Large-Scale Network-Service Disruption: Dependencies and External Factors

Supaporn Erjongmanee and Chuanyi Ji

***Abstract*—Large-scale service disruptions in communication have been observed in the past but are not well-understood. The goal of this work is to gain a better understanding of disruptions in communication services in response to large-scale external disturbances such as hurricanes. In particular, Hurricane Ike is drawn as a case study, and heterogeneous data is obtained from networks, storm, and system administrators. Using the data, we first study network-wide disruptions and dependences among different unreachable subnets. Our findings show that 120 out of 230 subnets in our data set were unreachable, among which 88 subnets became unreachable dependently at a time scale of seconds or less than three minutes. We then study dependencies between communication service-disruptions and external factors such as weather and power. Unreachable subnets are found to be weakly correlated with the storm. Power outages and lack of spare power are reported to be certain causes of communication disruptions. New research issues emerge for information acquisition across communication and power infrastructures as well as weather, and information sharing among organizations.**

***Index Terms*—Internet application, human information processing, heterogeneous databases.**

I. Introduction

MAINTAINING network availability is essential yet challenging since communication infrastructure is not stand-alone but exposed to external disturbances such as natural disasters, and supported by resource networks such as power. Hence, complex dependencies exist among communication networks, external disturbances, and power systems.

For example, the blackout in the northeastern United States and Canada in 2003 [1] showed that the Border Gateway Protocol (BGP) routing infrastructure shivered in accordance with power outages. Severe weather in Brazil in 2009 caused large-scale power outages which resulted in subsequent spikes in BGP routing messages [2]. Those observations, however, were made at an aggregated scale of states and countries, and detailed study was not available at a smaller spatial scale. Other prior works have only focused on communication networks, i.e., disruptions due to natural or man-made disasters; nonetheless, dependencies to other entities have not been studied sufficiently [1]–[4]. Hence, large-scale and/or dependent disruptions remain as mystery. A better understanding of what happened at a more refined scale is needed for enhancing communication infrastructure.

The importance of studying large-scale and dependent disruptions lies in their impact and possible recovery schemes. For example, dependent disruptions of communication involving a large number of subnets are often perceived as more destructive than individually and randomly occurring unreachability [1]. Dependent disruptions from across difference types of networks are considered damaging to the critical infrastructures [1]. A better understanding of large-scale communication disruptions and their dependences to other factors also lays a foundation for network survivability [5] in face of disasters.

The objective of this work is to gain a better understanding on large-scale network-service disruptions, and their dependencies, through a case study on how communication networks respond to large-scale external disturbances such as a hurricane. Large-scale external disturbances push communication networks to the extreme, and thus expose disruptions and dependencies that may not be observable in day-to-day operation. Therefore, the first aspect essential for our understanding is on the scale and the dependencies of service disruptions for communication networks; for example, to what extent communication services were disrupted, and whether the communication disruptions occurred dependently among different networks. The second aspect is on the dependencies between communication networks and external entities such as weather and power; for example, how strongly communication service disruptions correlated with the hurricane. These aspects constitute network- and external-factors summarized below:

(a) Service disruptions and their dependencies in communication networks:
   - Subnet unreachabilities that correspond to unavailability of communication networks,
   - Affected organizations who own the subnets;

(b) Dependencies of communication disruptions with external factors:
   - Weather,
   - Power.

The problem we study here is to analyze communication disruptions and their dependencies, to study dependencies between communication disruptions and severe weather, and to seek root causes between communication disruptions and power outages directly from network administrators. The problem requires using heterogeneous data/information from networks, storm, and power. This provides an opportunity for us to understand what available data can and cannot do

Manuscript received January 7, 2011; revised May 7, 2011. The associate editor coordinating the review of this paper and approving it for publication was K. Van der Merwe.

S. Erjongmanee is with the Computer Engineering Department, Kasetsart University, Bangkok, 10900, Thailand (e-mail: fengspe@ku.ac.th).

C. Ji is with the Department of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA, 30332 USA (email: jic@ece.gatech.edu).

Digital Object Identifier 10.1109/TNSM.2011.201100106.

in revealing the complex dependencies. We shall focus on disruptions relating to Hurricane Ike as a case study and draw inference from heterogeneous measurements.

Ike, the strongest hurricane in 2008, made landfall at Galveston, Texas on September 13 [6], causing strong winds, flooding, and heavy rains. Large-scale network disruptions were reported and suggested that some disruptions might be caused by power failures [3]. Nonetheless, no detailed study is provided.

We begin studying communication disruptions and dependencies use publicly available measurements on networks and organizations in Section II. Communication disruptions are signified as unreachability of subnets and remotely inferred using BGP measurements. We identify and represent temporal and network dependence of disruptions in the logical space of subnets and organizations. We then study in Section III whether network disruptions occurred dependently with the hurricane, using storm data and subnet geo-locations obtained from National Hurricane Center and GeoIP City respectively. However, we soon encounter a difficulty that geo-locations of subnets are either inaccurate or unavailable, especially at a fine spatial scale of a few miles. We introduce a probability measure to incorporate uncertainty in subnet geo-location data. We then correlate unreachable time and storm-hitting time of subnets. The correlation shows that subnet unreachability is weakly dependent on the hurricane. A surprising observation emerges from our study that some subnets became unreachable hours before the hurricane arrived at the area. These findings show that there exist hidden causes that were not obtainable from publicly available data and could be unfolded from studying root causes.

We seek root causes directly from organizations who own subnets in Section IV. Three organizations revealed that power outages were the cause for their unreachable networks. The other organization turned off their network prior to the storm due to lack of spare power. Hence, power was a hidden variable that was unobservable from public data. This raises further research issues on dependencies between communication and power networks, which would require new types of data and sharing of root-cause information among organizations. We discuss, in Section V, research issues for information acquisition and information sharing to raise awareness among interdisciplinary communities such as communication networks, power, and meteorology.

II. Network Disruption and Dependency

Our study begins with network service disruptions and their dependencies during the hurricane. We use publicly available network measurements and organization data.

A. Network Unreachability

We consider network-service disruptions as unreachable subnets due to an external disturbance, e.g., a hurricane [3], [4]. When a subnet becomes unreachable, BGP routers withdraw routes to the subnet by sending multiple withdrawals to each other [7]. When the subnet becomes reachable, multiple announcements are sent among BGP routers to establish new routes. Thus, BGP update messages (withdrawals and announcements) have been used to remotely infer subnet unreachabilities [4], [8], [9].

We collect BGP update messages belonging to 230 subnets in Texas area from Route Views, where the monitoring point is at University of Oregon [10], during September 10-20, 2008. This period is from the announcement of evacuation to one week after a landfall of Hurricane Ike. We also collect BGP updates between August 1-September 9, to provide a baseline of subnet unreachability during normal operations. The update messages form 230-dimensional time series and result in 27 megabytes of data. Unreachable subnets are identified in the following two steps:

- A hierarchical clustering algorithm [11] is applied to reduce the 230-dimensional time series to a small number of clusters, where each cluster contains subnets with a similar pattern of BGP updates (see Appendix A for details).
- An unreachable status of a representative subnet from each cluster is inferred from its BGP updates as follows:
  - Identifying bursts of update messages: A burst of BGP updates is considered as anomalous. A threshold value is applied so that if the inter-arrival time between two BGP updates is smaller than the threshold, these two BGP updates are grouped in the same burst. A threshold value is chosen if inter-arrival times of BGP updates are grouped based on their similarities (see Appendix A for details).
  - Identifying unreachable subnets and their unreachable durations: Within each burst, if all the prior connectivity from the BGP routers are withdrawn, the subnet is considered as unreachable, and the time instance when this unreachability first occurs is the “initial time" of subnet unreachability. The unreachable duration is the period between the initial time and the next BGP announcement. The identified unreachable subnets are then validated by confirming that they are not caused by artifacts, i.e., BGP resets using data provided in [12].

Figure 1 shows the inferred unreachable subnets in the chosen time duration. Among these unreachable subnets, 120 subnets became unreachable in an intense period from 10:00 a.m. September 12 to 10:00 a.m. September 13. We consider these 120 unreachable subnets as impacted by Hurricane Ike, and shall validate this assumption in Section II-D using data from normal operations.

Organization identity of a subnet is pertinent for understanding how different social entities were impacted by the hurricane. We obtain organizations who own subnets using Whois database [13]. The 120 unreachable subnets are found to belong to seven Internet service providers (ISPs), 33 individual organizations, and 38 Autonomous Systems (ASes). For example, the subnet with prefix (subnet address) 129.109.0.0/16 belongs to the University of Houston Medical Branch at Galveston and AS14373.

The time instances when subnets first become unreachable (initial times) bear particularly valuable information for understanding network disruptions due to the hurricane. Hence, the union of 120 (a) initial times, (b) prefixes, (c) organizations, and (d) ASes form the network data for this study.

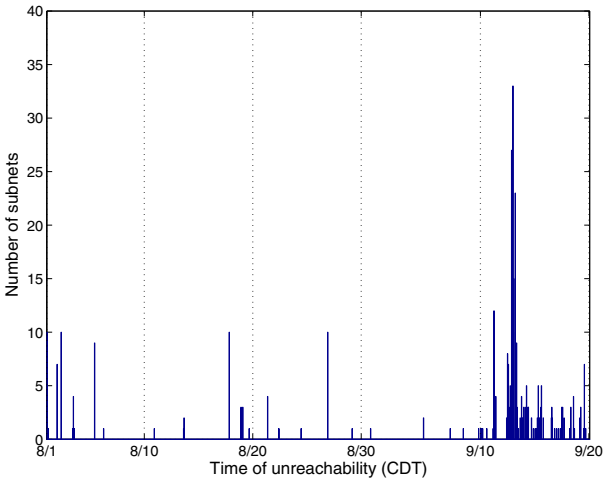


Fig. 1. Unreachable subnets between 8/1/08 - 9/20/08.

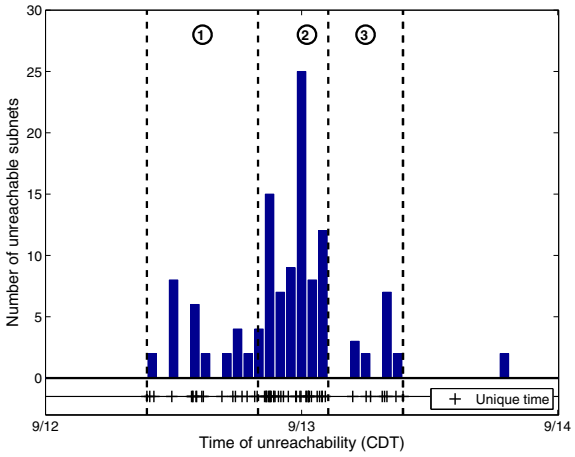


Fig. 2. Unreachable subnet groups with unique initial times between 9/12/08 - 9/14/08. Intervals 1-3 with different average inter-unreachable times.

B. Temporal Independence

We now analyze temporal dependencies, i.e., whether subnets became unreachable statistically dependently in time. From the 120 initial times, we observe that there are two types of unreachable subnets. One consists of subnets that became unreachable at a large time scale, i.e., tens of minutes or hours apart. The other type contains subnets that became unreachable in groups, i.e., at a small time scale, where the initial times of the unreachable subnets differed within seconds or less than three minutes.

We first group unreachable subnets that became unreachable in the 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. We obtain 56 groups, each with a unique initial time, presented at the bottom of Figure 2.

The unique initial times in Figure 2 occurred in the following three disjoint intervals, each of which has a different average inter-unreachable time: (1) Gradually from 10:16 a.m. September 12 to 8:25 p.m., (2) immensely between 8:25 p.m. September 12 and 3:00 a.m. September 13, and (3) less frequently from 3:00 a.m. to 10:00 a.m. September 13. These intervals are so chosen that they are sufficiently separated by gaps when no unreachability occurred, and exhibit sufficiently different average inter-unreachable times. The sample-average inter-unreachable times for the three durations are 37, 12, and 60 minutes respectively[1].

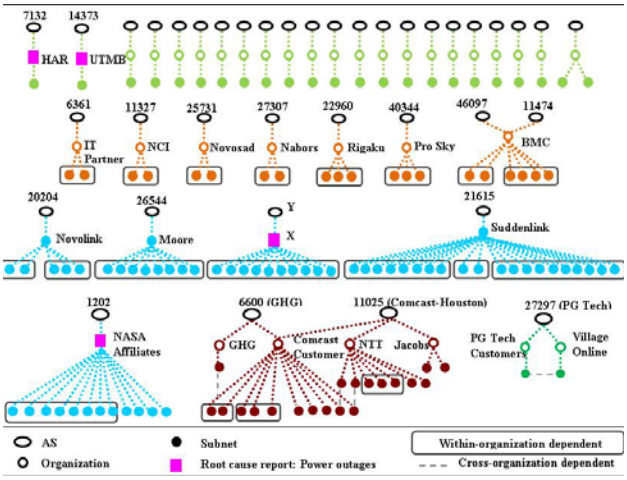


Fig. 3. Subnet-organization-AS hierarchy. X and Y are organization and AS of the anonymous ISP.

We pose a hypothesis (H_0) as follows. The unique time epochs occur statistically independently, with different average inter-unreachable times at different intervals, according to a non-uniform Poisson point process. Within each interval, the time epochs are assumed to exhibit the same arrival rate. A chi-square statistic is constructed within each interval by comparing the samples with the mean. The sum of independent chi-square statistics from the three intervals results in another chi-square statistic of degree 3. This allows us to use Pearson's chi-square test on the hypothesis H_0. H_0, if accepted, shows that these subnets became unreachable independently in time. Details of the test is given in Appendix B.

The data set $\{t_i\}_{i=1}^{56}$ is used to test H_0, where t_i is i^{th} unique initial time of subnet unreachability, $1 \leq i \leq 56$. Consider a confidence level 95%. Let $\chi^2_{0.05,3}$ be a threshold value for a chi-square statistic of degree 3 at confidence level 95%. $\chi^2_{0.05,3} = 7.82$ resulting from $\Pr(\chi^2 < \chi^2_{0.05,3}) = 0.95$, where $\Pr()$ is a chi-square distribution. The chi-square statistic obtained from the non-uniform Poisson point process and the data set is $\chi^2_3 = 1.71 < \chi^2_{0.05,3}$. Thus, the hypothesis cannot be rejected, and subnets became unreachable at distinct time epochs (statistically) independently.

C. Temporal and Network Dependence

We now examine subnets that became unreachable in groups, i.e., at the small time scale of less than three minutes, to understand the temporal and logical dependence of subnet unreachability. Figure 3 shows the dependence of unreachable subnets, organizations, and ASes in a hierarchy of the logical network space. There are 22 organizations, each of which has independent unreachable subnet(s) as shown in the first row of Figure 3. For the remaining subnets, there are 22

[1]Note that statistical methods, e.g., change-detection algorithms can be used to find time durations rigorously for more complex data sets.

TABLE I
EXAMPLE OF WITHDRAWAL BURSTS BELONGING TO TWO SUBNETS (A = BGP ANNOUNCEMENT, W = BGP WITHDRAWAL).

Type	BGP router	66.212.124.0/24 Arrival time (9/13)	66.212.127.0/24 Arrival time (9/13)
A	203.181.248.168	7:23:06 a.m. *ASPATH:* 7660 2516 1239 3356 297 270 1202	7:23:06 *ASPATH:* 7660 2516 1239 3356 297 270 1202
W	164.128.32.11	7:23:35 a.m.	7:23:09 a.m.
W	203.181.248.168	7:23:36 a.m.	7:23:36 a.m.
W	209.123.12.51	7:24:17 a.m.	7:25:18 a.m.

TABLE II
COMPARISON OF SUBNET UNREACHABILITY BETWEEN THE PRE-IKE AND IKE PERIODS.

	Pre-Ike	Ike
Number of unreachable subnet groups	0.65	56
Number of unique unreachable subnets	2.05	120
Number of within-organization dependent subnets	0.25	18
Number of cross-organization dependent subnets	0.03	4
Average inter-unreachable time	16.42 hours	0.41 hours
Maximum unreachability duration	0.11 days	7.20 days
Number of non-ISP organizations affected	0.18	33

groups, each of which contains two or more subnets that became unreachable within a small time scale of less than three minutes.

- Within-Organization Dependence

We examine each group to determine whether and which subnets from the same group became unreachable dependently. We begin by identifying unreachable subnets from the same organization.

Patterns in BGP withdrawal bursts of unreachable subnets provide basic information for the dependence. Table I 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 routers 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.

To determine whether two BGP withdrawal bursts exhibit similar pattern, we compute the correlation coefficient using inter-arrival times of BGP updates inside the bursts, from the same BGP routers and to the two subnets.

Let k be a BGP 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, m is number of BGP routers, and $37 \leq m \leq 42$ for our data set. Let T_{ik} and T_{jk} respectively be the inter-arrival times of BGP updates from BGP router k to subnets i and j. The correlation coefficient of the two burst patterns is $c_{ij} = \frac{1}{m}\sum_{k=1}^{m} \frac{(T_{ik}-\hat{E}[T_{ik}])(T_{jk}-\hat{E}[T_{jk}])}{\hat{\sigma}_{T_{ik}}\hat{\sigma}_{T_{jk}}}$, where $\hat{E}[\cdot]$ and $\hat{\sigma}^2$ are the sample mean and the sample variance. If the correlation coefficient is close to one, two BGP bursts have a similar pattern.

We find 18 groups whose correlation coefficients of BGP bursts are between 0.9150-1.000. These 18 groups correspond to 77 within-organization dependent subnets, 14 organizations, 14 ASes, and are shown in the last three rows of Figure 3. The organizations range from local businesses, e.g., I.T. Partners, NCI, to major ISPs, e.g., Novolink, Suddenlink. Besides BGP burst patterns, the majority of subnets in each group have the same ISP (84.58%) and exhibit the same unreachability duration (92.06%).

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

- Cross-Organization and Within AS Dependence

We study the remaining groups that contain subnets from different organizations but the same AS. As these subnets belong to the same AS, they exhibit the patterns of withdrawal bursts similar to those with within-organization dependence. The same criterion applies to determine the dependence of these unreachable subnets from their correlation coefficients.

There are four groups of cross-organization dependent subnets whose correlation coefficients of BGP bursts are between 0.9829-1.000. The subnets from these four groups belong to five organizations and three ASes as shown in the last row of Figure 3. All subnets in each group have the same ISP, and the same unreachability duration. The cross-organization dependence uniquely occurs at the local ISPs, for example, Comcast-Houston and PG Tech.

Our findings in this section show that AS is another logical variable that characterizes subnets from different organizations to become unreachable dependently. Combining the findings in the previous section, Figure 3 summarizes dependent and independent unreachable networks in a logical space of subnets, organizations, and ASes.

D. Comparison with Normal Operations

We validate our findings through comparing with subnet unreachabilities in the pre-Ike interval (August 1-September 9) when no major network disruptions were reported [14]. We separate the 40-day pre-Ike BGP update messages described in Section II-A into 40 one-day data sets. The average number of unreachable subnets is obtained per day, and then used as the baseline to compare with the unreachabilities from the Ike interval (September 10-20). We find that the one-day Ike period between 10:00 a.m. September 12-10:00 a.m. September 13 indeed presents anomalous subnet unreachabilities whereas the other days appeared to be normal. We exclude the data on

unreachable subnets after September 13 when the hurricane weakened to tropical storm and left the state of Texas.

Table II provides the comparison of unreachable subnets between the pre-Ike and the Ike periods. Compared to the pre-Ike duration, the unreachabilities occurred during Ike are 158.54 times more in quantity, at 40.05 times higher rate, and 7.09-day longer duration. Moreover, 183.33 times more unreachable subnets during Ike are from small organizations at the network edge, showing the impact of external disturbances.

Our results tell which subnets were unreachable from BGP routers where the unreachability corresponds to two scenarios. The first is when a subnet was disconnected from the rest of the network. The second is when 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 [15]. Therefore, BGP update messages can be used to infer reachability of prefixes but do not provide sufficient information to delineate these cases.

A significant number of unreachable subnets at the network edge illustrates, nevertheless, qualitative dependence between unreachable subnets and the hurricane, and motivates detailed study of storm correlation in the next section.

III. External Factor: Correlating with Storm

We now study how strongly subnet unreachability correlates with the storm. This requires storm and subnet geo-location data.

A. Storm Data

We obtain hourly observation data of Hurricane Ike from the public and forecast advisories [16], and the best track data [17], between 10:00 a.m. September 12 and 10:00 a.m. September 13. The storm data consists of latitude and longitude coordinates of storm center, storm speed, and wind radii of hurricane force winds of more than 74 miles per hour.

We reconstruct the storm path and the coverage using hurricane data, where 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 storm center and the coverage moved between 12-13 miles per hour [16]. As the subnet unreachabilities occurred at the scale of minutes, we interpolate the hourly storm data so that the reconstructed storm path and coverage are at the scale of 15 minutes (see Appendix C for detail).

B. Geo-Locations of Subnets

Geo-locations of subnets are needed to relate the subnet unreachability to the hurricane. However, accurate data is not yet available for geo-locations of IP addresses [18]. We adopt geo-location data provided by GeoIP City from Maxmind [19] since Maxmind specifies the accuracy of their measurements. Specifically, Maxmind provides latitude and longitude locations of IP addresses and reports that approximately 79% of their provided geo-locations are "correctly resolved within 25 miles from a true location" [19]. Such accuracy can be interpreted using probability. Consider an IP address i. Let X be a true but unknown geo-location and X_i be a geo-location provided by Maxmind. The probability $\Pr(|X - X_i| \leq 25)$ that X differs from X_i by at most 25 miles is approximately 0.79.

We obtain the geo-location of every IP-address for each prefix (e.g., 256 locations for /24 subnet) from GeoIP. To incorporate inaccurate information provided by GeoIP, each geo-location then becomes the center of a disk with 25-mile radius where a true location lies. The geo-location of a subnet is presented as a region from the union of geo-location disks belonging to all its IP-addresses. The true (unknown) geo-location of the BGP router can be anywhere within such a subnet region.

GeoIP geo-locations of our data set show that there are 91 subnets, each of which has all IP addresses at a unique latitude and longitude location. The remaining subnets have 2, 3, 4, 10, 15, 18, and 30 latitude and longitude locations per subnet. As a result, we obtain the 120 subnet regions.

C. Relating Network Unrechability and Storm

We now use the reconstructed storm coverage and the subnet geo-locations to study how subnet unreachability relates to the storm. In particular, we seek the time when the storm coverage appeared in subnet regions.

As the geo-location data is inaccurate, we define a coverage probability as follows to infer when a subnet was in the storm coverage. Consider subnet i with region S_i, $1 \leq i \leq 120$. Let R_t be the storm coverage at time t that consists of a hurricane center and wind radii, for $t \in$ [10:00 a.m. September 12, 10:00 a.m. September 13]. As the storm moves, the fraction of storm coverage R_t appears in region S_i. If an entire subnet region is in the storm coverage, the subnet is surely in the storm coverage. In contrast, if there is no overlapping between the storm coverage and a subnet region, the subnet is not in the storm coverage. Hence, the amount of overlap between the storm coverage and the subnet region characterizes the likelihood for the actual subnet location to be in the storm coverage. Thus, we define $p_{(i,t)} = \frac{|S_i \bigcap R_t|}{|S_i|}$ to be the coverage probability that the storm appears at the location of subnet i at time t, where $|\cdot|$ denotes the area. Coverage probability indicates how likely for the hurricane to impact subnet unreachability directly.

Figure 4 shows the color-coded values of the coverage probability for 120 unreachable subnets. Each subnet corresponds to one horizontal line in time. The coverage probability starts to increase at 6:15 p.m. September 12 until $p_{(i,t)}$ reaches a maximum value, and then decreases as the storm moves out of subnet region S_i, $1 \leq i \leq 120$. This shows spatially and temporally how the hurricane passed through subnets. The majority of subnets (73%) were completely inside the storm coverage between 9:30 p.m. September 12 and 6:15 a.m. September 13. All subnets were out of the storm coverage by 9:45 a.m. September 13. There are 18 subnets that were never covered by the storm.

We define the "hitting time" when the storm coverage first overlaps with a subnet region. This corresponds to the time when the coverage 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 part of Figure 4.

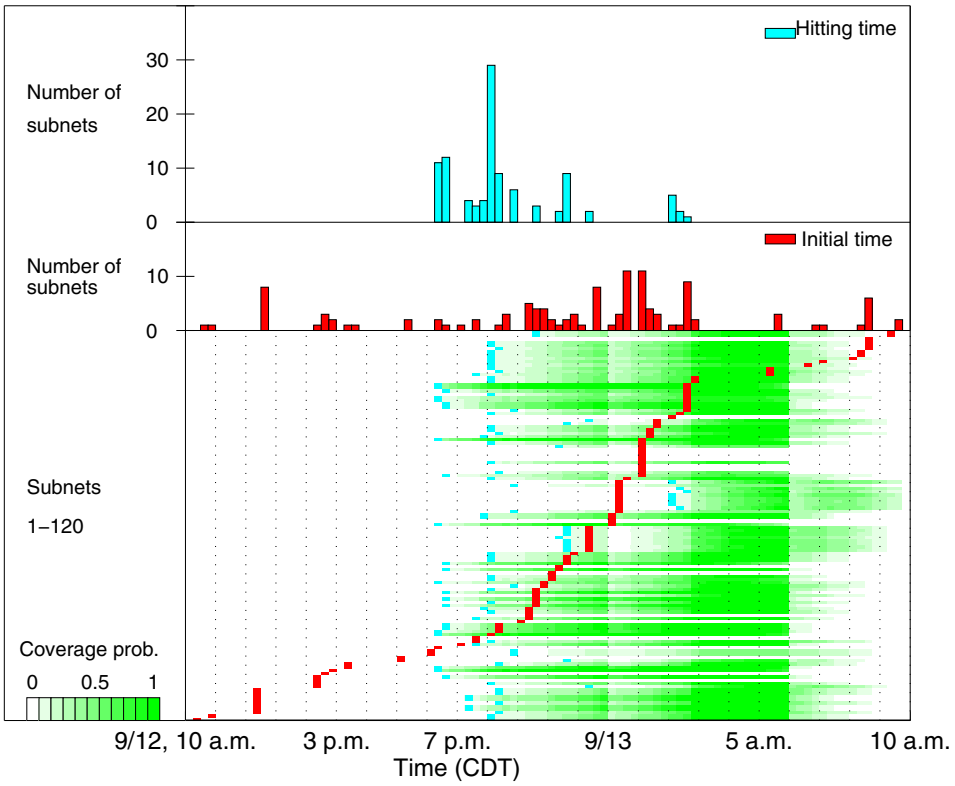
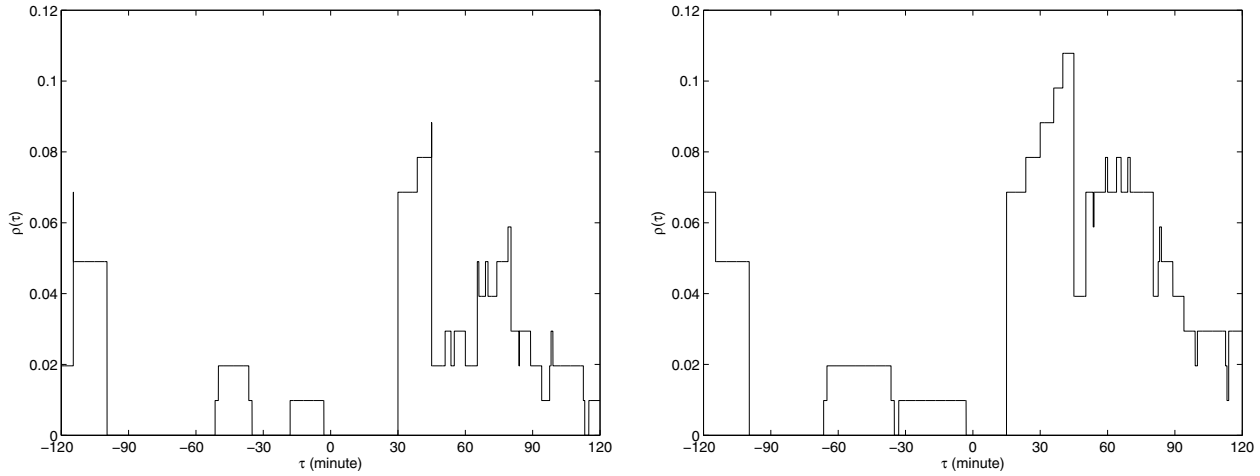


Fig. 4. Distributions of initial times, hitting times, and coverage probabilities.



(a) 15 minutes

(b) 30 minutes

Fig. 5. Sample correlation coefficients for T = 15 and 30 minutes.

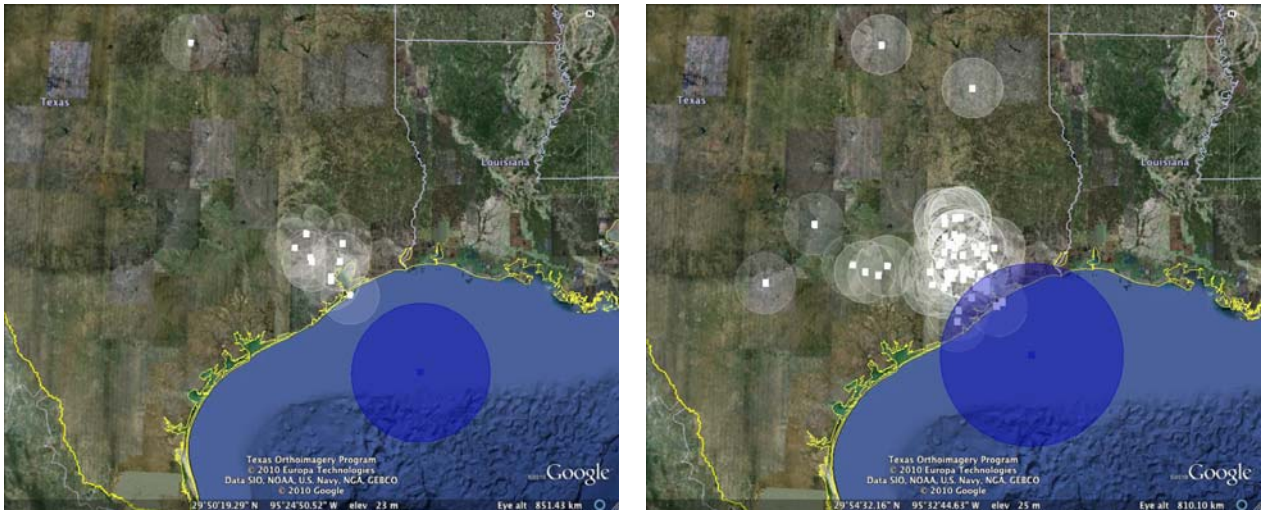
The hitting times of 102 subnets[2] occurred between 6:15 p.m. September 12 and 2:45 a.m. September 13 with the maximum at 8:00 p.m. September 12.

The empirical distribution of the initial times when the subnets first became unreachable is shown in the middle of Figure 4 for comparison. The initial times span a longer duration than the hitting times, i.e., from 10:16 a.m. September 12 to 9:20 a.m. September 13 with the maximum between 0:30-1:15 a.m. September 13, which was 4.5 hours later than the maximum of the hitting times. This implies that the majority of the subnet unreachabilities occurred after the storm arrived the subnet regions. However, there are subnets that became unreachable prior to the arrival of the storm. We shall further discuss this in Section III-E.

[2]Note that 11 subnets that were located close to the coast exhibited similar hitting times as their subnet regions began with the coastline rather than the 25 miles extending into the ocean.

D. Correlating Network Unreachability with Storm

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 $1 \leq i \leq 102$. Let $I_i(t)$ and $I_{hi}(t)$ be two indicator functions for subnet i, where



(a) 9/12, 10:00 a.m.-6:00 p.m.

(b) 9/12, 6:00 p.m. - 9/13, 2:00 a.m.

Fig. 6. Storm coverage at the end of each sub-interval (blue circle) and geo-locations of subnets (white circles and squares) for different sub-intervals

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

$$I_{hi}(t) = \begin{cases} 1 & \text{if } t \in [t_{hi} - T, t_{hi}], \\ & \text{i.e., the storm hit the subnet region } i \\ & \text{when } t \text{ falls in this interval;} \\ 0, & \text{otherwise.} \end{cases} \quad (2)$$

$T > 0$ 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:

$$\rho(\tau) = \frac{1}{n} \sum_{i=1}^{n} I_i[t_{ui}] I_{hi}[t_{ui} - \tau], \quad (3)$$

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 5(a) presents $\rho(\tau)$ where τ varies between $[-2,2]$ hours when T = 15 minutes. The maximum sample correlation coefficient is 0.0882 at τ = 45.02 minutes. To illustrate the robustness of this result with respect to T, Figure 5(b) presents $\rho(\tau)$ for T = 30 minutes. The maximum sample correlation coefficient is 0.1078, when τ = 40.10 minutes. To 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.

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 IV.

E. Non-Causality of Subnet Unreachability

One interesting observation emerges from our study where Figure 4 shows that subnets could become unreachable before the storm approached, i.e., the initial time appears earlier than the hitting time. Such a phenomenon corresponds to the non-causality of the unreachable subnets. To further illustrate the non-causality, we consider time duration 10:00 a.m.-6:00 p.m. September 12, when the storm had not appeared in any subnet region. Twenty subnets already became unreachable as shown in Figure 6(a). These subnets were non-causal to the storm by definition. When the storm entered the majority of subnet regions during 6:00 p.m. September 12 - 2:00 a.m September 13, more subnets became unreachable as shown in Figure 6(b).

These observations suggest that there may exist hidden variables which are neither taken into consideration nor inferable from our data. Hence, it is necessary to seek root causes directly from organizations who experienced the network disruptions.

IV. External Factor: Seeking Root Causes

We sought the root causes by actively contacting the organizations who own the unreachable subnets. All 40 organizations were contacted through extensive field work, e.g., emails and surface mails as well as phone calls. Four organizations offered the root causes.

(1) Houston Advanced Research Center (HARC): The subnet became unreachable at 2:10 p.m. September 12. Using the exact geo-location obtained from HARC, their subnet became unreachable nine hours prior to the storm coverage,

and lasted for more than a week. The network administrators shut down the network as a precaution due to lack of spare power. In addition, HARC experienced power outages starting on 7:45 a.m. September 13, after the Ike landfall. The power outages lasted until September 18, and their ISP experienced unknown network connectivity problem on September 19. HARC was able to regain their Internet connections at 10:30 a.m. September 20. There was no physical network damage caused by the hurricane. This illustrates how power outages could strongly impact network reachability.

(2) X is an ISP in Texas (the name of the ISP is not disclosed for privacy): The subnets were observed unreachable three times for short durations at 12:54 a.m., 2:10 a.m., and 8:54 a.m. September 13. Using GeoIP geo-locations, the storm coverage never appeared in six subnets, and the other four subnets became unreachable between 2.66-6.41 hours after the storm coverage appeared in their subnet regions. The root causes were reported to be power outages.

(3) NASA-Johnson Space Center: Using GeoIP geo-locations, one subnet had no storm coverage appearance, another subnet became unreachable 4.19 hours prior to the storm coverage, and the remaining unreachable subnets occurred between 3.04 and 8.18 hours after the storm coverage. The network administrators acknowledged that the root cause of some of their unreachable subnets was power outage.

(4) University of Texas Medical Branch at Galveston (UTMB): Our data observed that the network became unreachable at 8:06 p.m. September 12. Using the geo-location from UTMB, the subnet became unreachable about 1.85 hours after it appeared in the storm coverage. The network experienced power failures when the rising water flooded the power generators. The network infrastructure itself was not damaged.

In summary, three root causes were power outages, and one root cause was the lack of spare power. The communication networks themselves experienced little damage for the four organizations. One subnet was verified to become unreachable several hours prior to the storm coverage. Hence, power is another external factor, where power outages caused service disruptions in communication.

V. Conclusion and Future Direction

In summary, 32 subnets in our data set were found to become unreachable independently in time. Eighty-eight subnets became unreachable dependently in time, where 77 are from the same organization, and 11 from same AS but different organizations. The initial unreachable time of the subnets correlates 0.09 with the hitting time when the storm coverage appeared at subnet locations. Subnet unreachability exhibits non-causality, e.g., 20 subnets became unreachable before the hurricane entered any subnet region. Power outages were reported to be a type of root causes. Lack of spare power was reported to be the cause of a non-causal unreachable subnet.

As these findings are from Hurricane Ike, we also have studied network-service disruptions from Hurricane Katrina in our prior work [4]. Our experiences from using publicly available data on large-scale network-inference raise challenges and point to possible future directions, which are more general than case studies.

(a) The first research issue emerged is information acquisition across communication and power infrastructures as well as weather. For example, what types of data, and at what spatial and temporal scale can provide a deeper understanding of large-scale communication service-disruptions, and intricate interdependencies among power- and communication-infrastructures as well as external disturbances?

BGP measurements provide reachability information from control plane. More detailed data is needed for refined inference of unreachability. In addition, other types of data can provide additional information for inferring and validating unreachability, e.g., trace-route data collected by iPlane [20]. How to combine measurements of different types at different network scales is relevant to inference for communication networks.

Detailed data is also needed for relating network-service disruptions with power outages. Power outages often affect network operations within organizations. The publicly available BGP measurements, however, treat a subnet/organization as a black box. This results in disparity between the available and the needed data for power and communication networks. An open issue is how to obtain power and network measurements at a fine enough spatial and temporal scale to identify dependent disruptions.

Geo-locations of communication networks are pertinent to observing dependence among communication, power, and external disturbances. External disturbances such as hurricane can occur locally. Thus, geo-locations of communication networks need to be accurate at a fine spatial scale of, e.g., a mile, to relate to external disturbances and identify how external disturbances and power networks impact communication infrastructure. How to obtain accurate geo-location data for networks at such a small scale is an open research issue [18].

The storm data used in this work was made available hourly. An hourly time scale is too crude to be comparable with subnet unreachabilities that occurred at the scale of minutes. The model we use for storm movement is simplistic. More refined weather data and model of storm are needed to study the impact of Hurricanes.

Hence, information acquisition is needed at a finer spatial and temporal scale across heterogeneous networks, geo-locations, and weather. Such data can reflect the ground truth [21] on what happened and where.

(b) The second research issue is information sharing across multiple organizations and providers. Organizations possess pertinent information of geo-locations and root causes. However, as shown in this study, information needs to be shared on root causes to unfold “what is behind the disruptions." The reports from the four organizations provided exactly that, raising awareness on possible root causes for what we might have overlooked. However, the information contributed by the four organizations may not be sufficient for large scale disruptions. New awareness and incentives arise, where information sharing across organizations is imperative to understanding and strengthening the interdependent communication and power infrastructures.

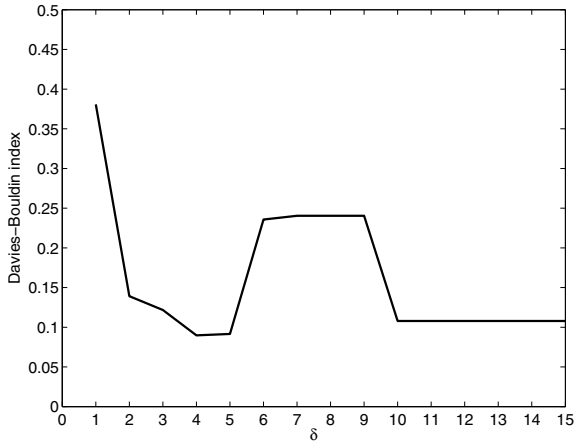


Fig. 7. Davies-Bouldin index for BGP measurements.

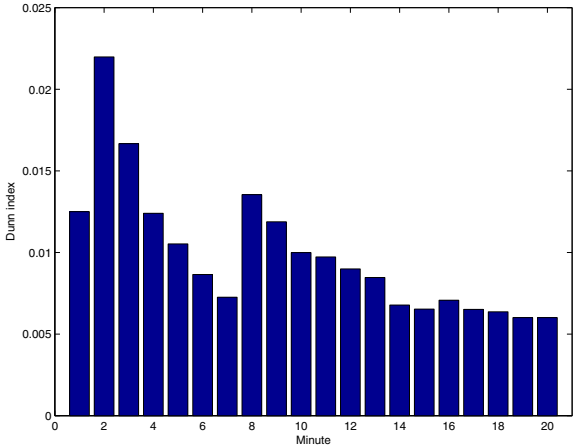


Fig. 8. Dunn index for different values of τ.

ACKNOWLEDGMENT

The authors would like to thank Houston Advanced Research Center (HARC), an anonymous ISP at Texas, NASA-Johnson Space Center, University of Texas Medical Branch at Galveston (UTMB), for providing the root causes of their network disruptions; Robert Berg and Lans Rothfusz (NOAA) for hurricane data; Gary May, David Green, Richard Davies and Yun Wang (Georgia Tech) for help with seeking the root causes, developing graphic tools, data processing and evaluating correlation; Bradley Huffaker (CAIDA) for helpful discussions on geo-location data; and Mason Porter (Univ. of Oxford) for helpful comments. The authors thank the anonymous reviewers for helpful suggestions, and the Associate Editor for help with timely reviews. Support from NSF-0952785 is gratefully acknowledged. This work started in September 2008 after Hurricane Ike.

APPENDIX A
UNREACHABILITY IDENTIFICATION

Identifying unreachable subnets is an important first step and has been studied in the prior work including our own [4], [8], [9]. Here, we use a similar approach as in [4] that includes two steps. First, the 230-dimensional time series are reduced to a small number of clusters where each cluster contains subnets with a similar pattern of BGP updates. Secondly, the BGP updates of a representative subnet from each cluster are examined for its unreachable status.

Clustering: We choose the average-linkage hierarchical clustering algorithm to group similar patterns of BGP measurements [11]. To measure the difference between two BGP time series, we convert the discrete time series of BGP update arrivals for subnet i to a continuous waveform $w_i(t)$, where $w_i(t) = 1$ if a BGP announcement arrives at time t; $w_i(t) = 0$ if a BGP withdrawal arrived. For time t between two consecutive BGP update arrivals at times t_1 and t_2 ($t_1 \leq t < t_2$), $w_i(t) = w_i(t_1)$.

Let $d(w_i(t), w_j(t))$ be the difference between $w_i(t)$ and $w_j(t)$, where $d(w_i(t), w_j(t)) = \int_{t=0}^{T} |w_i(t) - w_j(t)| dt$ for $1 \leq i, j \leq n$, and n is a total number of subnets. $d(w_i(t), w_j(t))$ is the duration that subnets i and j have different BGP updates. The set $\{d(w_i(t), w_j(t))\}_{i,j=1}^{n}$ is the input to the average-linkage hierarchical clustering algorithm [11].

Let δ be a threshold for the clustering algorithm. If $d(w_i(t), w_j(t)) < \delta$, subnets i and j are considered to have similar BGP pattern and assigned to one cluster. Figure 7 shows the Davies-Bouldin index [22] that measures cluster compactness for different values of δ. The criterion for selecting a threshold value is to choose a sufficiently small δ that corresponds to a small Davies-Bouldin index and thus compact clusters. From Figure 7, δ is chosen to be four minutes. This means that for two subnets whose BGP updates differ less than four minutes, the two subnets are considered to have a similar BGP pattern and assigned to the same cluster. The clustering algorithm reduces the spatial dimensionality by grouping 230 subnets into 105 (45.65%) clusters.

Identifying unreachable subnets: The clusters have a small average intra-cluster dispersion of 0.0001. Therefore, a representative subnet is randomly chosen from each cluster, and the BGP pattern of this subnet is examined to determine whether and when it is unreachable.

The identification of unreachable subnets are done into two steps: burst detection and unreachabilitiy inference.

Burst detection: We detect bursts of multiple BGP update messages since BGP update bursts show symptom of subnet unreachabilities. 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.

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 [23], [24]. To compensate long bursts that might occur in our data set, we select the range of τ to be between 1-20 minutes.
- Compute Dunn index [25] to measure compactness of

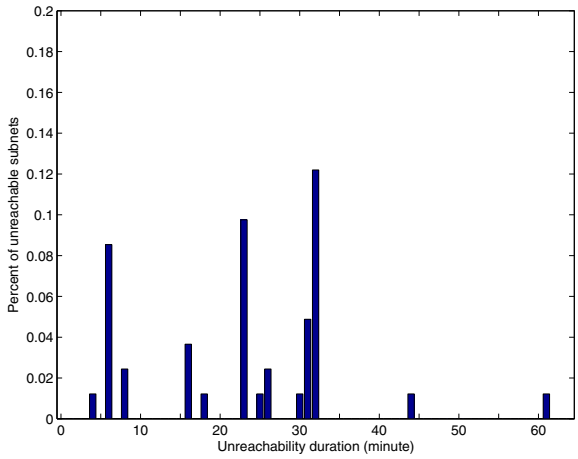


Fig. 9. The empirical distribution of unreachability duration between 8/1/08 - 9/9/08.

BGP update bursts. The Dunn index $D = \frac{d_{min}}{d_{max}}$, where d_{min} and d_{max} respectively are 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 the consistent Dunn index. The minimum value is preferred because the larger value of τ, the more BGP updates outside the underlying bursts would be incorrectly included in our detected bursts. We look for the consistent allocation of BGP update bursts despite the increasing values of τ. Figure 8 shows the values of Dunn index for different values of τ. The Dunn index decreases from 0.0085 at τ = 13 minutes to 0.0068 at τ = 14 minutes, and the Dunn index stays almost consistently between 0.0060-0.0071 for the values of τ at 14-20 minutes. Hence, since the minimum τ that results to the consistent Dunn index is prefered, we select the value of τ to be 14 minutes.

Thus, in this work, if any two BGP updates arrive within 14 minutes, we consider them to be in the same BGP update burst.

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

Note that BGP withdrawals can be lost. From the observation of this data set, 0.23% of bursts have number of missing withdrawals between 1-3. For this case, to determine whether a burst corresponds to an unreachable burst, we use an unreachability duration. The unreachable time duration is the time duration between the initial time and the next BGP announcement. Figure 9 presents the unreachability durations collected from the pre-Ike interval. The minimum unreachability duration during the pre-Ike interval (normal operations) is approximately 4.33 minutes. (Note that the remaining 50% (not shown in Figure 9) of unreachable subnets from the pre-Ike intervals had their unreachability durations longer than an hour.) Thus, if a burst has between 1-3 missing BGP withdrawals and is followed by an unreachability duration longer than 4 minutes, we identify a subnet with this burst as an unreachable subnet.

APPENDIX B
STATISTICAL TEST

The test is described below for non-uniform Poisson occurrences of unreachable subnets at the scale of more than three minutes.

1) Separate the unique initial times into three disjoint intervals: T_1=[*10:00 a.m.-8:25 p.m. September 12*], T_2= [*8:25 p.m. September 12-3:00 a.m. September 13*], and T_3=[*3:00 a.m.-8:00 a.m. September 13*].
2) Assume the unique initial times in T_k are from a Poisson point process with arrival rate λ_k, $1 \leq k \leq 3$. 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 III shows the values of N_k, T_k, and $\hat{\lambda}_k$. The interarrival times of unreachabilities for the three intervals are estimated to be 37, 12, and 60 minutes respectively.
3) Calculate the chi-square statistic $\chi^2_{d_k}$ with degree of freedom d_k for each interval T_k.
   - 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 IV.
   - 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 IV.
   - 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 $1 \leq k \leq 3$. Table IV shows values of $\chi^2_{d_k}$. As a result, $\chi^2_{d_1} = 1.3718$, $\chi^2_{d_2} = 0.3394$, and $\chi^2_{d_3} = 0.0044$.
   - 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 $1 \leq k \leq 3$.
4) Obtain the summation of the chi-square statistics, $\chi^2_m = \chi^2_{d_1} + \chi^2_{d_2} + \chi^2_{d_3} = 1.7156$. Due to the independence assumption in H_0, χ^2_m is a chi-square statistic of degree $m = d_1 + d_2 + d_3 = 3$ since the sum of independent chi-square random variables is also a chi-square random variable [26].
5) Perform the chi-square test: Given a confidence level 95%, obtain a threshold value $\chi^2_{0.05,3} = 7.82$, where $\Pr(\chi^2 < \chi^2_{0.05,3}) = 0.95$, and $\Pr()$ is a chi-square dis-

TABLE III
PARAMETERS OF DISJOINT INTERVALS. (UNIT OF T_k IS MINUTE.)

k	T_k	N_k	T_k	T_k/N_k	$\hat{\lambda}_k$	$\chi^2_{d_k}$
1	9/12 10:00 a.m. - 9/12 8:25 p.m.	17	625	37	0.0272	1.3718
2	9/12 8:25 p.m.-9/13 3:00 a.m.	32	395	12	0.0810	0.3394
3	9/13 3:00 a.m. - 9/13 10:00 a.m.	7	420	60	0.0167	0.0044

TABLE IV
CALCULATION OF CHI-SQUARE STATISTICS. $S_{(k,c_j)} = \frac{(O_{(k,c_j)} - E_{(k,c_j)})^2}{E_{(k,c_j)}}$, $1 \le j,k \le 3$.

		T_1			T_2			T_3		
j	c_j	$O_{(1,c_j)}$	$E_{(1,c_j)}$	$S_{(1,c_j)}$	$O_{(2,c_j)}$	$E_{(2,c_j)}$	$S_{(2,c_j)}$	$O_{(3,c_j)}$	$E_{(3,c_j)}$	$S_{(3,c_j)}$
1	0	14	12.6654	0.1406	20	20.7405	0.0264	15	15.0472	0.0001
2	1	6	8.6125	0.7925	15	15.1222	0.0010	5	5.0157	0.0001
3	>1	5	3.7221	0.4387	8	7.1374	0.1043	1	0.9371	0.0042
$\chi^2_{d_k}$				1.3718			0.1317			0.0044

tribution. Since the chi-square statistic $\chi^2_3 = 1.7156 < \chi^2_{0.05,3}$, the hypothesis cannot be rejected.

In the last disjoint interval, there were only seven unique initial times. Thus, we only have one sub-interval 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 did not expect this uncertainty in this last interval to affect the non-rejection of the hypothesis. We plan to resolve this uncertainty in the future work.

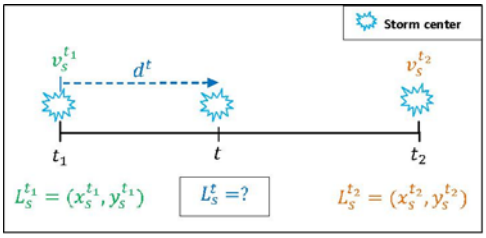


Fig. 10. One-dimensional storm motion during times t_1, t, and t_2, where $t_1 \le t < t_2$.

APPENDIX C
INTERPOLATION OF STORM DATA

From the collected storm data, let C_s^t be storm composition at time t where the subscript s specifies 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][3]. 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 10. Let d^t be the distance that storm center travels between t_1 and t, where $t_1 \le t < t_2$. 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 [27]:

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

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

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($y_s^{t_2} - y_s^{t_1}$)).

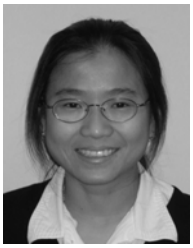
The interpolation is chosen to be at 15-minute interval and constrained to match the observation data of an hour. We observe that this constraint could be violated if an interpolation interval is too small, e.g., less than 10 minutes.

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

REFERENCES

[1] J. H. Cowie, A. T. Ogielski, BJ Premore, E. A. Smith, and T. Underwood, "Impact of the 2003 blackout on the Internet communications," Renesys Corporation, Nov. 2003.
[2] J. Cowie, "Lights out in Rio," Renesys Corporation, Nov. 2009.
[3] M. Brown, "Ike hammers Texas Internet," Renesys Corporation, Sept. 2008.
[4] S. Erjongmanee, C. Ji, J. Stokely, and N. Hightower, "Knowledge discovery from sensor data," M. Gaber, *et al.*, eds. *Lecture Notes of Computer Science*, vol. 5840, Springer, June 2010.
[5] S.C. Liew and K.W. Lu, "A framework for characterizing disaster-based network survivability," *IEEE J. Sel. Areas Commun.*, vol. 12, no. 1, Jan. 1994, pp. 52-58.
[6] R. Berg, "Tropical cyclone report: Hurricane Ike (AL092008) 1-14 September 2008," National Hurricane Center, Jan. 2008.
[7] Y. Rekhter, T. Li, and S. Hares, A Border Gateway Protocol 4. *RFC 1771*, 1995.
[8] A. Feldmann, O. Maennel, Z. M. Mao, A. Bergerm, and B. Maggs, "Locating Internet routing instabilities," *Comput. Commun. Rev.*, vol. 34, no. 4, Oct. 2004, pp. 205-218.
[9] K. Xu, J. Chandrashekar, and Z. L. Zhang, "A first step towards understanding inter-domain routing," in *Proc. ACM SIGCOMM Workshop Mining Netw. Data*, Philadelphia, PA, Aug. 2005.
[10] Route Views project, [Online]. Available: http://www.routeview.org.
[11] L. Kaufman and P. J. Rousseeuw, *Finding Groups in Data: An Introduction to Cluster Analysis.* John Wiley and Sons, 1990.
[12] P. Cheng, X. Zhao, B. Zhang, and L. Zhang, "Longitudinal study of BGP monitor session failures," *Comput. Commun. Rev.*, 2010.
[13] Whois database, [Online]. Available: http://www.arin.net/whois.
[14] The NANOG mailing list archives - August 2008, [Online]. Available: http://mailman.nanog.org/pipermail/nanog/2008-August/thread.html.
[15] M. Caesar, L. Subramanian, and R. Katz, "Towards localizing root causes of BGP dynamics," Technical Report CSD-03-1292, Computer Science Department, University of California-Berkeley, Nov. 2003.
[16] National Hurricane Center - Hurricane Ike advisory archive, [Online]. Available: http://www.nhc.noaa.gov/archive/2008/IKE.shtml.

[17] National Hurricane Center - A digital record of the complete best track data, [Online]. Available: ftp://ftp.nhc.noaa.gov/atcf/archive/2008/bal092008.dat.gz.
[18] B. Huffaker, "Geocompare: A comparison of public and commercial geolocation databases," *ISMA Workshop on Active Internet Measurements*, La Jolla, CA, Feb. 2011.
[19] GeoIP City, [Online]. Available: http://www.maxmind.com/app/city.
[20] iPlane: An information plane for distributed services, [Online]. Available: http://iplane.cs.washington.edu.
[21] R. Oliveira, D. Pei, W. Willinger, B. Zhang, and L. Zhang, "In search of the elusive ground truth: The Internet's AS-level connectivity structure," in *Proc. ACM Sigmetrics*, Annapolis, MD, June 2008.
[22] D. L. Davies and D. W. Bouldin, "A cluster separation measure," *IEEE Trans. Pattern Recognition Machine Intelligence*, vol. 1, no. 2, Apr. 1979, pp. 224-227.
[23] C. Labovitz, G. R. Malan, and F. Jahanian, "Internet routing instability," *IEEE/ACM Trans. Netw.*, vol. 6, no. 5, 1998, pp. 515-528.
[24] C. Labovitz, A. Ahuja, A. Bose, and F. Jahanian, "Delayed Internet routing convergence," *IEEE/ACM Trans. Netw.*, vol. 9, no. 3, 2001, pp. 293-306.
[25] J. C. Dunn, "Well separated clusters and optimal fuzzy partitions," *Cybernetics Syst.*, vol. 4, 1974, pp. 95-104.
[26] J. A. Rice, *Mathematical Statistics and Data Analysis*. Duxbury Press, 1995.
[27] E. Williams, "Aviation formulary v1.144," [Online]. Available: http://williams.best.vwh.net/avform.htm.



Supaporn Erjongmanee received the B.S. degree in Electrical and Computer Engineering from Carnegie Mellon University, Pittsburgh, in 2001, and the M.S. and the Ph.D. degrees in Electrical and Computer Engineering from Georgia Institute of Technology, Atlanta, respectively in 2003 and 2011. Currently, she is working at the Computer Engineering Department, Kasetsart University, Bangkok, Thailand. Her research interests include network management, machine learning, and large-scale database analysis.

Chuanyi Ji received the B.S. (Honors) degree from Tsinghua University, Beijing, China, in 1983, the M.S. degree from the University of Pennsylvania, Philadelphia, in 1986, and the Ph.D. degree from the California Institute of Technology, Pasadena, in 1992, all in electrical engineering. She is an Associate Professor in the Department of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta. She was on the faculty at Rensselaer Polytechnic Institute, Troy, NY, from 1991 to 2001. She spent her sabbatical at Bell Laboratories, Lucent Technologies, in 1999, and was a visiting faculty at the Massachusetts Institute of Technology, Cambridge, in Fall 2000. Her research areas are in network management and security, machine learning and their interfaces. Her research interests lie in understanding and managing complex networks, machine learning theory and algorithms, and information theory. Chuanyi Ji received an NSF Career Award in 1995, and an Early Career Award from Rensselaer Polytechnic Institute in 2000.