Disaster Network Evolution Using Dynamic Clustering of Twitter Data

Yilang Wu†, Krishna Kant*, Shanshan Zhang*, Amitangshu Pal*, Junbo Wang†
† *Aizu University, Japan*
* *Temple University, Philadelphia, PA*

***Abstract*—Ad hoc smartphone networks can be used to augment communications degraded by disasters provided that the individual ad hoc clusters can reach some "connection gateways" to get out to the Internet. This capability can be provided by devices in the surrounding area that retain cellular connectivity in addition to the connectivity provided by the specially deployed emergency equipment, if any. The disconnected areas may not be known until they are back online; however, we need a mechanism to estimate them so that the gateway devices can be best recruited to provide the connectivity. This needs to be done in a dynamic environment because of the significant mobility in the wake of the disaster. In this paper, we propose a mechanism to estimate regions that are likely to be dense but disconnected, and with significant connected devices in and around them. Such regions are most likely to benefit from the ad hoc network. Because of the lack of direct information on people (or smartphone) density, we attempt to do this by analyzing the twitter data. We use our approach on the twitter data available on hurricane Sandy in 2012.**

***Index Terms*—Ad hoc Network, Twitter, Network Evolution, Spatial Clustering**

1. Introduction

With increasing frequency and intensity of natural disasters, and increasing impact of all types of disasters on large urban areas, it is important to focus on the key issue of facilitating communications during and after the disaster. The communications networks become stressed in the aftermath of a disaster both due to heavier traffic and potentially reduced capacity due to damage to the infrastructure (e.g., cell towers). Thus it is important to assess the density of the people vis-a-vis the communications capability and capacity in the disaster region. Disaster events generally lead to substantial non-routine movement of people, which results in people densities to change over time. In addition, the damage and repair processes also typically evolve with time especially for events that go on for days (e.g., wild-fire, floods, hurricanes, etc.) The dynamic evolution makes the emergency provisioning of communication resources quite difficult because of lack of predictability in mobility, damage, or even repair processes. For example, the mandatory/voluntary evacuations before the disaster are expected to result in large scale movement of people, but their destinations beyond the immediate area of danger can be highly varied. During the disaster, unplanned movements may occur for a variety of reasons including the impact of additional failures. During hurricane Sandy, there was a major power failure in Manhattan over a large area. Interestingly, much of it was repaired rather rapidly but an area of western Manhattan remained dark for several days, causing movement out of this area.

This paper is concerned with assessing areas with large people density that may be facing failure and/or severe overload of the communications network, so that it is possible to expeditiously provide them with additional external connectivity via either the surviving infrastructure around that area, or via additional communications equipment. We propose an approach to efficiently estimate dense regions that are likely to be disconnected, which can be used on large dynamically evolving data sets. Unfortunately, our currently available data sets, including the one for hurricane Sandy analyzed here, contain only sporadic communications outages which limits large scale validation of our approach.

Section 2 places this problem in the context of our ongoing work on smartphone integration into the emergency response networks. Section 3 discusses the problem of estimating how many potentially affected and unaffected smartphones are in the area. Then, section 4 introduces the method of dynamic clustering of twitter data to solve this problem. Section 5 discusses about the recruitment process of the surviving smartphones to work as gateways to the ones that lost communication. Finally, section 6 concludes the discussion.

2. Ad Hoc Communications Network

Given the ubiquity of smartphones, we exploit them to augment damaged communications. We assume that each smartphone has our app, called EDARWIN, [1] installed and this app can be initiated during disasters. The network is intended for emergency message/data transmissions from people (or phones) that do not have any direct cellular coverage to an emergency control center (ECC). The ECC is assumed to be outside the disaster area and fully operational (including Internet connectivity). We envision the ad hoc network to consist of many disjoint clusters of smartphones in the disaster area, each of which is ultimately able to

This research was supported by NSF grant CNS-1461932. *E-mails: y-wu@u-aizu.ac.jp, kkant@temple.edu, zhang.shanshan@temple.edu, amitangshu.pal@temple.edu, j-wang@u-aizu.ac.jp*

reach one or more "connection gateway" that has Internet connectivity and thus can reach ECC. These issues are discussed briefly next.[1]

2.1. Building SmartPhone Clusters

When a disaster affected area loses cellular connectivity, it starts to build out a cluster of all EDARWIN enabled smartphones in the area. Since a mere loss of signal does not indicate a disaster, we exploit the Wireless Emergency Alerts (WEA) signals to trigger the buildout. WEA is the US centric implementation of an international standard called Common Alerting Protocol (CAP) and is supported by nearly all current smartphones [2]. For building the cluster, we exploit the WiFi hot-spot mode, since it is available universally on almost all smartphones, and offers a much more practical solution than say, ad hoc mode WiFi [1]. Building WiFI-hotspot based network is a bit challenging due to the need for complementary modes (hot-spot vs. client) among any two communicating smartphones, and energy conservation by keeping the smartphones in a deep sleep model when not communicating actively. The details of the discovery of neighboring devices, building the ad hoc network, and energy efficient data transfer issues are discussed in [1]. Because of the challenges involved and significant delays of multi-hop transmissions, the network is suitable for emergency data transfers only, rather than voice communications. However, bulk transmission – such as transmitting multiple pictures/sounds captured by the phone to help with rescue/safety assessment – can be done easily through this network since the biggest delay/energy expenditure is in setting up a continuous communication path rather than the data transfer.

2.2. Connection Gateways

A connection gateway is ideally a especially designed emergency communication node (ECN) that provides both WiFi and long distance (e.g., satellite) connectively. ECNs are usually mounted on emergency communications vehicles (ECVs) so that they can be positioned as needed and then moved as the situation evolves. However, such deployments are generally difficult due to physical inaccessibility of disaster hit area and can take a long time to deploy. Fortunately, in most disasters, the communications network is only partly damaged, e.g., some cellular towers (or their backend connections) are damaged, while others are not. Therefore, we focus on scenarios where the smartphones that still have cellular coverage can act as connection gateways. That is, there are two types of phones in the disaster area: *Affected Vulnerable Devices* (AVDs), i.e., those without direct cellular connection, and *Unaffected Vulnerable Devices* (UVDs), i.e. those with cellular connection. Here "Vulnerable" means that all devices have received the WEA signal and presumed to be in the disaster area. Recruiting UVDs for providing the gateway service and ensuring that they are not burdened unnecessarily are crucial for the scheme to be successful. Note that in order to act as a gateway, the UVDs with EDARWIN app would need to turn on their WiFi radio and discover AVDs in their vicinity in an energy efficient manner. (According to our scheme in [1], UVDs are always in client mode so that they will not discover other UVDs.) Each UVD will also keep track of who it has talked to recently (for few hours) to avoid repeated connection establishment and authentication message exchange.

We assume that UVDs are recruited by ECC, possibly with the involvement of cellular carriers who could provide them incentive to participate; however, we do not focus on those aspects here. Instead, we consider the problem of how to best make use of the willing UVDs and minimize burden on them. Obviously, the burden on UVDs depends on how many AVDs need to be served around them and how many other UVDs are present. While the number and location of UVDs can be determined, this is not possible for AVDs until the ad hoc network has been built and connected to ECC. In the following, we discuss the problem of estimating AVD density and identifying suitable UVDs to recruit by using the twitter data.

3. Mobile Density Estimation

We estimate the density of AVDs and UVDs within an area by tracking the movements in and out of the area starting with the normal situation. Given the assumption of almost everyone carrying a smartphone, tracking of GPS coordinates of all people (really, smartphones) is theoretically possible, but the data is not publicly available. Instead, we try to estimate it using the publicly available twitter data that is labeled with GPS coordinates. Given the density estimates, the ECC can recruit willing UVDs suitably (based on their ability to dedicate bandwidth for gateway function and other aspects if known, such as the battery level). ECC may also rotate the gateway role among UVDs to minimize battery impact, and would need to make changes based on UVD mobility, i.e., do the least disruptive handoff from one UVD to another. In the following, we address the issue of density estimation and tracking its evolution. The actual use of density estimates is beyond the scope of this paper and a part of our continuing research.

User density estimation from available twitter data faces several difficulties and sources of error. First, not everybody tweets and the twitter API provides only 1% of the tweets. However, studied over a rather long time window (e.g., an entire day or a good fraction of a day), it can provide a reasonable estimate of the user presence. Since the data collected does include unique user IDs, it is possible to identify all tweets of a given user, from which user density can be computed. For mobile users the location depends on tweet time, which complicates matters. We use "average" location of the user for density estimations. The second difficulty is that in the environment of interest here, no tweets will be recorded in an area where the cellular communications are down. In other words, the tweets only tell us about areas

1. Once reached, the ECC also authenticates devices and confirms the disaster, so that the ad hoc network is not misused.

that do have cellular coverage, i.e., the UVDs around the areas where communications are lost. The AVD density can be estimated based on the last known density updated for increase or decrease in density in the surrounding areas.

We have done a preliminary analysis of the twitter data for NY/NJ area around Manhattan for hurricane Sandy in 2012. The hurricane actually hit NY/NJ area on Oct 29, and its immediate effects lingered for several days. Fig. 1 shows the change in density over successive days starting with Oct 25 (i.e., difference between Oct 25 and 24) until Nov 2. Similarly, Fig. 2 shows the changes over Oct 29 thru Nov 01. A red square indicates a net density gain over the previous day, and a blue square shows net loss, with darker shade showing a larger change.

Let us focus on Manhattan which is the most heavily populated area. It is seen that Figs 1a and 1b do not show normal movements because of almost complementary nature of the density change across 2 days. However, Fig 1c shows an appreciable change. This is because on Oct 27, the path of the storm is known with high certainty and people move to safer areas. In particular, the midtown area is somewhat depleted. Oct 29 shown in Fig 2a shows a large depletion in the western part. This is due to the hurricane landfall and widespread power failure as a result. Figs 2b and 2c show the power failure persisting in extreme western part, and on Nov 01 (in Fig 2d), the situation returning to normal. Unfortunately, neither the density data for Oct 25-Nov 1 (not shown), nor the differences shown in Figs 1 and 2 show any significant predictability. That is, *given the density for any given day and past history, it is not possible to predict what will happen the next day (except during normal period before the hurricane)*. While this is not a problem for using UVDs as gateways and changing them as needed; it is a problem for deploying emergency communications equipment which might take more than a day to plan, reach the site and deploy. Therefore, we shall focus exclusively on recruiting UVDs for providing the connectivity.

While the power failure did happen in this instance and is confirmed by the tweets, the reports of Internet failure were very sporadic as discussed later. In fact, if there were widespread communication outages, we would not have tweets from those areas. Thus, tweets from the affected area can only be inferred by their absence! Therefore, assuming that the areas of dark blue represent areas where additional support for emergency communications will be useful. We thus attempt to determine areas of large density depletion, and find the red areas in and around them where UVDs can be recruited. In the following, we discuss an efficient spatial clustering algorithm for estimating such areas.

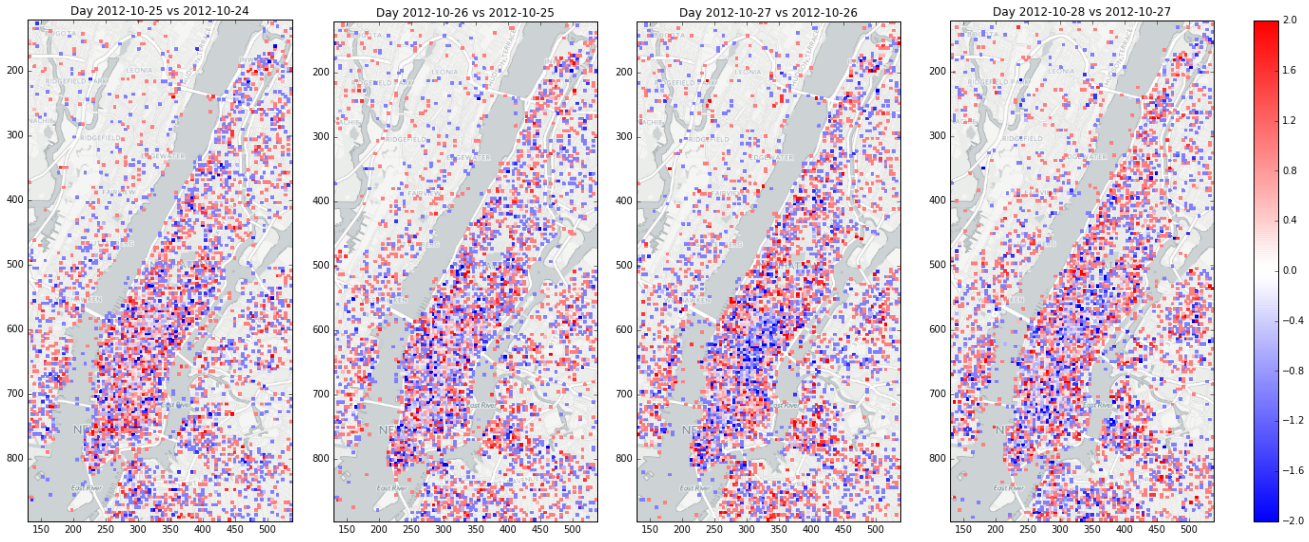
4. Spatial Clustering

In order to determine areas of large density depletion, here we propose a spatial clustering method to efficiently load and cluster the spatial twitter data that comes in large scale and dynamic changes. The spatial clustering is usually computationally expensive, making it inefficient to deal with the large volume dynamic data. Taking the traditional grid-based spatial clustering for example, it takes about $|Cell|^2$ (the $|Cell|$ is the total amount of grid cells) steps for it to execute the spatial clustering, which causes big delay and distorted result regarding the newly loaded data. The hierarchical-partition results in a tree-like data structure, which handles sparsely and densely populated spatial data space more efficiently. The tree-based data structure has been widely used in database technology to organize the spatial data for rapid query. In the proposed method, spatial clustering is reconstructed in a tree-based structure to load the dynamic spatial data and determine the spatial cluster by scanning the dense regions along the tree.

The dynamic spatial clustering is feasible only if the spatial data in computation is up-to-date regarding the large volume dynamic data, and the spatial clustering is efficient in process. Figure 3 shows the partition of point-type spatial data; the data space is partitioned into a tree-based structure while each page (node) represents the portioned sub-region in the map. We choose the K-d-B-tree (k-dimensional Balanced Tree) as the tree-based structure for its search efficiency similar to a balanced k-d tree, and optimal external memory accesses for block-oriented storage ([3]). Figure 3-(b) shows the partition into a 2-d-B-tree ($k = 2$ in this case). There are four necessary operations used to keep the tree-based data structure update to the dynamic spatial data loading from twitter. The splitting and reorganizing operations are used to keep the property of k-d-B-tree, such as keeping the tree balanced and avoiding the overfull in the block storage for each point page. The data loading and discarding the spatial big data is achieved by inserting and deleting operations. In Figure 3-(c), the spatial clustering can be achieved by searching the pages (nodes) with higher density of emergency-related data points.

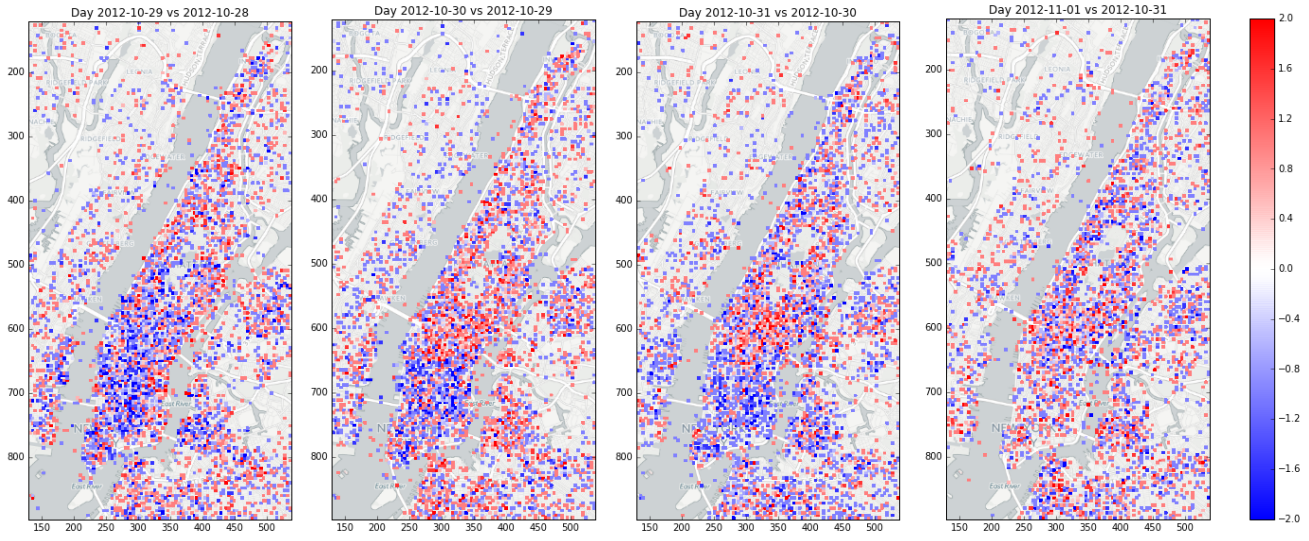
As shown in Figure 4, the dense region scanning uses a depth-first search in three steps. A stack is constructed to store the intermediate parameters of each region's pages ID, density value, and boundary. A depth-first scan is designed in three steps. It first pushes all regions' intermediate parameters onto the stack, and then pops the intermediate parameters to compare them with the threshold using the judgement logic (as shown in Table in the middle of Fig. 4. It then determines the dense regions according to the enumerated eight finite states of scanning results. The proposed clustering method is much more efficient than the traditional grid-based one. Let A denote the set of all spatial data points, and let us denote the size of singular block memory for the k-d-B-Tree as *Block*. The total number of leaf nodes (pages) of k-d-B-Tree equals to $\frac{|A|}{Block}$, and its height equals $log_2 \frac{|A|}{Block}$. Since the k-d-B-Tree is a balanced binary tree, its total number of nodes (pages) is given by $(2^{log_2 \frac{|A|}{Block}+1} - 1)$. The current dense region scanning has to parse all nodes in k-d-B-Tree, therefore, its computational complexity will be close to $O(\frac{|A|}{Block})$.

The proposed spatial clustering clusters the disaster-related region that is dramatically changing during emergency. Fig. 5 shows results from spatial clustering of the



(a) Oct25/24 Difference (b) Oct26/25 Difference (c) Oct27/26 Difference (d) Oct28/27 Difference

Figure 1: Density Changes over Oct 25 - Oct 28, 2012



(a) Oct29/28 Difference (b) Oct30/29 Difference (c) Oct31/30 Difference (d) Oct31/Nov 01 Difference

Figure 2: Density Changes over Oct 29 - Nov 01, 2012

Hurricane Sandy twitter data. The dynamic spatial clustering computes two density functions. One is the density of tweets with disaster-related keywords, such as "power" and "Internet" which have abnormal signal spikes. Fig. 5-a and Fig. 5-b illustrate the daily hotspots based on the first density measure. The two figures show the situation before and after the storm hit NY/NJ area. It is clearly seen that the affected area expanded significantly after the storm. In fact, the tweets about power and Internet before the storm are likely to be about potential rather than real events. It appears that the Internet failures were very sporadic instead of large scale. This is actually a more desirable scenario for the smartphone based ad hoc networks for the outage areas since it would be relatively easy to find UVDs that can be recruited. Nevertheless, in the following, we shall try to identify the affected areas (no Internet coverage, poor coverage, or heavy congestion) based on the drop in tweet density itself.

The other density function is the density drop of any given sequential time slots. Fig. 5-c and Fig. 5-d respectively show the drop in tweet density before and after the storm with the drop extent being shown by various overlayed rectangles. The data clearly shows a significant drop in tweets following the hurricane.

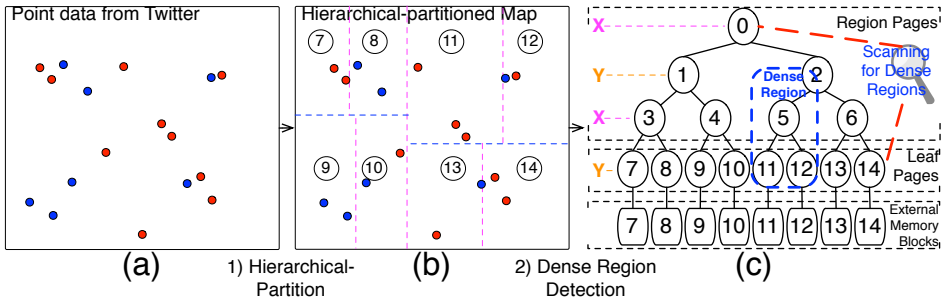
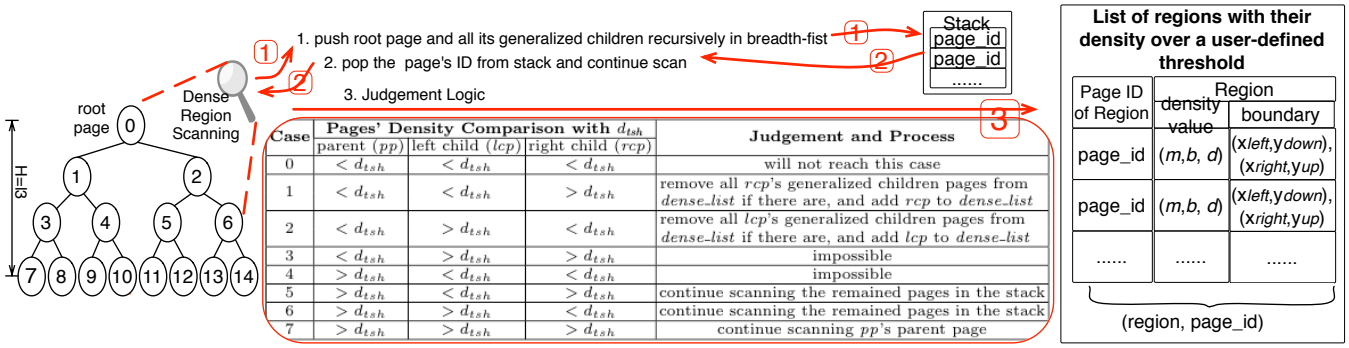
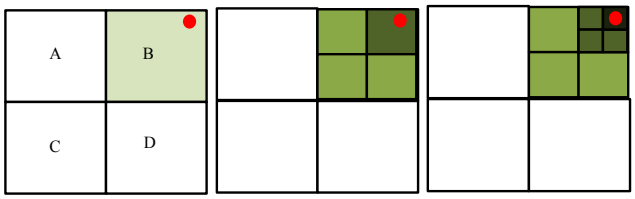


Figure 3: Constructing Dynamic Spatial Data into a tree-based Data Structure



Case	Pages' Density Comparison with d_{tsh}			Judgement and Process
	parent (pp)	left child (lcp)	right child (rcp)	
0	$< d_{tsh}$	$< d_{tsh}$	$< d_{tsh}$	will not reach this case
1	$< d_{tsh}$	$< d_{tsh}$	$> d_{tsh}$	remove all rcp's generalized children pages from $dense\_list$ if there are, and add rcp to $dense\_list$
2	$< d_{tsh}$	$> d_{tsh}$	$< d_{tsh}$	remove all lcp's generalized children pages from $dense\_list$ if there are, and add lcp to $dense\_list$
3	$< d_{tsh}$	$> d_{tsh}$	$> d_{tsh}$	impossible
4	$> d_{tsh}$	$< d_{tsh}$	$< d_{tsh}$	impossible
5	$> d_{tsh}$	$< d_{tsh}$	$> d_{tsh}$	continue scanning the remained pages in the stack
6	$> d_{tsh}$	$> d_{tsh}$	$< d_{tsh}$	continue scanning the remained pages in the stack
7	$> d_{tsh}$	$> d_{tsh}$	$> d_{tsh}$	continue scanning pp's parent page

Figure 4: Efficient Spatial Clustering by Scanning Dense Regions (over Threshold d_{tsh}) along the Tree

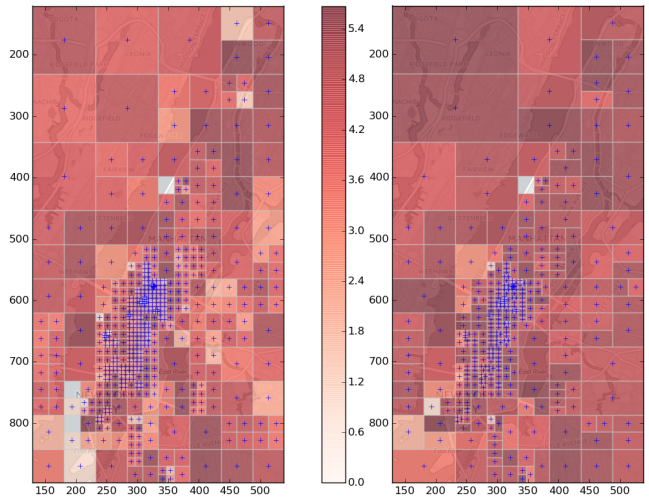


(a) Original Grid (b) After one Split (c) After two Splits

Figure 6: Grid Splitting Illustration (Red circle denotes sub-cells where AVD density $> \tau$.

5. Recruitment Process for UVDs

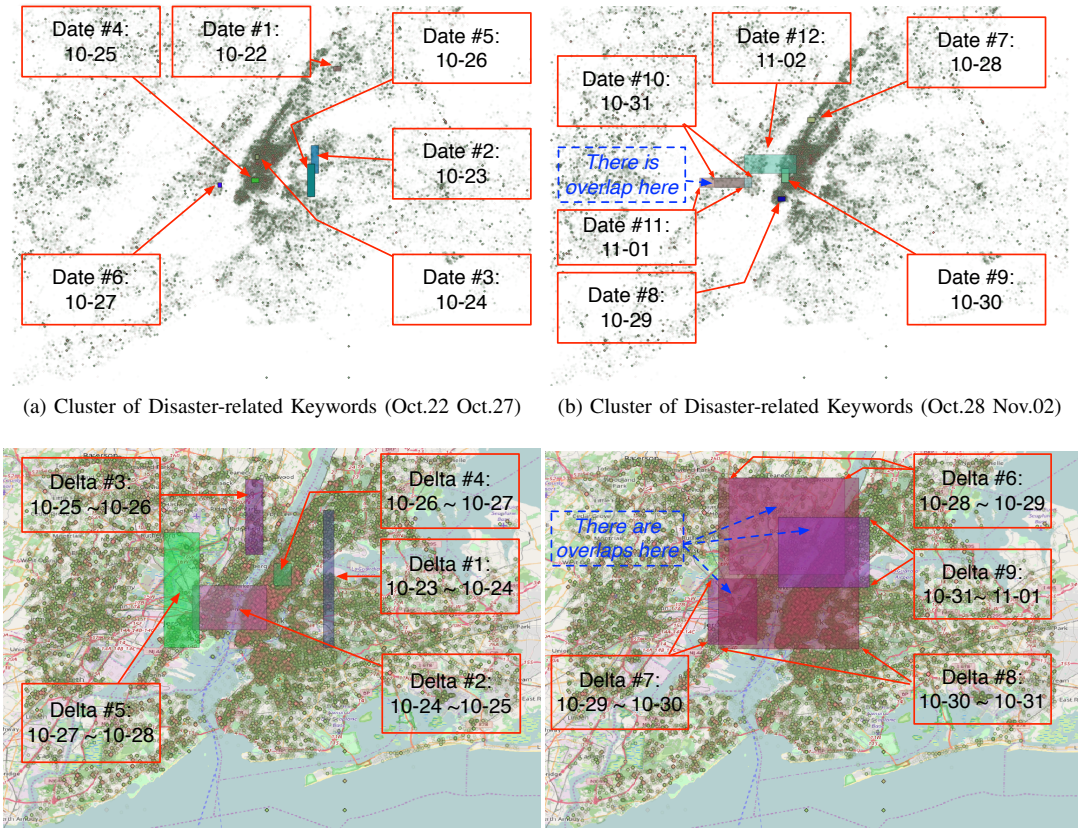
Since the density of the AVDs are extremely non-uniform, we develop a scheme which divides the entire affected region into a nonuniform grid so that any UVD can be recruited from each cell. We propose a heuristic to develop this scheme. We first divide the region into a uniform grid where the size of a grid cell is $D_{max} \times D_{max}$ meters as shown in Fig. 6a. The size D_{max} is chosen so that all AVDs within a cell can be reached within a few hops. This is essential since the performance of a multi-hop network in terms of throughput and delays degrades rapidly with the number of hops, and more than 5-10 hops becomes impractical in practice. We next check whether in any cell, the density of AVDs is more than some threshold τ. If so, such elements are subdivided into 4 smaller sub-elements as shown in Fig. 6b, provided that all the smaller sub-elements consist of atleast one UVD, and the element dimensions remain more than $D_{min} \times D_{min}$. This process goes on until in all the sub-cells the AVD densities are less than τ.



(a) After Grid Formation (b) After Grid Merging

Figure 7: Heatmaps after the non-uniform grid formation and subsequent mergers

When this process completes, we choose one UVD from each of the smallest grid cells. Such divisions ensure that the load density on all cells (or on the UVDs) are limited. We simulate this scheme on the tweets from unique users from 10/23–10/26 (4 days before Sandy hurricane) as the total population. The locations of the users are extracted from their last tweet. The total number of such unique users is found to be 17,314. Fig. 7a shows the corresponding results after the non-uniform grid formation, with D_{max} = 800 meter, d_{min} = 100 meter and τ = 50. Here the blue crosses



(a) Cluster of Disaster-related Keywords (Oct.22 Oct.27)

(b) Cluster of Disaster-related Keywords (Oct.28 Nov.02)

(c) Cluster of Density Drop (Oct.23 Oct.27)

(d) Cluster of Density Drop (Oct.28 Nov.01)

Figure 5: Spatial Clustering of Tweets for hurricane Sandy

show the UVDs required to provide connectivity. Obviously, the grid cells from the densely populated areas of Manhattan are much smaller compared to those from other areas.

It is easy to see that many cells can be *merged* into a larger elements while still ensuring that the AVD density does not exceed τ. Fig. 7b shows result of recursive merging which ends when no more elements can be combined. As can be seen, the merging reduces the number of UVDs significantly (from 538 to 320).

6. Conclusions and Future Collaboration

In this paper, we examine the problem of determining the most prominent disaster affected areas that where connectivity may be lost but there is potential to provide connectivity via devices around them that have connectivity. We presented an efficient algorithm for identifying such clusters using analysis of the twitter data. The work described here is very preliminary and ongoing. It needs to be examined in much more depth in order to determine the robustness of the method and its usefulness in real disaster scenarios. A basic limitation in this regard is lack of suitable data sets with known areas of communication loss, against which the algorithm can be tested. Japan, unfortunately, has experienced several large earthquakes and we will examine what data is available and how it can be used for network evolution. In particular, understanding people's movement pattern in different phases of a disaster will be useful for adapting the communication networks and placing of ECNs.

We will also study the interaction between the transportation and telecommunication networks in large urban areas which tend to be intertwined during disasters. In particular, the lack of communications degrades routing decisions and worsens congestion; whereas the congestion and blocked roads prevent ECVs to be positioned in the most needed areas. Building a successful smartphone based ad-hoc network could break this circular dependency and allow for faster repair of communications infrastructure and better transportation to assist with evacuations and rushing emergency supplies and health care.

References

[1] A. Pal, M. Raj, K. Kant, and S. Das, "A smartphone based network architecture for post-disaster operations using wifi tethering," submitted, available at http://www.kkant.net/papers/TMC_paper.pdf.

[2] "Guide to implementing the integrated public alert and warning system (ipaws)," "https://www.cseppportal.net/Training Documents/IPAWS_HowToGuide_21JUL2014.pdf".

[3] J. T. Robinson, "The kdb-tree: a search structure for large multidimensional dynamic indexes," in *Proc. of 1981 ACM SIGMOD*. ACM, 1981, pp. 10–18.