994	652	2435	343
Number of CoCoRaHS reports	0	0	0	0	191	0
Midwest Ice Storm	Arkansas	Kentucky	Missouri			
Number of NCDC reports	3638	2820	4191	-	-	-
Number of CoCoRaHS reports	0	268	1043	-	-	-

foot of localized rainfall, causing abrupt flooding that led to two deaths. This illustrates the example of non-uniform precipitation where local measurements can be useful for warnings if they are available.

The locations of CoCoRaHS measurements are latitude and longitude locations of their volunteers. Figures 29(c) and 29(d) respectively show the CoCoRaHS stations from the New England and the Midwest ice storms. In general, CoCoRaHS volunteers make the measurements between 7:00-9:00 a.m. daily and enter the data on the CoCoRaHS website. Figures 30(a) and 30(b) illustrate that most of CoCoRaHS new snow measurements are reported daily between 7:00-9:00 a.m. Hence, the CoCoRaHS data is spatially available in scale of latitude and longitude coordinates, and temporally available in scale of days.

Figure 29 shows the comparison between number of NCDC (29(a), 29(b)) and CoCoRaHS (29(c), 29(d)) stations. For the states with CoCoRaHS measurements, there were considerably more number of CoCoRaHS than NCDC stations. If measurements similar to CoCoRaHS could be collected more frequently, more complete weather data would enhance the dependence study of network disruptions on weather.

7.1.2 Power

Shared power information can be obtained from I-Grid [53]. I-Grid is an online distributed power monitoring system; it uses sensors to send signals through the Internet to the centralized server to update the active status of the sensors. Some of I-Grid reports are allowed

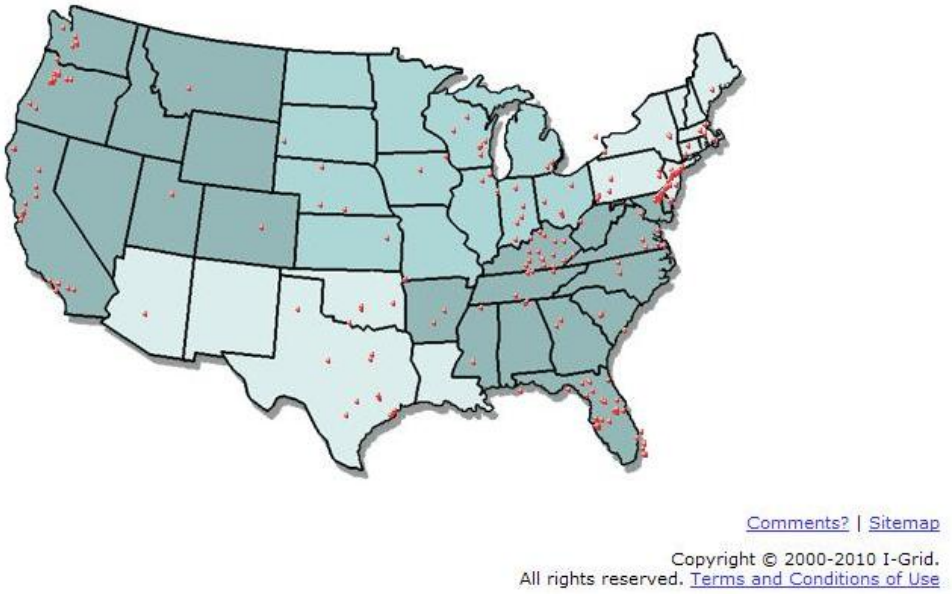


Figure 31: I-Grid sensors (source: www.igrid.com).

for public access. Although most data is available for members who own I-Grid monitoring sensors, I-Grid demonstrates how power statuses can be monitored distributively.

I-Grid reports themselves contain initial times and durations of power outages that the sensors experience. I-Grid reports are spatially available at the state level where the I-Grid sensors are located. As shown in Figure 31, the I-Grid sensors are not uniformly distributed. As an example of I-Grid reports, at 3:15:11 a.m. December 12, 2008, one sensor in New York sustained power outages for one day and seven hours. Thus, the time scale of I-Grid data is in seconds. Table 16 presents the number of I-Grid reports available in different states.

There were 10 I-Grid reports for the New England ice storm between December 11-12, 2008, and 51 reports for the Midwest ice storm between January 27-28, 2009. Hence, the I-Grid reports are (a) insufficient in spatial scale (state level), (b) refined in temporal scale (in seconds), and (c) available in small quantity.

According to the power data in this study, the major cause that results in spatial and temporal insufficiency of power data is the small numbers of available reports. With the national security concern related to the power data, it is rare to have power-outage data largely available to the public.

Table 16: Number of I-Grid reports.

	Connecticut	Massa-chusetts	Maine	New Hampshire	New York	Vermont
New England ice storm	0	4	1	0	5	0
	Arkansas	Kentucky	Missouri			
Midwest ice storm	1	49	1	-	-	-

7.1.3 Other Communities

Below, we list information sharing networks for other communities.

Internet Service Providers

There are a number of information sharing networks among ISPs. First is the mailing list of North American Network Operators' Group (NANOG). NANOG is a forum for backboned Internet service providers (tiers-1 and 2) that discuss technical issues and exchange insight knowledge [68]. Recently, the public were showned how powerful such a forum can be when NANOG participants worked together and figured out the source of the global network outage in February, 2008 [47].

Cooperative Association for Internet Data Analysis (CAIDA) is a collaboration between government and industries set up to enhance the macroscopic Internet infrastructure by analyzing network traffic and developing technologies to improve the Internet performance [30].

Daily Lives

Social networks prove to be popular and effective among people because the networks can convey messages rapidly and conveniently. Currently, there are more than 100 social networks and hundred millions of social network users worldwide. Examples of well-known social networks are Facebook, MySpace, Twitter, Orkut, and LinkedIn.

Social networks can be used to spread the news and establish a group of people who share the same interest [3, 70, 77]. They can also be used for advertisement and businesses [64, 76, 89]. In April 2009, AT&T used Twitter to notify their customers of network outages caused by fiber cuts as well as updates of the recovery process [71]. Some power service

providers also use Twitter to report and update their customers on the restoration of their network outage [50, 69]. In September 2009, social networks were used among locals to monitor flood situation [41].

Emergencies

Currently, there are several applications developed for information sharing during emergencies. Inspired by Hurricane Katrina in 2005, Microsoft developed Vine [82], so it can be used as a communications tool between family, friends, and communities in an emergency. Vine allows users to stay informed with news and weather reports from National Oceanic and Atmospheric Administration (NOAA), send or receive alerts, and report their statuses and situation to their family and friends. The idea of Vine is similar to the existing social networks but focus more on the use in emergency, especially reaching terms of necessary information.

The Department of Homeland Security also has an information sharing network among Federal, State, local governments and private sector partners [38]. The Homeland Security Information Network (HSIN) contains all emergency response plans and risk assessment. HSIN shows an example of how interdependency information can be shared among different emergency responders, e.g., law enforcement, critical infrastructure resources. The information must be shared in timely-fashion and multi-direction accessible for all agencies.

Other applications for emergency use are GeoChat and Mesh4X from InSTEDD [54]. InSTEDD is a collaboration among governments, industries, international and non-government organizations, and local communities to enhance emergency communications. Their goal is to allow all emergency responders to be able to communicate no matters what types of equipment and applications they use, and to fuse and share useful information such as news and disaster maps among different organizations.

Information sharing networks have proven to be effective; however, there are issues to be addressed for the use of sharing information on network disruptions or power outages.

7.2 Sharing Disruption Information

After a disaster, information can be shared among disaster responders or network administrators in the impacted region, to report their network or power outage statuses. Focusing on network administrator community, the information sharing network should consider the followings:

Identity: Identity of users in this information sharing networks should be open and can be used as a verification source to the public. Name of the users should be enlisted with their affiliation. Despite the negative reputation for organizations with unreachable subnets, users should be reminded that infrastructure damage due to natural disasters is involuntary. Hence, users should agree in disclosing their identities.

Pool of knowledge and timely update: Not only can users share the causes of their service disruptions and responses, they can also discuss their emergency preparedness plan, such as exchanging the knowledge on what types of equipment or fuel used, as well as seeking advice from one other. Furthermore, the information sharing network can be used as a status update where users can report their current situation , e.g., "power is back" or "the Internet is back." This would allow users to keep up with the updates from the others who may also be in the same neighborhood, instead of receiving updates only from their service providers.

Ease of cooperation: When many network administrators come together to share the disruption information, it draws attention to the public. This information sharing network may serve as another platform for network administrators to ask for assistance and cooperation from government or service providers. At the same time, the information sharing network would be a good resource for government or service providers to keep track of or make public announcements to network administrators.

7.3 Ideal Data

In this research, we face the challenges of collecting and using three types of data: network, weather, and power. Here, we propose the ideal data that would be the best suit for our study of the dependence between network disruptions, weather, and power outages. The

followings describe the characteristics of our ideal data:

1. Refined spatial and temporal scales: The refined scale of data helps with the accuracy of our dependence study. Currently, the weather data is available in the scale of latitude and longitude. However, the network and power data in the spatial scale of 25-mile radius disks and counties are inadequately comparable. Likewise, the network data is currently available on a time scale of seconds whereas the weather and the power data are available at the level of hours and days respectively. Thus, the network, weather, and power data should ideally be on a more spatially refined (latitude and longitude coordinates) and temporally refined (seconds) scales.

2. Coexistence in space and time: Network disruptions, weather measurements, and power data are expected to take place in the "same" physical region and occur "close" in time. To complete our dependence study, we rely on network, weather, and power data from the same event. Thus, "same" and "close" can be corresponded to refined scale such as 5-mile region or 15-minute interval. Power outage occurs at 10 miles from where the disrupted network is located, or power outage that occurs an hour after the network disruption occurs are unlikely to be related. Thus, the most suitable data for our dependency study needs to be the measurements that occurred in the same space and time.

3. Sufficient quantity: Not only should the data be refined and coexisting in space and time, but many of such measurements would also be needed. When the size of the data set is small, this raises the question of the validity of our dependence study. However, when the natural disaster occurs, the measurements may not be available due to the evacuation. This can be shown by unavailable NCDC measurements from some coastal stations immediately before the landfall of Hurricane Ike. Thus, this poses the challenges of how to obtain all three sufficiently large data sets of network, weather, and power.

Our discussion of ideal data lead to the following open issue: how to obtain more of ideal network, weather, and power data. The intent of our study is to analyze network

behaviors after natural disasters with respect to logical network, weather, and power resources. Better scale and greater availability of data would assist our dependence study of network disruptions, weather, and power outages and help provide better understanding of network-service disruptions caused by large-scale disturbances.

7.4 Summary

In this Chapter 7, we describe existing information sharing networks used in collecting weather, i.e., NCDC and CoCoRaHS, and power data, i.e., I-Grid. The information sharing networks used in other communities such as ISPs, daily lives, or emergencies are also provided. In addition, we discuss the use of information sharing in network disruptions or power outages. Last, we describe the ideal data that would be most suitable for our dependence study of network-service disruptions caused by large-scale disturbances.

In the next Chapter 8, we present the contributions and the future directions of this research.

CHAPTER VIII

CONCLUSION AND FUTURE RESEARCH DIRECTIONS

The contributions of this research and the future research directions are discussed as follows.

8.1 Research Contributions

The contributions of this thesis include the empirical study of network-service disruptions caused by large-scale disturbances and the use of real and publicly available heterogeneous data and statistical learning. We incorporate network-, weather-, and power-related data into the analysis of network-service disruptions with respect to logical networks and external factors such as weather and power resources.

In Chapter 2, we develop an approach using unsupervised and semi-supervised learning with large-scale network measurement and user inputs. This research introduces the use of user inputs. Despite few are available, user inputs generally provide accurate network statuses. Thus, features of user inputs are included in the identification of unreachable subnets, along with network measurements, to obtain the classifiers for network statuses.

In Chapter 3, we introduce the organization variables in this network study and discover that networks became unreachable within organizations, cross organizations, and cross ASes. Thus, temporal dependence of network disruptions also lies within logical dependence, i.e., organizations and ASes.

In Chapter 4, by including the weather data, we discover the uncommon belief that subnet unreachability and storm are weakly correlated. This work demonstrates the new understanding that not all network disruptions are strongly dependent on the storm. It is found that networks could become unreachable before the hurricane hit the disaster area.

In Chapter 5, our ground truth searches reveal that network disruptions are dependent on power resource. Rather than speculation, this research explicitly shows and confirms the fact that network disruptions indeed depend on power resources.

In Chapter 6, we study the new, publicly available, and aggregated power outage data,

and the result shows that network disruptions and power outages are strongly correlated.

8.2 Future Research Directions

The future research directions include a detailed study of the dependence between network disruptions and power resources. This research studies the publicly available and aggregated power outage data. If more refined spatial and temporal scale of network and power data can be obtained, the detailed study can be pursued on the dependence of network disruptions on power outages. The power data can be studied to understand how power outages occurred. Moreover, the temporal and spatial correlation between network disruptions and power outages can be further analyzed. Last, the weather data can be included to provide the complete study of dependence among network disruptions and external factors such as weather and power.

APPENDIX A

ACCURACY OF WHOIS DATABASE

To verify organization names and geo-locations, we use the Internet search engines (e.g., Google, Yahoo) to look for present physical addresses of 50 organizations, and compare with those obtained from Whois. Our assumption is that the information online is currently updated. If the search finds that some organizations no longer exist, we continue the search to track down histories of these non-existing organizations, i.e., whether they have their organization name changed, or whether they have merged with another organization.

Among 50 organizations, the Internet-search result shows that 28 organizations (56%) is up to date and have the same names and addresses as Whois database. The other organizations either have different names or different addresses.

Four subnets have false addresses according to the residential-privacy protection reported in [15]. One organization information (OAO) cannot be found from the Internet, so we resolve its organization name by performing traceroute and find that it belongs to another organization (NASA). One subnet organization has its name misspelled (Novalink); however there are the other two unreachable subnets with almost similar organization name (Novolink) and with the same initial time of subnet unreachability.

Furthermore, we observe that organization names and addresses in Whois database are likely to remain the same over time (e.g., more than 10 years) although organizations themselves change their information several times. We suggest network administrators to update their information in Whois database more frequently.

APPENDIX B

INTERPOLATION OF STORM DATA

From the collected storm data, let C_s^t be storm composition at time t where the subscript s specified Hurricane Ike. Let $C_s^t = \{L_s^t, v_s^t, r_s^t\}$, where at time t, L_s^t=(x_s^t, y_s^t) is a latitude and longitude coordinate of a hurricane storm center, v_s^t is a storm speed, and r_s^t is wind radii of the hurricane force wind from four quadrants [NE, SE, SW, NW][1]. For example, at t = 10:00 p.m. September 12, $C_s^t = \{L_s^t$=(28.6, -94.4), v_s^t=12 miles per hour, r_s^t=[126.5, 103.5, 63.25, 86.25] miles$\}$.

Let $C_s^{t_1}$=$\{L_s^{t_1}$=$(x_s^{t_1}, y_s^{t_1})$, $v_s^{t_1}$, $r_s^{t_1}\}$ and $C_s^{t_2}$=$\{L_s^{t_2}$=$(x_s^{t_2}, y_s^{t_2})$, $v_s^{t_2}$, $r_s^{t_2}\}$ be the compositions of the storm at times t_1 and t_2 respectively, where $t_1 < t_2$. In our collected storm data, $t_2 - t_1$=1 hour. Assume that the storm moves in one-dimension for simplicity as illustrated in Figure 32. Let d^t be the distance that storm center travels between t_1 and t, where $t_1 \leq t < t_2$. Given storm speeds at $v_s^{t_1}$ and $v_s^{t_2}$, by the Newton's laws of motion, $d^t = v_s^{t_1}(t - t_1) + 0.5(t - t_1)(v_s^{t_2} - v_s^{t_1})$. Then, the latitude and longitude coordinate of storm center at time t, $L_s^{t_1}$=$(x_s^{t_1}, y_s^{t_1})$ is derived from [86]:

$$x_s^t = \arcsin(\sin(x_s^{t_1})\cos(d^t) + \cos(x_s^{t_2})\sin(d^t)\cos(ang(L_s^{t_1}, L_s^{t_2}))), \tag{10}$$

$$y_s^t = \left\lfloor((y_s^{t_1} - \arcsin(\sin(ang(L_s^{t_1}, L_s^{t_2}))\sin(d^t)/\cos(x_s^t))) + \pi)/2\pi \right\rfloor - \pi, \tag{11}$$

$$ang(L_s^{t_1}, L_s^{t_2}) = \arccos(\sin(x_s^{t_1})\sin(x_s^{t_2}) + \cos(x_s^{t_1})\cos(x_s^{t_2})\cos(y_s^{t_2} - y_s^{t_1})). \tag{12}$$

The interpolation is chosen to be at 15-minute interval and constrains to match the observation data of an hour. We observe that this constraint could be violated if an interpolation interval is too small, e.g., less than 10 minutes.

[1]Unit of wind radii that National Hurricane Center provides is nauticle mile (nmi). 1 nmi = 1.15 miles.

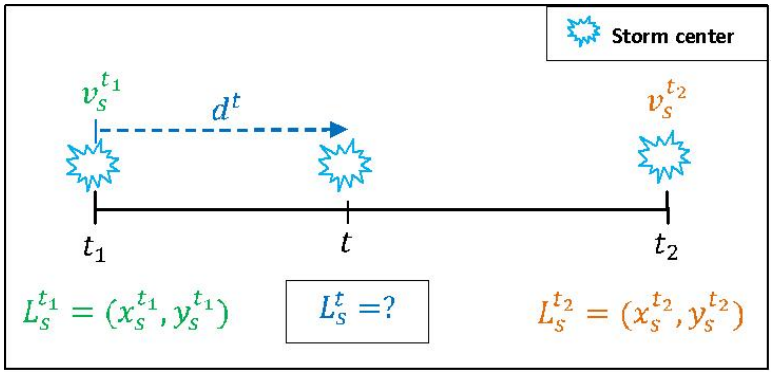


Figure 32: One-dimensional storm motion during times t_1, t, and t_2, where $t_1 \leq t < t_2$.

REFERENCES

[1] "California wildfires affect Internet service." http://www.satellitefamily.com/news-california-wildfires-affect-internet-service.asp.

[2] "Storm leaves at least 1 million without power in Northeast," Dec. 13, 2008. http://www.cnn.com/2008/US/weather/12/12/storm.new.york.massachusetts/index.html.

[3] "Tracking candidates on Twitter," Feb. 7, 2008. http://blog.twitter.com/2008/02/tracking-candidates-on-twitter.html\#links.

[4] "Kentucky ice storm: Nearly 1M still without power," Jan. 31, 2009. http://www.huffingtonpost.com/2009/01/31/kentucky-ice-storm-nearly_n_162777.html.

[5] "Haiti earthquake: Communications still a challenge," Jan. 18, 2010. http://www.von.com/news/2010/01/haiti-earthquake-communications-still-a-challenge.aspx.

[6] A DIGITAL RECORD OF THE COMPLETE BEST TRACK DATA ftp://ftp.nhc.noaa.gov/atcf/archive/2008/bal092008.dat.gz.

[7] ADVISORY ARCHIVE, H. http://www.nhc.noaa.gov/archive/2008/IKE.shtml.

[8] ANDERSEN, D., FEAMSTER, N., BAUER, S., and BALASKRISHMAN, H., "Topology inference from BGP routing dynamics," in *Proc. ACM SIGCOMM Workshop on Internet Measurements*, (Marseille, France), pp. 243–248, Nov. 2002.

[9] BAHL, P., CHANDRA, R., GREENBERG, A., KANDULA, S., MALTZ, D. A., and ZHANG, M., "Towards highly reliable enterprise network services via inference of multi-level dependencies," in *ACM SIGCOMM*, (Kyoto, Japan), Aug. 2007.

[10] BALCAN, M.-F., BLUM, A., CHOI, P. P., LAFFERTY, J., PANTANO, B., RWEBANGIRA, M. R., and ZHU, X., "Person identification in webcam images: an application of semi-supervised learning," in *International Conference on Machine Learning Workshop on Learning with Partially Classified Training Data*, (Bonn, Germany), 2005.

[11] BBC NEWS, "Asia communications hit by quake," Dec. 27, 2006. http://news.bbc.co.uk/2/hi/asia-pacific/6211451.stm.

[12] BERG, R., "Tropical Cyclone Report: Hurricane Ike (AL092008) 1-14 September 2008," Jan. 29, 2008.

[13] BEVEN-II, J. L., AVILA, L. A., BLAKE, E. S., BROWN, D. P., FRANKLIN, J. L., KNABB, R. D., PASCH, R. J., RHOME, J. R., and STEWART, S. R., "Annual summary-Atlantic hurricane season of 2005," Mar. 2008 http://www.buddeblog.com.au/frompaulsdesk/chiles-telecommunications-after-the-earthquake/.

[14] Bibolini, L., "Chile telecommunications after the earthquake," Mar. 1, 2010. http://www.buddeblog.com.au/frompaulsdesk/chiles-telecommunications-after-the-earthquake/.

[15] Bicknell, L., "Whois by the numbers." ARIN Public Policy Meeting, Oct. 2006 https://www.arin.net/participate/meetings/reports/ARIN_XVIII/PDF/thursday/WHOIS_Bicknell.pdf.

[16] Blum, A. and Chawls, S., "Learning from labeled and unlabeled data using graph mincuts," in *Proc. of International Conference on Machine Learning (ICML)*, (Williamstown, MA), pp. 19–26, June 2001.

[17] Blum, A. and Mitchell, T., "Combining labeled and unlabeled data with co-training," in *Proc. of Conference on Computational Learning Theory (COLT)*, (Madison, WI), pp. 92–100, 1998.

[18] Brown, M., "Ike hammers Texas Internet." Renesys Corporation, Sept. 2008.

[19] Brown, M., Popescu, A., Underwood, T., and Zmijewski, E., "Aftershocks from the Taiwan earthquake: Shaing up Internet transit in Asia." NANOG42, Feb. 2008.

[20] Burges, C. J. C., "A tutorial on support vector machines for pattern recognition," *Data Mining and Knowledge Discovery*, vol. 2, no. 2, pp. 121–167, 1998.

[21] Caesar, M., Subramanian, L., and Katz, R., "Towards localizing root causes of BGP dynamics," Tech. Rep. CSD-03-1292, Computer Science Department, University of California-Berkeley, Nov. 2003.

[22] Castelli, V. and Cover, T., "The relative value of labeled and unlabeled samples in pattern recognition with an unknown mixing parameter," *IEEE Transactions on Information Theory*, vol. 42, no. 6, pp. 2101–2117, 1996.

[23] Cebridge News http://support.cebridge.net/news.php.

[24] CenterPoint Energy, "CenterPoint Energy begins power restoration and damage assessment following Hurricane Ike - Crews restore power to 112,000 customers in the first 8 hours of recovery," Sept. 13, 2008. http://www.centerpointenergy.com/newsroom/newsreleases/e75964c19ed5c110VgnVCM1000005a1a0d0aRCRD/.

[25] CenterPoint Energy, "CenterPoint Energy responds to Hurricane Ike - Crews to assess damage, begin power restoration for 2.1 million customers," Sept. 13, 2008. http://www.centerpointenergy.com/newsroom/newsreleases/bab2d436c885c110VgnVCM1000005a1a0d0aRCRD/.

[26] Chang, D. F., Govindan, R., and Heidemann, J., "The temporal and topological characteristics of BGP path changes," in *Proc. of IEEE International Conference on Network Protocols (ICNP)*, (Atlanta, GA), pp. 190–199, Nov. 2003.

[27] Chapelle, O., Scholkopf, B., and Zien, A., *Semi-Supervised Learning*. MIT Press, 2006.

[28] Committee on the Internet Under Crisis Conditions: Learning from the Impact of September 11, *The Internet Under Crisis Conditions Learning from September 11*. The National Academics Press, 2003.

[29] Community Collarborative Rain, Hail, and Snow Network http://www.cocorahs.org.

[30] Cooperative Association for Internet Data Analysis http://www.caida.org/.

[31] Cowie, J., "Lights out in Rio." Renesys Corporation, Nov. 2009 http://www.renesys.com/blog/2009/11/lights-out-in-rio.shtml.

[32] Cowie, J., Popescu, A., and Underwood, T., "Impact of Hurricane Katrina on Internet infrastructure." Renesys Corporation, Sept. 2005 http://www.renesys.com/tech/presentations/pdf/Renesys-Katrina-Report-9sep2005.pdf.

[33] Cowie, J. H., Ogielski, A. T., and Premore, B., "Impact of the 2003 blackout on the Internet communications." Renesys Corporation, Nov. 2003 http://www.renesys.com/tech/presentations/pdf/Renesys_BlackoutReport.pdf.

[34] Curie, M. and Petty, J., "NASA's Johnson Space Center to reopen Monday after Ike," Sept. 18, 2008.

[35] Davies, D. L. and Bouldin, D., "A cluster separation measure," *IEEE Trans. Pattern Recognition and Machine Intelligence*, vol. 1, pp. 224–227, Apr. 1979.

[36] Demiriz, A. and Bennett, K., "Optimization approaches to semi-supervised learning," in *Applications and Algorithms of Complementarity* (Ferris, M. C., Mangasarian, O. L., and Pang, J. S., eds.), pp. 809–814, 1999.

[37] Department of Homeland Security, "National infrastructure protection plan."

[38] Department of Homeland Security, "National infrastructure protection plan partnering to enhance protection and resiliency." http://www.dhs.gov/xlibrary/assets/NIPP_Plan.pdf.

[39] Duda, R., Hart, P., and Stork, D., *Pattern Classification*. John Wiley & Sons, 2001.

[40] Dunn, J., "Well separated clusters and optimal fuzzy partitions," *J. Cybernetics*, vol. 4, pp. 95–104, 1974.

[41] Emerson, B., "Atlanta flood monitored through social media." The Atlanta Journal-Constitution, Sept. 22, 2009. http://www.ajc.com/news/atlanta-flood-monitored-through-144253.html.

[42] Erjongmanee, S., Ji, C., Stokely, J., and Hightower, N., "Large-scale of inference of network-service disruption upon natural disasters," in *Knowledge Discovery from Sensor Data* (Gaber, M., ed.), vol. 5840 of *Lecture Notes of Computer Science*, Springer, June 2010.

[43] Feamster, N., Andersen, D., and Balakrishnan, H., "Measuring the effects of internet path faults on reactive routing," Proc. ACM SIGMETRICS International Conference Measurements and Modeling of Computer, (San Diego, CA), pp. 133–139, June 2003.

[44] Federal Emergency Management Agency, "2008 Federal emergency declarations." http://www.fema.gov/news/disasters.fema?year=2008#sev2.

[45] Federal Emergency Management Agency, "2009 Federal emergency declarations." http://www.fema.gov/news/disasters.fema?year=2009#sev2.

[46] Feldmann, A., Maennel, O., Mao, Z. M., Bergerm, A., and Maggs, B., "Locating Internet routing instabilities," *ACM SIGCOMM Computer Communication Review*, vol. 34, pp. 205–218, Oct. 2004.

[47] Fildes, J., "Unsung heroes save net from chaos." BBC News, July 22, 2009. http://news.bbc.co.uk/2/hi/technology/8163190.stm.

[48] Gay, M., "Big storms are taking heavy toll on Midwest." The New York Times, Sept. 15, 2008. http://www.nytimes.com/2008/09/16/us/16midwest.html?_r=1&oref=slogin.

[49] GeoIP City http://www.maxmind.com/app/city.

[50] Georgia Power Twitter Page http://twitter.com/georgiapower.

[51] Global and U.S. integrated surface hourly data http://cdo.ncdc.noaa.gov/pls/plclimprod/poemain.accessrouter?datasetabbv=DS3505.

[52] Goldsein, H., "Engineers race to restore communications after Haiti quake," Jan. 19, 2010. http://spectrum.ieee.org/tech-talk/telecom/internet/engineers-race-to-restore-communications-after-haiti-quake.

[53] I-Grid http://www.igrid.com.

[54] InSTEDD: Innovative Support to Emergencies, Diseases, and Disasters http://instedd.org.

[55] Joachims, T., "Transductive inference for text classification using support vector machines," in *Proc. of International Conference on Machine Learning (ICML)*, (Bred, Slovenia), pp. 200–209, June 1999.

[56] Kandula, S., Katabi, D., and Vasseur, J. P., "Shrink: A tool for failure diagnosis in ip networks," in *ACM SIGCOMM Workshop on Mining Network Data (MineNet)*, (Philadelphia, PA), Aug. 2005.

[57] Kaufman, L. and Rousseeuw, P. J., *Finding Groups in Data: an Introduction to Cluster Analysis*. John Wiley & Sons, 1990.

[58] Labovitz, C., Ahuja, A., and Bose, A., "Delayed Internet routing convergence," *IEEE/ACM Transaction on Networking*, vol. 9, pp. 293–306, June 2001.

[59] Labovitz, C., Malan, G. R., and Jahanian, F., "Internet routing instability," *IEEE/ACM Transaction on Networking*, vol. 6, pp. 515–528, Oct. 1998.

[60] Lee, G. J. and Poole, L., "Diagnosis of TCP overlay connection failures using Bayesian networks," in *ACM SIGCOMM Workshop on Mining Network Data (MineNet)*, (Pisa, Italy), pp. 305–310, Sept. 2006.

[61] Li, J. and Chua, C. S., "Transductive inference for color-based particle filter tracking," in *Proc. of International Conference on Image Processing (ICIP)*, vol. 3, (Barcelona, Spain), pp. 949–952, Sept. 2003.

[62] mailing list archives August 2008, T. N. http://mailman.nanog.org/pipermail/nanog/2008-August/thread.html.

[63] Martin, K. J., "Written statement of Kevin J. Martin, Chairman Federal Communications Commission, at the hearing on public safety communications from 9/11 to Katrina: Critical public policy lessons, before the subcommittee on telecommunications and the Internet." House Committee on Energy and Commerce, U.S. House of Representatives, Sept. 2005.

[64] Miller, C. C., "Marketing small business with Twitter." NYTimes.com, July 22, 2009. http://www.nytimes.com/2009/07/23/business/smallbusiness/23twitter.html?_r=2&em=&adxnnl=1&adxnnlx=1248368448-71iG71CwCD1gS6JJkEslXA.

[65] National Climatic Data Center http://www.ncdc.noaa.gov.

[66] National Weather Service, "Public information statements, spotter reports," Dec. 12, 2008. http://www.erh.noaa.gov/aly/Past/2008/Dec_11-12_2008/PNS.txt.

[67] Nigam, K., *Using Unlabeled Data to Improve Text Classification*. Doctoral thesis, 2001.

[68] North American Network Operators' Group (NANOG) Mailing List Archive http://mailman.nanog.org/pipermail/nanog/.

[69] Pacific Power Oregon Twitter Page http://twitter.com/pacificpower_OR.

[70] Popkin, H., "Net big enough for Michael Jackson and Iran." MSNBC.com, June 26, 2009. http://www.msnbc.msn.com/id/31571885/ns/technology_and_science-tech_and_gadgets/wid/11915829/.

[71] Reardon, M., "AT&T uses Twitter during service outage," Apr. 9, 2009. http://news.cnet.com/8301-1035_3-10216712-94.html.

[72] Rekhter, Y., Li, T., and Hares, S., "A Border Gateway Protocol," *RFC 1771*, 1995.

[73] Rice, J. A., *Mathematical Statistics and Data Analysis*. Duxbury Press, 1995.

[74] RouteViews Project http://www.routeviews.org.

[75] Shahshahani, B. and Landgrebe, D., "The effect of unlabeled samples in reducing the small sample size problem and mitigating the Hughes phenomenon," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 32, pp. 1087–1095, Sept. 1994.

[76] Stay, J., "Facebook for business: Opportunities and limitations," July 28, 2008. http://www.insidefacebook.com/2008/07/28/facebook-for-business-what-it-needs-what-it-has/.

[77] Stone, B. and Cohen, N., "Social networks spread defiance online." NYTimes.com, June 15, 2009. http://www.nytimes.com/2009/06/16/world/middleeast/16media.html.

[78] The Select Bipartisan Committee to investigate the preparation for and response to Hurricane Katrina, "A failure of initiative: Final report of the Select Bipartisan Committee to investigate the preparation for and response to Hurricane Katrina," Tech. Rep. H. Rpt. 109-377, U.S. House of Representatives, 2005.

[79] Underwood, T. http://www.merit.edu/mail.archives/nanog/2005-08/msg00938.html.

[80] University of Texas Medical Branch at Galveston http://www.utmb.edu/ike/gallery/1.asp.

[81] U.S. Energy Information Administration, "Major disturbances and unusual occurrences." http://www.eia.doe.gov/cneaf/electricity/epm/appenb.pdf.

[82] Vine. Microsoft Research http://www.vine.net.

[83] Wang, L., Zhao, X., Pei, D., Bush, R., Massey, D., Mankin, A., Wu, S. F., and Zhang, L., "Observation and analysis of BGP behavior under stress," in *Proc. of ACM SIGCOMM Internet Measurement Workshop on Internet Measurements*, (Marseille, France), p. 183.

[84] Warrick, J., "Crisis communications remain flawed," Dec. 10, 2005. http://www.washingtonpost.com/wp-dyn/content/article/2005/12/09/AR2005120902039.html.

[85] Whois database http://www.arin.net/whois.

[86] Williams, E., "Aviation formulary v1.144." http://williams.best.vwh.net/avform.htm.

[87] Xu, K., Chandrashekar, J., and Zhang, Z. L., "A first step towards understanding inter-domain routing," in *Proc. of ACM SIGCOMM Workshop on Mining Network Data*, (Philadelphia, PA), Aug. 2005.

[88] Yarowsky, D., "Unsupervised word sense disambiguous rivaling supervised methods," in *Proc. of the Thirty-third Annual Meeting of the Association for Computational Linguistics (ACL)*, (College Park, MD), pp. 189–196, June 1995.

[89] Zeidler, S., "Looking for a job? try LinkedIn or Twitter." Reuter, Aug. 12, 2009. http://www.reuters.com/article/newsOne/idUSTRE57B2EX20090812.

[90] Zhang, J., Rexford, J., and Feigenbaum, J., "Learning-based anomaly detection in BGP updates," Tech. Rep. YALEU/DCS/TR-1318, Yale University, Apr. 2005.

[91] Zhu, X., "Semi-supervised learning literature survey," Tech. Rep. Computer Sciences TR 1530, University of Wisconsin-Madison, July 2008.

VITA

Supaporn Erjongmanee was born in Bangkok, Thailand. She received the B.S. degree in Electrical and Computer Engineering from Carnegie Mellon University, Pittsburgh, in 2001, and the M.S. degree in Electrical and Computer Engineering from Georgia Institute of Technology, Atlanta, in 2003. Currently, she is working toward the Ph.D. degree at the School of Electrical and Computer Engineering, Georgia Institute of Technology, under the supervision of Dr. Chuanyi Ji. Her graduate research focuses on study of network-service disruptions caused by large-scale disturbances, using heterogeneous data and statistical learning.