Social Media Driven Big Data Analysis for Disaster Situation Awareness: A Tutorial

Amitangshu Pal*, Junbo Wang†, Yilang Wu, Krishna Kant*
*Information Science and Engineering, Temple University, Philadelphia, USA
Email: {amitangshu.pal, kkant}@temple.edu
†School of Intelligent Systems Engineering, Sun Yat-sen University, China
Email: {y-wu, j-wang}@ieee.org

***Abstract*—Electronic communications play a key role in the assessment of situation in the event of disasters and accordingly dispatching of aid and rescue resources. These communications are shifting more and more towards social media postings, particularly using the twitter platform. Extracting intelligence from the available data involves several challenges that we discuss in this paper. This includes (a) filtering out irrelevant data, (b) fusion of heterogeneous data generated by the social media and other sources, and (c) working with partially geo-tagged social media data in order to deduce the needs of the affected people. Bigdata techniques are essential to accomplish this because of large volume of data, much of which is not very relevant. Spatial analytics of the data plays a key role in understanding the situation but available only sparsely because many users do not want to be tracked. We also discuss the role of edge computing handling this analytics in a scalable manner.**

***Index Terms*—Spatial Big Data Analytics; Crowd Big Data; Disaster Management**

I. Introduction

Situational awareness is crucial in a disaster scenario and is often difficult to come by due to difficulty in obtaining the necessary information in coherent manner and organizing it. Part of the difficulty arises due to potential damage to and overloading of communications networks, but to a large extent it is not clear a priori what information is most relevant and how it should be gathered. Since disasters evolve over hours and days, tracking situational awareness becomes even more challenging. Lately, social media has emerged as a primary means for informing the ground realities and expressing the needs by people caught in the disasters. Since much of the social media access in disasters occurs from smartphones, it is possible, in theory, to find the spatial location of the data origin, but in reality location information is rather spotty due to privacy concerns. Twitter has established itself as the disaster communication vehicle of choice due to its modest networking requirements, ease of use, and brevity. For example, after 2011 Japanese earthquake there were more than 5,500 tweets per second after the disaster. Twitter has been used for a wide variety of disaster scenarios, including the three major Hurricanes in 2017, namely Harvey, Maria and Irma that affected Carribian and US East coast [1], 2019 Pan-European Floods [2], and 2019 US midwestern floods [3].

Fig. 1(a)-(b) shows the distribution of earthquake related tweets (with keywords 'earthquake', and 'jishin' which means disaster in Japanese) in the Kumamoto Earthquake that struck at Kumamoto City of Kumamoto Prefecture in Kyushu Region, Japan in 2016. The density of these keywords shows close correlation with the shake map observed to the east of Kumamoto City obtained by the Geospatial Information Authority of Japan and National Disaster Institute for Earth Science and Disaster Resilience [5]. Fig. 1(c) shows the power outage related geo-tagged tweets from New York city during the occurrence of Hurricane Sandy in 2012. The storm hit New York city hard on Oct. 29th night, leaving hundreds of thousands without power [6]. Fig. 1(d) shows the intensity map of the affected areas in the Eastern USA, demonstrating that New York and New Jersey areas were worst affected by the storm. In fact the regions of Lower Manhattan from Madison Square to the tip of the island was hit the hardest, with more than 0.24 million people without power as of noon on Nov. 1st. The distribution of the such disaster related tweets was well correlated with the actual areas of damage, which shows the usefulness of the tweet analysis.

In addition, various network performance related data, such as network usage, call drops, bandwidth utilization, signal strength measurements etc. can also be obtained from the radio access network (RAN), the core network (CN), and Internet service providers (ISPs). A significant amount of such data can be accumulated and used to gain valuable insights into where and how the network repair or capacity addition should be scheduled. In this paper our main objective is to explore situational awareness in a disaster area through various types of big data analysis.

The paper is organized as follows. Section II describes the background of spatial big data analysis. Section V-A–IV summarize the possibilities and challenges of big data analytics in emergency management networks. We have demonstrated an application of big data analysis in section VI through a case study. The paper is concluded in section VII.

II. Spatial Big Data Analytics

Junbo: Spatial Prediction: (1) (2)

Spatial analytics is quite important in disaster scenarios, by studying and discovering the relations between the data

Junbo Wang is the corresponding author.

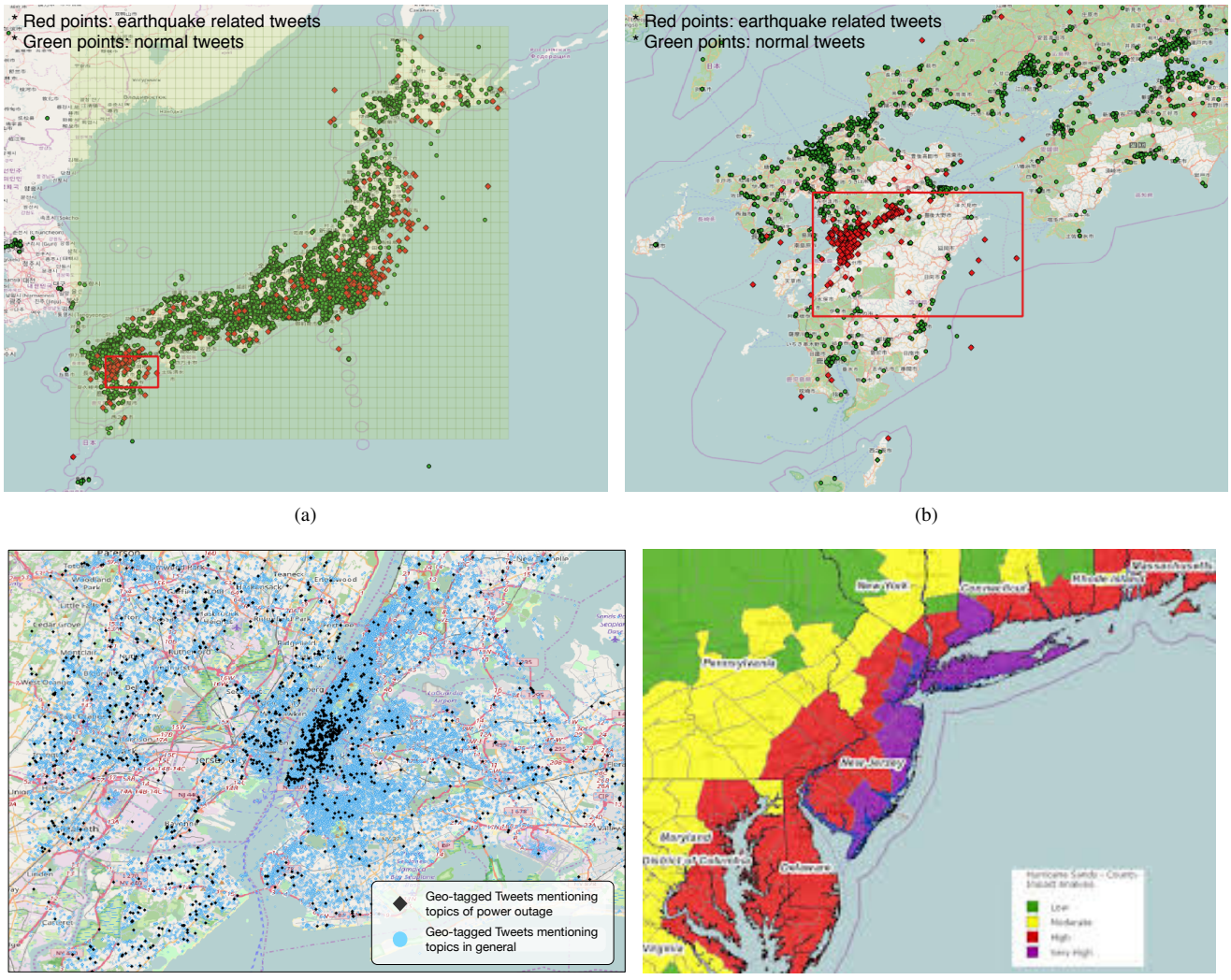


Fig. 1. Distribution of Kumamoto Earthquake (April-May, 2016). (a) Hotspot of Earthquake related Tweets after Kumamoto Earthquake and (b) its zoomed in view. This high volume of tweets overlaps with the region of the epicenter as obtained the ground reality. (c) Tweets Distribution of Power Outage during Hurricane Sandy 2012. (d) Illustration of the affected areas that were largely affected by Hurricane Sandy [4].

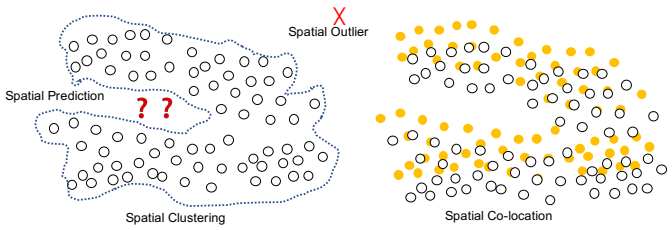


Fig. 2. Four Major Types of Spatial Analytics.

and the location where the data is generated or is intended for. Extracting interesting and useful patterns from the spatial information of data is important and yet difficult due to the complexity of spatial data types, spatial relations, and spatial auto-correlations [7]. In this section we introduce four major approaches in spatial analytics [8], namely spatial prediction, spatial clustering, spatial outlier detection, and spatial co-location pattern discovery.

A. *Spatial prediction*

Spatial prediction models can be used to support crime analysis, network planning, and services after natural disasters such as fires, floods, droughts, plant diseases, and earthquakes. Consider, for example, n points with locations denoted as $s_1, s_2, ..., s_n$, and a set of explanatory features $X = [x(s_1), x(s_2), ..., x(s_i), ..., x(s_n)]^T$ at these locations. Let $Y = [y(s_1), y(s_2), ..., y(s_i), ..., y(s_n)]^T$ denote the "situation" at these points, which refers to the learned function $Y = f(X)$ representing a quantity of interest. The function $f(X)$ is usually known only in certain locations, and we are interested in predicting it at others. This is illustrated in Fig. 2. Here, we want to predict the situation at the location of the red question mark based on the surrounding situations and the spatial correlation among the data.

Spatial prediction models mainly can be divided into two categories, i.e., spatial auto-correlation (dependency) and spatial heterogeneity (non-dependency) models.

1) Spatial auto-correlation: Spatial auto-correlation follows the first law of geography, i.e., "everything is related

to everything else, but near things are more related than distant things". For example, closer locations are likely will the similar situations both in terms of the needs of the people and conditions (e.g., wireless signal strength). Spatial auto-correlation can be further divided into the approaches based on spatial contextual information and the approaches based on prediction models [9].

The first approach is achieved through augmenting input data with additional spatial relation information [10], with additional spatial context information from multi-source [11][12] and so on. Once spatial contextual information is added, traditional data prediction algorithm can be used (even not special designed for spatial data). Instead of generating spatial contextual information, the other approach directly incorporate spatial dependency in the prediction model, and the main strategies include Markov random filed based models [13] and Kriging based models [14].

As an example, Kriging (Gaussian process regression) [14] is a typical method for spatial prediction, which utilizes an observed spatial relation to determine the range of auto-correlation. In the kriging method, it is assumed that each point i in a space is associated with a value z_i. Let u denote a point whose value, i.e., z_u, is unknown. Then let $V(u) = \{1, \ldots, N_n\}$ be a set of u' neighboring points, and z_i represents the known value in prior for each point $i \in V(u)$. In ordinary kriging, the unknown value z_u at point u is estimated as a weighted linear combination of the known values in $V(n)$ as shown in Eq. 1. To minimize the estimation error, kriging calculates a set of optimal weights by using Eq. 2. In Eq. 2, $h_{i,j}$ represents the distance between two points i and j, $\gamma(h_{i,j})$ is a function for the spatial correlation measure that represents the spatial variance in the distance between all pairs of sampled locations in space, and λ is the Lagrange multiplier to minimize the kriging error. The ordinary kriging method assumes that the mean is a constant for a neighborhood point, which can be represented as the estimation error at an unknown point u is zero, i.e., $E(\hat{(z)}_u - z_u) = 0$. The optimal weights in Eq. 2 are found by minimizing the variance of the estimation error, i.e., $Var(\hat{z}_u - z_u)$.

$$\hat{z}_u = \sum_{i \in V(u)} w_i z_i, \text{ where } \sum_{i \in V(u)} w_i = 1 \tag{1}$$

$$\begin{pmatrix} w_1 \\ \vdots \\ w_{N_n} \\ \lambda \end{pmatrix} = \begin{pmatrix} \gamma(h_{1,1}) & \cdots & \gamma(h_{1,N_n}) & 1 \\ \vdots & \ddots & \vdots & 1 \\ \gamma(h_{N_n,1}) & \cdots & \gamma(h_{N_n,N_n}) & 1 \\ 1 & \cdots & 1 & 0 \end{pmatrix}^{-1} \begin{pmatrix} \gamma(h_{1,u}) \\ \vdots \\ \gamma(h_{N_n,u}) \\ 1 \end{pmatrix} \tag{2}$$

2) Spatial heterogeneity: In contrast, spatial heterogeneity refers to the variation in the sample distribution across the study area [15]. It assumes that spatial data samples often do not follow an identical distribution in the entire big area, thus the learning model from the entire area may indicate poor predictions for some specific areas. To solve the above problem, the researchers investigate several kinds of solutions, including integrating spatial coordinate features into data mining, multi-task learning and so on. Multi-task learning is a common machine learning solution for heterogeneous data, and can group learning samples into several different learning tasks. To solve spatial heterogeneity problem, multi-task learning can be adopted by learning several local models based on specific local data to enhance spatial prediction [9]. The heterogeneity can also be described as spatial non-stationarity or spatial anisotropy.

B. Spatial clustering

Spatial clustering groups similar objects based on various measures such as distance, connectivity, or their relative density in space. As a part in *unsupervised learning* in machine learning and concept hierarchies, the cluster analysis in statistics aims to find interesting structures or clusters from data based on natural notions of similarities without using much background knowledge. Spatial clustering can be categorized as partitional clustering, density-based clustering, and grid-based clustering. Partitional clustering separates the whole data set into a set of disjoint clusters. For example, the popular K-means methods partitions n data points into k clusters where each data point belongs to the cluster with the closest *mean*. Density-based clustering groups the points that are closely packed together by neighborhood relationship, and marks outliers that lie alone in low-density regions. The most popular density-based clustering method is DBSCAN which finds groups of points that satisfy the following condition: given a radius (Eps) a cluster at least contains a minimum number of objects ($MinPts$), and all the points are density-reachable conditions. However, its computation complexity is high because it processes each data point individually. Grid based clustering differs from the above two in that it assigns a value in each cell of the grid covering several data points. Thus Grid-based clustering is generally quite efficient in big data processing so long as the grid cell is not too small.

C. Spatial Outlier Detection

Spatial outlier detection [16] discovers the data which are spatially distinct from their surrounding neighbors, such as the red cross mark in Fig. 2. In many real applications using geographic information, such as transportation, public, safety, and location based services [17][18], spatial objects cannot be simply abstracted as isolated points, because different properties, such as boundary, size, volume, and locations among the spatial objects, lead to neighborhood effects.

Spatial outlier detection is designed to discover some unexpected, interesting, and useful spatial patterns for further analysis. Spatial objects can be seen as spatial points with attributed values (non-spatial value such as temperature). The spatial outliers can be formularized as follows.

Given a data mapping function ($f : X \to \mathbb{R}$) from the data set $X = \{x_1, \ldots, x_n\}$ to the real number set $\mathbb{R}$, and a neighbor evaluation function g, which is to evaluate the neighbor information of each data $x_i \in X$ by averaging $NN_k(x_i)$ (the k nearest physical neighbors of x_i), and further a comparison function h, such as $h = f - g$ or the ratio $h = f/g$. For a set of data $i = 1, 2, \ldots, i, \ldots, n$, a data unit i is a spatial outlier if h_i is an extreme value of the set $\{h_1, h_2, \ldots, h_n\}$.

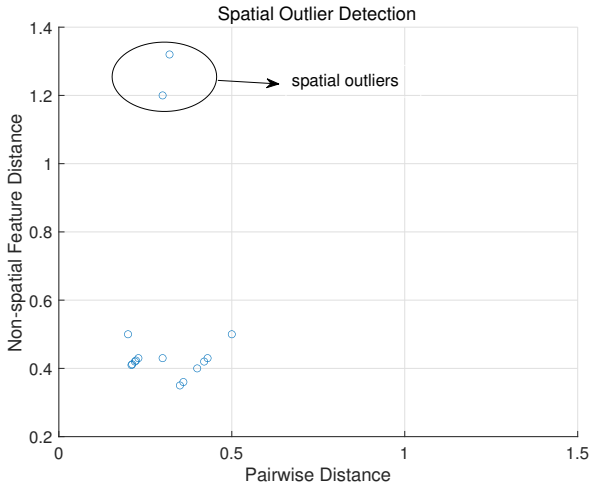


Fig. 3. Spatial Outlier Detection based on a Variogram Cloud.

There are two main approches for spatial outlier detection. One is graphic-based approach, such as variogram clouds in [19][20] and pocket plots in [19] [21]), which visulize data first and then find the corresponding spatial outliers. The other is quantitative tests, such as scatterplots [22][23] and Moran scatterplots [24]), which show the spatial association or non-association of spatially close objects.

For example, Bipartite tests are typical multi-dimensional spatial outlier detection methods, which use the spatial attributes to characterize location, neighborhood, and distance. Then it further find a spatially referenced object to its neighbors based on non-spatial attributes such as temperature. A Variogram Cloud can be used for spatial outlier detection [25] as shown in Fig. 3. This plot shows two pairs are above the main group of pairs are possible related to spatial outliers, where x axles represents the spatial distance of pairs of points, and y axles shows non-spatial features. Several other methods used for spatial outlier detection[26] includes kNN, graph-based method and so on. Also different statistic measures were used for representing spatial distance, e.g., Z-Score, Mahalanobis distance, LOF-based measure.

D. Spatial Co-location

Spatial co-location discovery [27] finds the subsets of features that are frequently located together in the same geographic area as shown in Fig. 2 (white and yellow circles). Spatial co-location mining problem can be formalized as follows [28]: Given a set F of K types of spatial features $F = \{f_1, f_2, ..., f_K\}$, and their instances $I = \{i_1, i_2, ..., i_D\}$, where D represent the amount of data. Each instance of data i_i is represented by a vector $< id, i_k, loc_i >$, including its id, a type of spatial feature i_k and its location. Spatial co-location mining is to efficiently find the colocated spatial features in the form of features or rules.

Co-location pattern discovery can be mainly classified into two categories: spatial statistics based and data mining based. The spatial statistics based approaches use various measures to characterize the relation between different types of spatial events (or features), whereas data mining methods find frequent and meaningful relations, positive associations, and stochastic plus asymmetric patterns among sets of items in a large transaction database and a spatial database. Measures of spatial correlation[28] include *cross*-K function with Monte Carlo simulation, mean nearest-neighbor distance and spatial regression models.

Data mining approaches can be further divided into a clustering-based map approach and association rule-based approaches, or their integration [29]. Take the vertical-view approach in [29] as an example to describe the process for spatial co-location mining, it works as follows: 1) Map the spatial data into K layers where K represents the types of the feature; 2) Find spatial cluster for each layer of point data; 3) segment all the layers with a finit number of regular cells, say M cells; 4) construct a $M \times K$ relational table with the binary $\{0, 1\}$; 5) Apply association-rule mining to the table to discover spatial co-location patterns.

III. Using Social Media for Emergency Situational Awareness

Recent years have seen an increased interest by the research community in using twitter data for situational awareness in the emergency and disaster contexts. Event detection is arguably the most active subtopic, where the objective is to detect new events from a real-time twitter stream. A typical approach for event detection is to define one or a few keywords (e.g., earthquake) of interest and to track if there are temporal bursts of the keywords' use in the tweets [51]. Extensions of this approach include general-purpose detection systems that track a large number of keywords [52], phrases [53] or detect emergence of clusters of similar tweets [54]. More recently, researchers started paying more attention to the spatial aspect of events [55]. For example, [56] considers burstiness of term "earthquake" in both time and space to detect spatial clusters of tweets that are candidates for an earthquake event. The unsupervised approach for event detection can be further enhanced by adding a classifier that is trained on previous events to recognize which bursty clusters are events and which are not [57].

Once an event has been detected, either by the aforementioned approaches, or by more traditional ways such as news or emergency department announcements, another commonly addressed research challenge is using twitter data to gain situational awareness. Considering the state of the art in natural language processing and data analytics, it is still not possible to build a fully automated system that could provide actionable knowledge to the responders. Instead, the emphasis has been on summarizing and visualizing disaster-related tweets to help human responders to quickly grasp the vast amounts of generated information. Representative systems are Senseplace2 [58], a visual analytics system that allows an operator to enter a query (in a form of a term or a hashtag), look at the map to observe where is the keyword common, click on a specific location, and view individual ranked tweets from the selected location, and Twitinfo [59], a tool that allows an operator to browse a large collection of tweets using a timeline-based

TABLE I
COMPARISONS OF SPATIAL DATA MINING ALGORITHMS

Types of Spatial Mining	Categories		Models and Algorithms	Details
Spatial Prediction	Spatial Autocorrelation	Special Contextual Information	Reference [10]	Spatial relation information
			Reference [11][12]	Spatial contextual information from multi-source
		Model-based Spatial Prediction	Reference [13]	Markov random field
			Reference [14]	Gaussian process (Kriging)
	Spatial Heterogeneity		Reference [30]	Location-dependent learning
			Reference [31]	Multi-task learning
Spatial Clustering	Density-based Approaches		Reference [32]	DBSCAN
			Reference [33]	Adaptive DBSCAN for massive data analysis
			Reference [34]	Enhanced DBSCAN with similarity measurement
			Reference [35]	Grid-based DBSCAN for fast processing
	Hierarchical-based Approaches		Reference [36]	Hierarchical clustering based on topology learning to reduce the computation complexity
			Reference [37]	Time-hierarchical clustering
			Reference [38]	Hierarchical aggregation for distributed clustering
			Reference [39]	Parallel hierarchical clustering
	Partition-based Approaches		Reference [40]	Partition-Density joint clustering
			Reference [41]	Adaptive partition for spatial analysis
Spatial Outlier Detection	Graphic-based Approaches		Reference [19][20]	Variogram clouds based solutions
			Reference [24]	Scatter-plot based solution
	Statistics-based Approaches		Reference [42]	Multi-attribute based solution
			Reference [43][44]	kNN-based solution
Spatial Co-location Pattern Discovery	Data Mining	Visualization-based	Reference [45]	Visualization and then mining
		Parallel-based	Reference [46]	Parallel solution on GPU
		Dynamic	Reference [47]	To solve dynamic varying patterns
	Spatial Statistics		Reference [48]	Cross-K function
			Reference [49]	Cross nearest distance
			Reference [50]	Q-test

display, drill down to sub-events, and explore via geolocation, sentiment, and popular URLs. More advanced visual analytics systems also include capability to cluster disaster-related tweets [60]. There are also summarization systems that have capability to classify tweets into some of the predefined categories [61]. As a representative system of this type, in [62] the authors categorize disaster-related tweets into one of a few predefined categories (e.g., personal, informative, other) and subtypes (e.g., caution, casualties) using a classifier which uses text features such as unigrams or bigrams and which is trained on a manually labeled data set of historical tweets. We should also mention that there are systems that integrate data from multiple sources, such as Ushahidi (www. ushahidi.com) [63], a platform that leverages Web 2.0 technologies to integrate data from phones, Web applications, email, and social media sites to provide publicly available crisis maps.

Other than Twitter, other social media platforms such as Facebook, Wikipedia, Flickr etc. are also used in different disaster scenarios. After the Sichuan earthquake in 2008, the use of Tianya (a popular online forum in China) is studied as a forum for online discussions on earthquake-related topics [64]. Reference [**?**], [65] have studied the peer-to-peer communication from a variety of other platforms especially Facebook after the Virginia Shooting in 2007, and southern California wildfires in 2007 [66]. During the 2013 Colorado Floods, different flood-related communications in Facebook and Twitter are examined in [67], [68].

When processing and analyzing such social media data for event detection and situational awareness, one should be aware of a multitude of challenges. One issue lies in varying credibility, reliability, and quality of twitter data. For example, geotagging of tweets is nontrivial because of the uncertainties in their location and timing [69]. For example, only a small fraction of tweets typically has an accurate GPS-quality location and there could be a significant and unknown lag between an event occurrence and its mention. Another challenge is that there are significant differences in the dynamics, spatio-temporal extent, and impact of different disasters, coupled with the ever changing use of social networks such as twitter. As such, one should be cognizant of these issues when performing titter data analysis and transferring knowledge from previous disasters.

IV. CHALLENGES IN INTEGRATING BIG DATA WITH EMERGENCY NETWORK

We envision the network configuration to be dependent on both automated data collection from the smart phones via specially designed emergency apps and human directed communication such as phone calls and social media. Acknowledging that twitter has established itself as the premier human communication mechanism during disasters and the wealth of publicly accessible disaster-related twitter data, we consider integration of the twitter-based information for the purposes of situational awareness and network configuration. While the objective of situational awareness is to inform a wide range of emergency responders, here we specifically focus on situational awareness that facilitates decisions about network configuration. We define network disturbances as any situation that negatively impacts ability of nodes to send and receive data. For example, this might include situations when the network demand exceeds capacity due to bursts of activity when the bandwidth is compromised, or when the portions

of the network are down or disconnected. Situational data available from different sources (such as social media) can be useful in detecting the disturbances, understanding their severity and causes, and take adaptive actions to recover from such network damages/failures. Below we list some of the key challenges regarding deriving situational awareness from disaster related data analysis.

A. Spatio-temporal Uncertainty in Available Data

One specific challenge in using the user data is their origination. Some mobile users may disable their location in their devices, or the location information from the base-stations may not be precise enough due to localization inaccuracies. Data originated from different locations during a disaster may have varying *data quality*, *precision*, and *accuracy*. For example, the location of the tweets is important as the tweets originating around the disaster area are more important and contain first-hand information. However, the users may not wish to share their location. The timing is important since we wish to consider it in dynamic network reconfiguration decisions. Unfortunately, tweets may refer to past events without precise time information. Thus, the challenges are both in terms of estimating location and time as accurately as possible, and using the available information suitably.

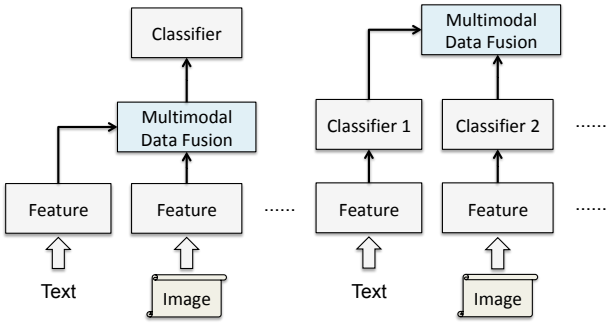
B. Data Ambiguity and In-homogeneity

The data generated by various sources is often non-homogeneous in nature, incomplete, or ambiguous. Data obtained from various social media is also prone to inaccuracies and inconsistencies. For example, the first hand twitter reports originating from the affected area are likely to be most useful in situation awareness and hence network configuration; however, because of potential damages to the Internet infrastructure in the affected area, such first hand tweets may be quite sparse. On the other hand, due to the popularity of twitter during disasters, much of the information generated by human-to-human communication media (e.g., word or mouth, landline phone, broadcast media such as radio or TV, etc.) increasingly ends up on twitter from non-disaster areas. In general, the origin of these tweets can be from anywhere; however, the regions around the disaster area are likely to be the most relevant. This brings in issues of *bigdata* since one must sort through a huge number of tweets in order to find the relevant ones. In fact, even in the general disaster area, most tweets may not be relevant for disaster response or network evolution and must be filtered out in real-time.

C. Multimodal data fusion

Yilang: Please expand this section...

Generally, information about the same situation can be collected from different types of resources, e.g., texts and images in Twitter and Instagram. For each kind of detector, it is represented as a modality, and it is rare that a modality can cover the complete information of the situation. Multimodal data fusion is required to integrate the information into a comprehensive view. Generally, there are two approaches for multimodel data fusion: feature-level fusion and decision-level fusion, also known as early fusion and late fusion. Fig. 4 illustrates these. Feature-level fusion merges features from different types of data resources together before classification. For example, in [70] a Topic Graph is proposed to integrate features from different modalities together, which is constructed by nodes (i.e., features or words) and edges among the nodes (i.e., correlation of features). For decision-level fusion, generally a classification score is given to each modality and the maximal one is treated as the final classification result. In [71], both of these methods were evaluated with text, video and audio contents, and the results from the both approaches increase around 10% precision comparing with the result with the single data resource. Most recently, deep learning is adopted to achieve model-based fusion for multi-modal data fusion. For example, strong modalilities can be automatically selected to achieve high accuracy of situation detection in [72].



(a) Feature-level Fusion. (b) Decision-level Fusion.

Fig. 4. Two Major Types of Multimodel Data Fusion

D. A Light Weight Architecture for Data Analysis

The emergency data is widely distributed on Internet – either published on the Web or gathered from distributed devices. Different from the enterprise-oriented solutions, such as AWS managed services, here we illustrate a light-weight architecture by using open source tools. This architecture is shown in Figure 5 and can be deployed either in the cloud or on-premises servers and local devices.

Since the volume of the emergency data is incremental, it requires automatic data crawling, storage, and transformation to standardize the data stream for a generalized data analysis. The data transformation part generalizes the data a matrix, for example, to take the imagery data pre-processing by calibrating, cropping and resizing the raw images which were generated from different devices, and to take the tabular data pre-processing by queries (SQL/NoSQL) from database or data frame filtering or combination. Since the data may contain ambiguous description about emergency, it is necessary to teach the machine the meaning of emergency in the big data stream. Therefore, we could label part of the data by categorizing or annotating them to be further used as training data set.

By using the container technology, we could compile such an architecture into a Docker/Kuberneties image, and deploy it in multiple devices. To support the emergency response,

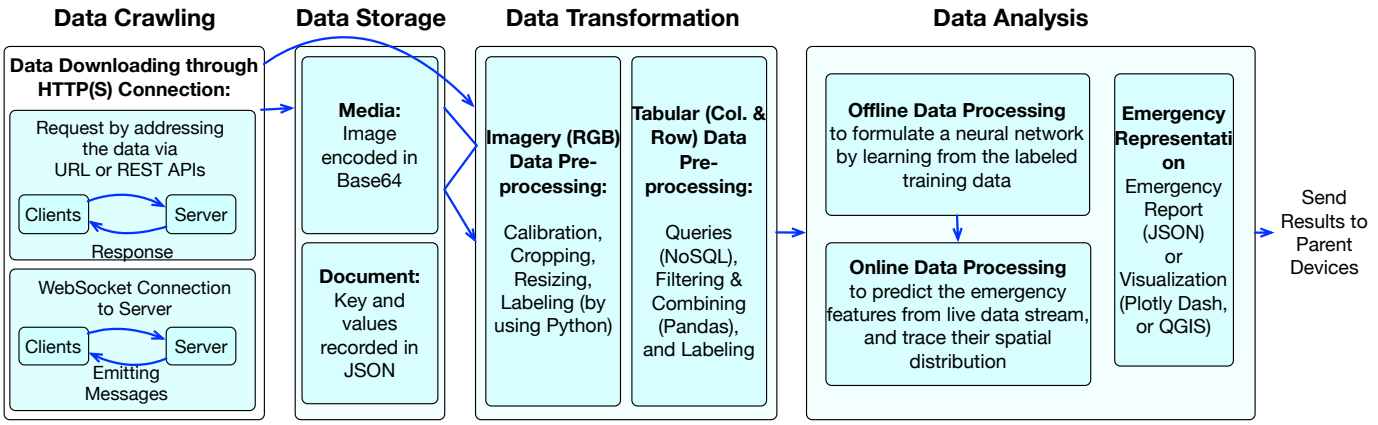


Fig. 5. A Lightweight Architecture for the Offline and Online Data Analysis

an online algorithm can continuously recognize specific emergency features from the images or text data that were received or crawled by the host device. It can also track the spatial distribution of the detected emergency through an incremental spatial clustering[73]. The offline data processing formulates specific AI models by learning from the training data. It runs in lower priority on the devices which has satisfied computation resources, and outputs the trained models to be deployed on devices for online data processing. The output of online processing outputs a JSON data which contains the spatial distribution of recognized emergency features and the original text or images. The output JSON could either be visualized locally for decision makers, or to be sent to parent device. The online data processing in the parent device could do the recognition of other specific emergency features, and to be fused with the previous results.

E. Spatial Analytics During Evolving Disasters

Even though spatial analytics have been studied for a long time, there are still new challenges when considering social big data generated in disaster scenarios. This is because in an evolving disaster scenario, the usage pattern and user's behavior changes over time, sometimes rather rapidly. Also, the incoming user data from the crowd is highly dynamic and the observed situation is intermittent, which becomes an obstacle when trying to achieve reliable data analysis to support decision-making after a disaster occurs.

To address the evolving spatial analytics, the authors in [73] have introduced an information decay based spatial clustering. The intuition behind this information decay factor is that in a disaster scenario the disruptions over a region cannot be satisfied immediately, and thus the importance of such information does not disappear instantly, and instead decays gradually over time. Decay model has been investigated in the spatial clustering for streaming data, i.e., evolving clustering. As the data comes in a streaming way, small clusters are first temporarily created to organize the received data in the clustering process. However, the existing work only applies decay model to the clusters, but not for each point data, which will affect the accuracy of situation representation.

F. Utilizing Non Geo-Tagged Tweets

Another key challenge of using Twitter data is to scarcity of the number of geo-located tweets, which typically varies in between 0.42% to 3.17% [74]. Utilizing the non geo-tagged tweets can also provide useful information if they can be related approximately to their origin. Some works [74], [75] have proposed to determine "local" words by exploiting the geographical distribution of the words in tweets over a region. Formally speaking, local words are the ones with high local focus and fast dispersion, i.e. they are frequently used at some central points and drop off in use rapidly as we move away from the central points [75]. For example *tube* is more frequently used in London than other places. By exploiting such distribution around 50%–87% of the tweets can be located within few tens of kilometers [74].

Geoparsing is another well-known technique for extracting the locations (also known as *toponyms*) inside a text, which can be exploited for deriving locations from non geo-tagged tweets. Using natural language processing techniques, locations in the level of streets or buildings can be derived, that can help identifying the origin of a particular situation. For example, a new tweet like "Having a moderate earthquake 5.8 mag here in Raoul Island, New Zealand" – provides sufficient location information to locate the origin of the incident. Related literature on geoparsing can be broadly divided into two categories [76], namely toponym recognition and toponym resolution. Toponym recognition techniques [77] extract single or consecutive words from texts and match them to a comprehensive set of pre-existing set of toponyms. The key limitation of these techniques is the ambiguity of the toponyms, as many location names have multiple occurrences worldwide. To overcome this limitation, toponym resolution based approaches [78] use different spatial indicators such as time zones, use location field, and other textual clues for ensuring more reliable location estimates.

Even in cases where the geo-locations are not found, the contents of the tweets can also provide important information regarding the situation. Different natural language processing techniques for keyword analysis to determine relevance, specificity (or fuzziness), and importance of the content can be explored to determine the usefulness of such tweets, whereas

the irrelevant ones can be filtered out. These substantially filtered, prioritized set of tweets can then be provided to human experts involved in situation monitoring, to determine how the infrastructure damage/repairs, movement of people, and potential communications needs are changing, and consequently how the relief assets (including those that support emergency communications network) should respond to them.

G. Big Data Analytics in a Fragile Communications Network

After collecting the raw data from various sources, big data platforms (such as Hadoop) need to sort through a huge amount of data in order to extract the most relevant ones. In fact, even in the general disaster area, most social media data may not be relevant for disaster response or network evolution and must be filtered out in real-time.

In the aftermath of a disaster, the communication systems can be wiped out which makes distributed processing challenging. A fragile and disruptive emergency communication network brings new challenges for spatial big data analytics since big data is often analyzed in a cloud center to reduce processing time, and the transmission delay from user's devices to the cloud could become dominant. This requires tradeoffs between local processing at the devices, intermediate processing at some edge computing nodes, and final processing in the cloud. However, distributing processing among these heterogeneous levels with varying storage, processing, and communications capabilities becomes quite challenging.

V. Situation Awareness in Different Disaster Applications

In the following, we study the literature available on situational awareness a wide range of disaster applications.

A. Situational awareness in communications networks

We define network disturbances as any situation that negatively impacts ability of nodes to send and receive data. For example, this might include situations when the network demand exceeds capacity due to bursts of activity due to inadequate bandwidth, or when portions of the network are down or disconnected. Network performance related data can be useful in detecting the disturbances, understanding their severity and causes, and take adaptive actions to recover from such network damages/failures. Given the high degree of robustness and redundancy of the public communications networks, large scale network failures are very rare, as evidenced by Kumamoto earthquake and hurricane Sandy events discussed in this paper. Also, if a large network outage does occur, it would decimate the social media traffic in the affected area; therefore, we do not focus on such events in this paper.

1) Detection of network disturbances: Detection of network disturbance can be performed by analyzing the spatial scan statistics [79] and its many extensions [80], [81], [82] to detect spatial, temporal, or spatial-temporal areas where the user's activity is different from the norm. Network abnormality or anomalies in a cellular network can be identified by examining the call records of the users in a region, their locations, mobility patterns etc. Similar anomalies can also be identified from the user's tweets that originate from the region of interest and their spatio-temporal behaviors. Spatial outlier based scanning can be applied in this context for spatio-temporal anomaly detection.

Spatial scan based algorithms have been traditionally used for disease mapping where the objective is to find regions containing significantly increased incidence of disease symptoms, but has since found many other applications as well. The spatial scan algorithms scan the spatial-temporal region of interest to find the most significant subregion and report its statistical significance. A notable application of scan statistics in the domain of social networks is analysis of spatial distribution of 803 flickr tags in the Bay Area [83] in order to distinguishing between place and event related tags. The key challenge for analyzing such scan statistics is computational because there is potentially a huge number of terms that could be tracked, which may require distributed processing across multiple clusters. For social media generated data, another challenge is to account for geolocation and temporal uncertainty in such data, and at the same time account for the expected mobility of the mobile users.

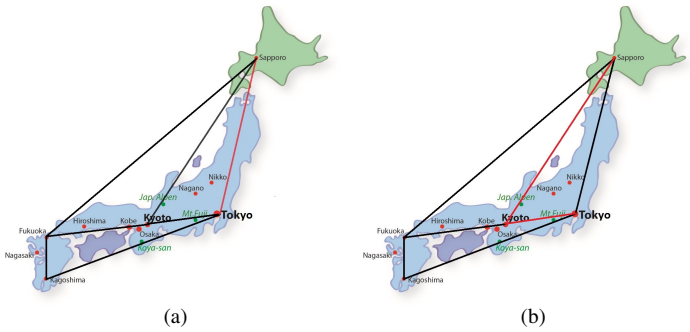


Fig. 6. Adaptation of the routes after sniffing potential congestion between Tokyo and Sapporo. (a) Route-1 is the direct route from Tokyo–Sapporo, whereas (b) Route-2 goes through Kyoto.

2) Congestion and traffic control: Big data analytics can be beneficial for traffic monitoring in both wireless and wired networks. Such analytics can be used to identify congestion in the communications infrastructure immediately before, during and after the disaster. Often the communications network experiences congestion when the event is imminent and during the event period. The reason for congestion could include both damage to and high demand for computing and communications. It is important to understand and manage such congestion while also backing up the state of potentially affected computing infrastructures to remote locations. Congestion remains crucial after the onset of the event related disruption. Social media data such as user tweets can also address the issue of characterizing failures in the network [84] – i.e., user complaints about the network functionality or slowness. Examples of such tweets are as follows [85]: *"I cannot get through to Miyagi...I'm worried."*, or *"I'm in Shibuya now. I cannot get through."* etc. Spatial clustering based schemes can be used to identify those regions where such complaints are significantly higher than in other regions.

In the context of wired networks the authors, in [86] have used the data plane programmability of the Openflow switches

to adopt more flexible network control. For example in Fig. 6 the costs of the routes are increased based on the tweets complaining about the network issues. The Openflow controller can switch the routes whenever it sniffs a link congestion. In Fig. 6 the route-1 in between Tokyo and Sapporo is switched to route-2 after the controller sniffs a potential congestion on route-1. The Openflow switches can also be reprogrammed for content based bandwidth control. For example, in case of potential network congestion, packets related to SMS, email or voice communication can be given higher priority than the video based communications.

3) Finding network isolations and resource allocation: Another application of situational awareness is to identify isolated regions that are functional but disjointed from the remaining network. In a cellular network, tracking the call records and the usage densities are good indicators of finding the network availability. Spatial outlier detection based techniques are very useful in such contexts, where the objective is to find the regions from where the usage is significantly lower as compared to the surrounding regions. However, this is a very challenging problem because of the need to analyze the available data over a large region encompassing the isolated area. Notice that such isolations can also happen due to other reasons, such as drainage of the smartphone batteries due to lack of power and mobility or evacuation of the users from a certain area. Careful analysis of the call density along with other useful information from multiple sources (such as evacuation notice) can be utilized for finding such network isolations.

Upon finding the isolated, disconnected regions, a variety of emergency equipment such as WiFi access points, satellite gateways, replacement cellular base stations, etc. mounted in fixed places or on Emergency Communication Vehicles (ECVs) can be deployed to bring the connection back. Movable base-stations or access points mounted on drones and balloons can also be deployed for meeting the communication gaps [87]. As the resource requirements in a disaster scenario change over time, spatial prediction of the user density and usage patterns are needed before such deployment operations to avoid further disruption and performance fluctuations.

B. Big Data Analysis for Power Outage Detection

Real time situational awareness for detecting power outages from social media data has received interest in recent years. Reference [88] have used key words searching to collect power outage-related tweets. They have developed a modified approach of Kleinberg's burst detection algorithm to promptly detect the power outages from the tweets. In [89] the authors have proposed a supervised Latent Dirichlet Allocation (sLDA) to detect power outages. To overcome the limitations of 140 character limit of the tweets, the authors have used a supervised topic modeling with text-rich heterogeneous information network. In [90] the authors have studied the reported cases of power outage related tweets during Hurricane Sandy. They have also proposed a k-means clustering scheme for the efficient allocation of power resources based on the available tweets. In [91] the authors have analyzed the brightness change in the satellite data along with the density of power outage related for identifying the severely impacted areas. The study have shown that Twitter data fused with satellite imagery can identify power outage information at a street-level resolution. In [92] the authors have used the key textual descriptions of power outages to filter the relevant Tweets, and built a predictive model that identifies those Tweets referring to real power outages. The procedure has been field tested on the users in real industrial settings; the results show that more than 93% of all the power outages detected by the scheme actually referred to the real outages. In [93] the authors have separated the tweets into power outage, communication outage and both power-communication outage related events by analyzing popular words, length of words, hashtags and sentiments that are associated with these tweets. The study has claimed that using simple classifiers like such as boosting and support vector machine can successfully classy the outage related tweets from the unrelated ones with close to 100% accuracy. The study has also claimed that by employing transfer learning models such as Bidirectional Encoder Representations from Transformers (BERT), different categories of outage-related tweets can be classified with an accuracy close to 90% in less than seconds of training and testing time.

C. Big Data Analysis for Event Detection during Natural Disasters

Social media data for situational awareness in crisis scenario are discussed in [105], [106], [107]. In [94] the authors have analyzed tweets regarding resource needs and resource availability for the efficient management of post-disaster operations, using supervised classification and unsupervised pattern matching and information retrieval approaches. The authors have conducted experimental study on tweets posted during the Nepal earthquake in April 2015 and the earthquake in Italy in August 2016. The study shows that classification approaches perform better if good quality training data are available from prior events, whereas in the absence of such training data, unsupervised retrieval methods outperform supervised classification approaches. In [95] the authors have proposed a Deep Neural Network (DNN) to identify informative tweets and classify them into topical classes. They have also proposed an online stochastic gradient descent based algorithm to train the DNNs in an online fashion during disaster situations. Reference [96] has provided a comparison between matching-based [108], [109] and learning-based [110], [95] approaches for effectively identifying relevant messages from matching keywords and hashtags in social media data. Learning-based approaches typically build a model from a set of labeled tweets, whereas matching-based approaches search the tweets having relevant keywords and hashtags. In [97] the authors have proposed an Integer Linear Programming (ILP) technique that generates summaries of big volume of twitter messages around some identified sub-events, that helps crisis responders to fulfill their information needs. Reference [111] has generated *verified* summaries from the information posted on Twitter during disasters. Enhancing real-time situational awareness through filtering and summarization of social media

TABLE II
REPRESENTATIVE STUDIES OF BIG DATA ANALYSIS FOR EMERGENCY EVENT DETECTION

Types	Key points	Representative Works	Details
Situational awareness in comm. networks	Spatial scan related analysis for finding the network disturbances, congestion and network isolation	Reference [84], [85]	Characterizing network failures from user complaints about network functionality or slowness
		Reference [86]	Used data plane programmability of the Openflow switches to adopt flexible network control
		Reference [87]	Studied the optimal delay in a fog/edge-computing platform constructed by vehicle-based movable & deployable ICT resource units
Power outage detection	Situational awareness for detecting power outages from social media data using keyword searching	Reference [88]	Developed a modified approach of Kleinberg's burst detection algorithm to promptly detect the power outages from the tweets
		Reference [89]	Developed a supervised Latent Dirichlet Allocation to detect power outages
		Reference [90]	Proposed a k-means clustering scheme for the efficient allocation of power resources based on the available tweets
		Reference [91]	Shown that Twitter data fused with satellite imagery can identify power outage information at a street-level resolution
		Reference [92]	Developed a predictive model for identifying Tweets referring to real power outages
		Reference [93]	Separated the tweets into power outage, communication outage and both power-communication outage related events
Event detection during natural disasters	Analysis of tweets regarding resource needs, availability; filtering, summarization and classification of informative tweets from others	Reference [94]	Analyzed tweets regarding resource needs and resource availability
		Reference [95]	Developed a DNN to identify and classify informative tweets into topical classes
		Reference [96]	Compared matching-based and learning-based approaches for effectively identifying relevant messages from matching keywords and hashtags in social media data
		Reference [97]	Proposed an ILP to generate summaries of twitter messages
		Reference [98]	Enhanced real-time situational awareness through filtering and summarization of social media data
		Reference [99]	Developed a probabilistic spatio-temporal model to find the center of the target event
Data analysis during Covid pandemic	Analysis of Covid related tweets regarding public awareness,sentiment analysis, and classification of informative tweets from others	Reference [100]	Identified Covid related hashtags, along with the linguistic analysis of the tweets in different hashtag groups
		Reference [101]	Characterized public awareness regarding Covid by analyzing tweets in the affected countries
		Reference [102]	Implemented a neural network for sentiment analysis using multilingual sentence embeddings
		Reference [103]	Discussed the diffusion of Covid related information with a massive data analysis on Twitter
		Reference [104]	Proposed a multi-view clustering for analyzing tweets using clustering hashtags

data is reported in [98]. The authors have reported the study of twitter data during 2012 Sandy Hurricane from New York, Philadelphia, Boston, and Washington DC. In [99] the authors have devised a classifier of tweets based on some keywords, their numbers, contexts etc, and developed a probabilistic spatio-temporal model that can find the center of the target event location. The authors have implemented this approach as an earthquake reporting system in Japan; the study has shown that it can promptly detect 93% of earthquakes of Japan Meteorological Agency (JMA) seismic intensity scale 3 or more.

D. Big Data Analysis Related to Covid-19

Recently COVID-19 (also known as the novel coronavirus) has emerged as a world-wide pandemic that has affected several millions of people in the last few months [112]. The continuing evolution of COVID-19 in the USA had placed substantial stress on the resources necessary to deal with it. This includes hospital-beds, doctors, nurses, paramedics, personal protection equipment (PPE), ventilators, ambulances, police, test kits, testing supplies, common medications currently being prescribed, etc. In [100] the authors have identified Covid related hashtags, and have grouped them into six categories (namely general Covid, quarantine, panic buying, school closures, lockdowns, and frustration & hope. They have also presented a linguistic analysis of the tweets in different hashtag groups and have observed that words such as family, life, health and death are common across hashtag groups. In [101] the authors have characterized public awareness regarding Covid by analyzing tweets in the most affected countries. Specifically, the authors have examined the (a)

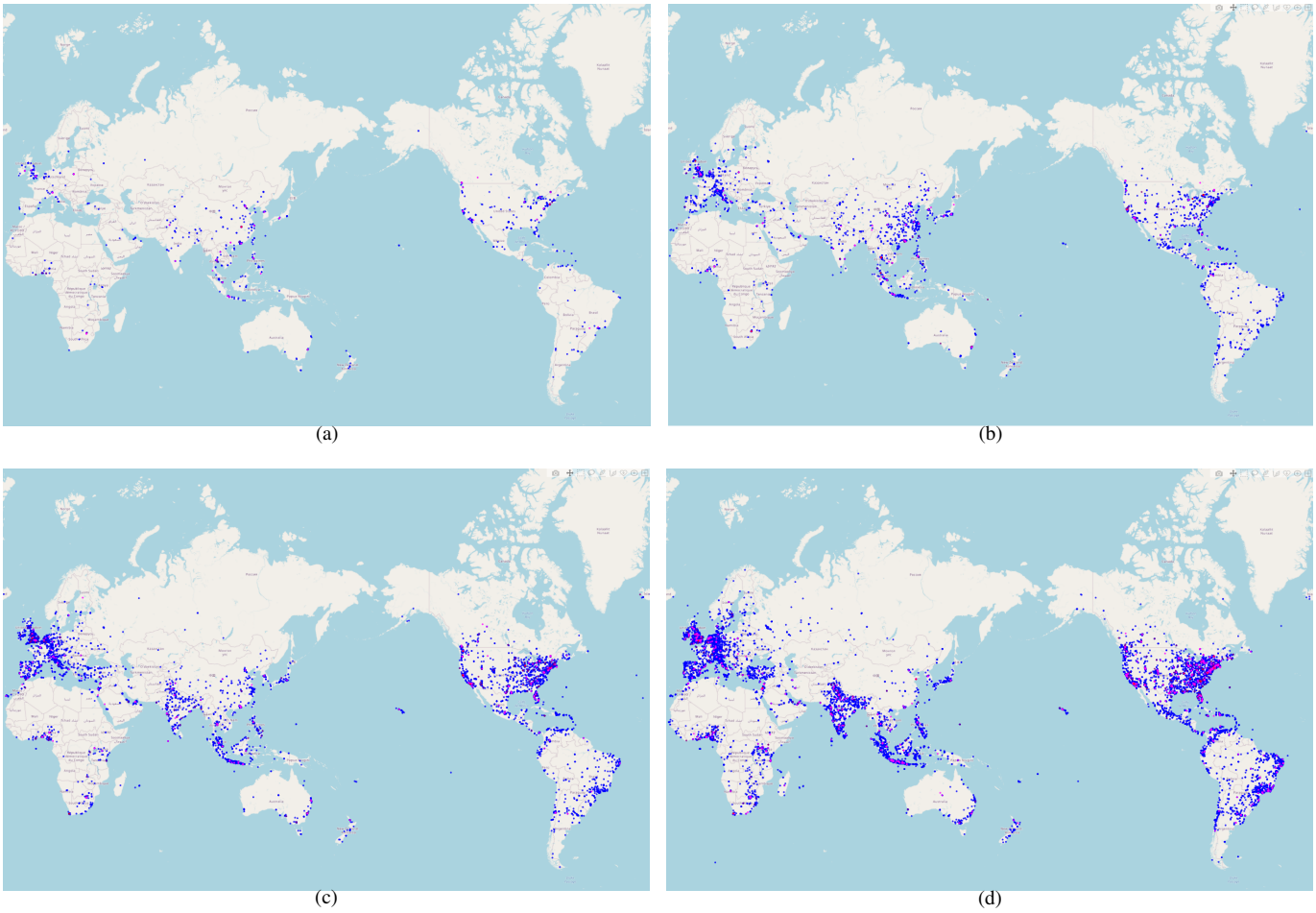


Fig. 7. Spatial densities of Covid-19 related tweets from months (a) January to (d) April, 2020.

temporal evolution of Covid related trends, (b) the volume of tweets and recurring trends in these tweets, and (c) the user sentiments towards preventive measures. In [102] the authors have implemented a neural network for sentiment analysis using multilingual sentence embeddings; they have observed that in almost all countries the lock-down announcements correlate with a deterioration of mood, which recovers within a short time span. The authors in [103] have addressed the diffusion of Covid related information with a massive data analysis on Twitter, Instagram, YouTube, Reddit and Gab. The have also fit information spreading with epidemic models characterizing the basic reproduction numbers for each platform. The authors in [104] have analyzed Covid related tweets using clustering hashtags, and have proposed a multi-view clustering technique which incorporates multiple different data types that can be used to describe how users interact with hashtags. A review of available methodologies for developing data-driven strategies to combat the Covid pandemic is discussed in [113], along with their difficulties and challenges. Representative studies on big data analysis on social media data for emergency event detection is summarized in Table II.

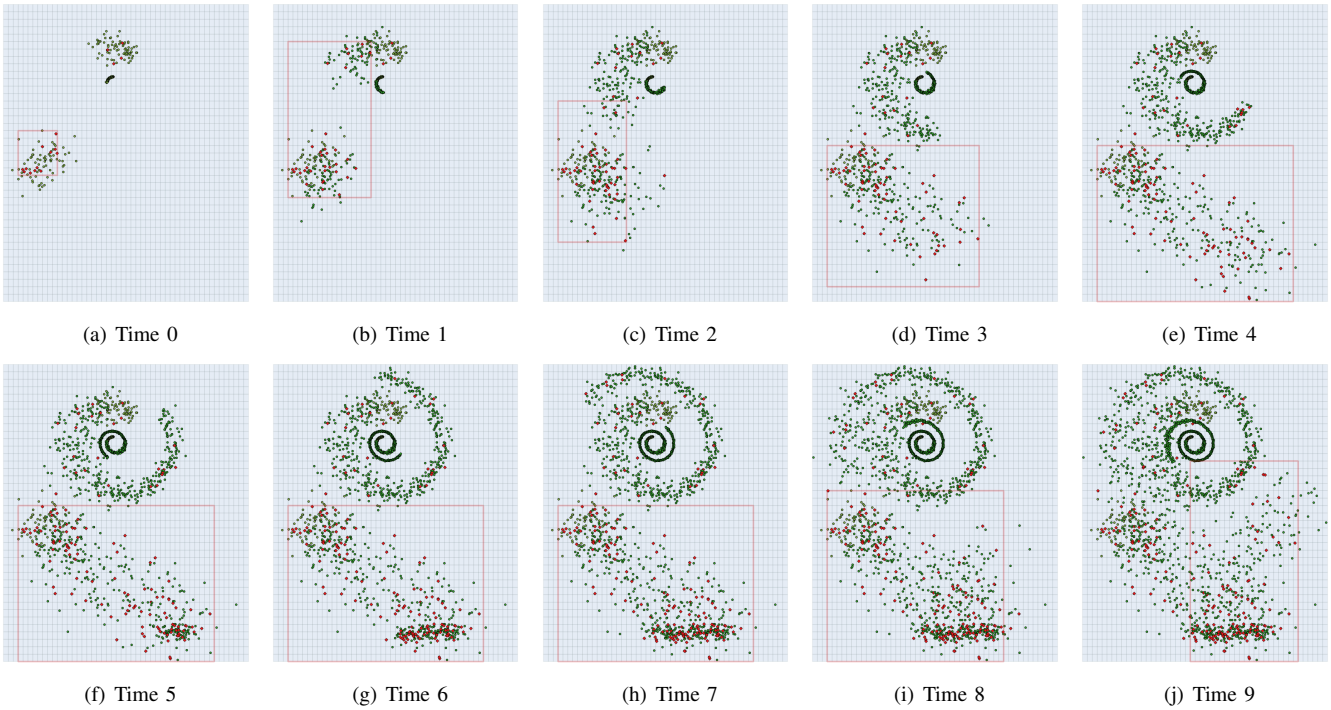
VI. Temporal Evolution of Spatial Features

It's been observed repeatedly that much of the data has a popularity pattern: Very hot when the data is generated, and then the popularity wanes. The data may become hot again. This trend is true for the social media data as well. In this context, we define the "information energy" of a tweet as the intensity of the tweet that is the highest power when a tweet originates, and then gradually fades over time. Information energy for a specific location can be accumulated with the other messages (or tweets) describing the same situation. Assume that the information energy for a point object p in spatial big crowd data at time instance t_c is denoted as $E_\epsilon(p, t_c)$. Also assume that the *temporal decay of the information energy* (TDIE) for each spatial data follows an exponential decay. That is,

$$E_\epsilon(p, t_c) = E_\epsilon(p, t_p) \cdot \eta^{-\lambda \cdot (t_c - t_p)} \tag{3}$$

where t_p denotes the time stamp when spatial data/object p appears, and η is the base of the exponential decay.

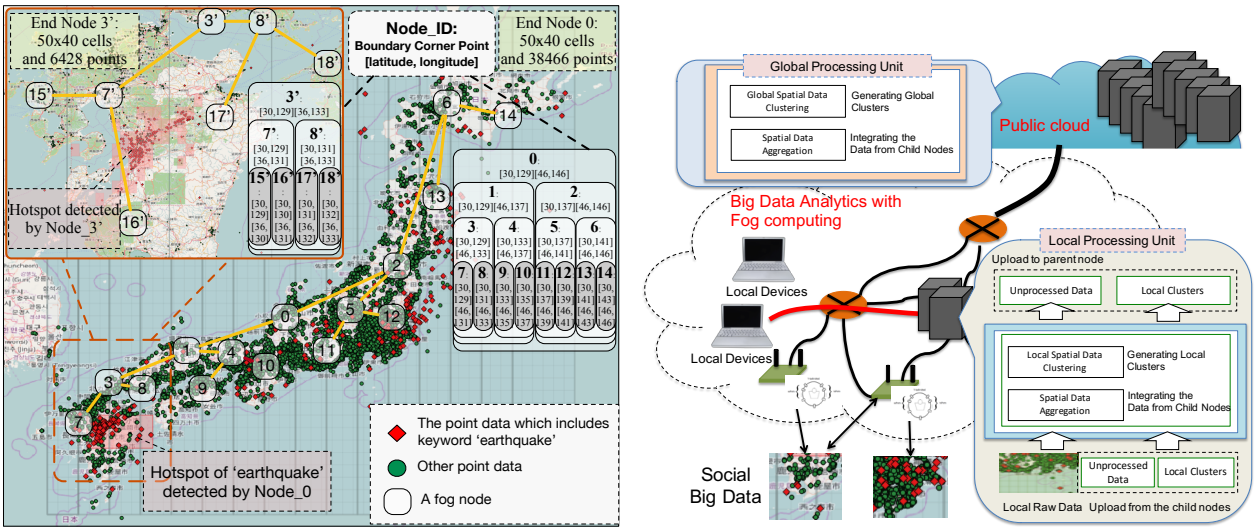
To find the spatial hotspots during an evolving disaster, we choose a density measure based on *Kulldorff's spatial scan statistic* [114], which is commonly used in finding the significant spatial clusters in case of an emerging outbreaks. With this the incremental spatial clustering in evolving disaster (or outbreak) scenario has two main functions, the spatial data aggregation (SDA) and spatial data clustering (SDC). The SDA handles decay and reinforcement of the information



(a) Time 0 (b) Time 1 (c) Time 2 (d) Time 3 (e) Time 4

(f) Time 5 (g) Time 6 (h) Time 7 (i) Time 8 (j) Time 9

Fig. 8. Position of Hot-spot ($\eta^{\lambda} = 2$) for Helix Movement Data Set.



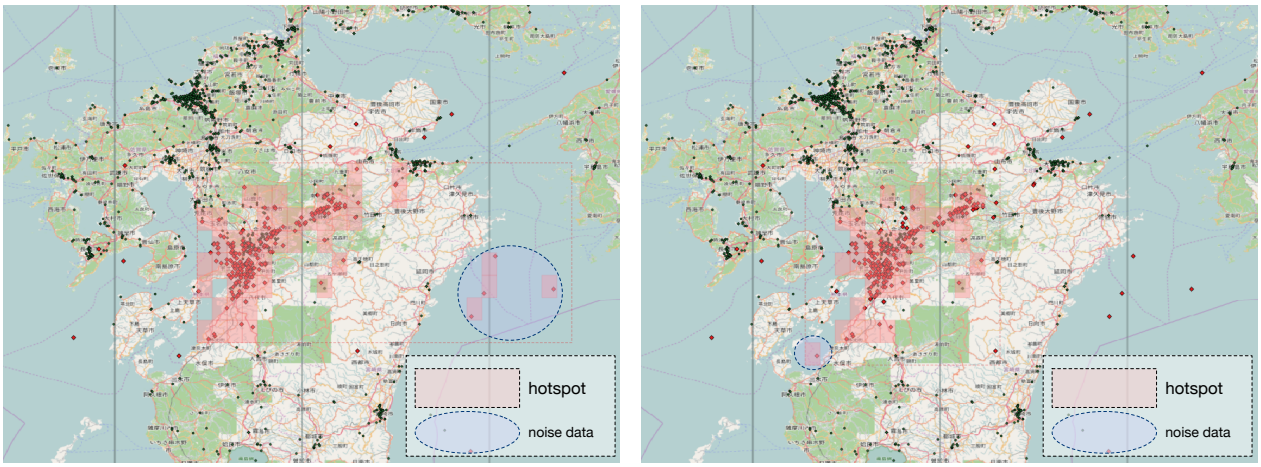
(a) Deployment of Edge Nodes in the Evaluation. (b) Edge Computing Based Spatial Analytics.

Fig. 9. Edge Computing for Spatial Clustering.

weight over regions. The SDC tracks the boundary and movement of the dense regions of the targeted evolving disasters.

We demonstrate such temporal and spatial evolution of tweets that are related to Covid-19 pandemic during that has affected millions of people around the world till date [112]. Fig. 7 shows the spatio-temporal evolution of tweets during the January-April 2020 timeline. From this figure we can observe that the temporal density variation of the tweets across different sub-continents have grown over time starting from January, which clearly matches its temporal spread. For example, during the February timeline, the spatial density of USA, East Asia and European countries were more as compared to Indian sub-continent, however, the cases in India started growing in March-April period. The tweet densities in Australian continent is quiet sparse which also matches with the small number of cases in that regions.

However, we noticed that the number of Covid related tweets with geo-tags are extremely sparse (Please give the statistics), thus, we could not conduct any spatial aggregation and clustering analysis in daily or weekly basis. We therefore simulate the incremental spatial clustering using a synthetic database obtained from [115]. The database is composed of several datasets that model the temporal evolution of the information contents in a two dimensional space. The datasets were generated by Gaussian distributions whose mean and/or variance change over time. We use the *"3C2D2400Spiral"*



(a) Clustering Results from Centralized Spatial Clustering.

(b) Clustering Results from Distributed Spatial Clustering.

Fig. 10. Clustering Results of centralized and distributed spatial clustering schemes.

dataset, which presents a helix alike movement of 3 clusters. These three clusters could be considered as three groups of population with dynamic ratios of the situation ϵ over the time series. We visually illustrate the effect of our incremental clustering on the helix movement dataset using Fig. 8 to illustrate the position, movement and coverage of the hotspots when $\eta^{\lambda} = 2$. From this figure we can observe that the movement of the hotspot is rather continuous, which is because of the use of TDIE concept. This continuous movement basically replicates the evolving nature of the disaster.

A. Role of Distributed Spatial Clustering

Offloading tasks from the cloud to the local computing resources can reduce transmission delay and cost. We demonstrate this by studying the effect of distributed spatial clustering scheme [116] on the Twitter data obtained during the Japan earthquake in 2016. Fig. 9 and Fig. 10 show some evaluation results from our recent studies in [116]. The evaluation was performed by simulating 18 processing nodes (defined as Edge nodes) deployed over Japan to process the Twitter dataset generated when Kumamoto-city suffered the earthquake in 2016. We have collected a total of 37 million tweets during and after the earthquake, among them 38466 tweets (around 0.1%) were geo-located. From these tweets we have extracted the ones containing terms of 'earthquake' both in English and Japanese language.

The edge nodes are placed in a tree-like architecture rooted at Node-0, as shown in Fig. 9(a). We divide the entire segment of the map into several sub-regions of rectangular grids. Each edge node processes the tweets of certain sub-regions; the leaf nodes cover the smallest sub-regions, whereas the root node covers the largest. For example in Fig. 9(a), Node-0 covers the rectangular region ranging from bottom–left coordinate [30,129] to top-right coordinate [46, 146]), and its children Node-1 and Node-2 cover sub-regions [30,129],[46,137] and [30,137],[46,146] respectively, and so on.

The maximum number of tweets were originated from the Kumamoto region, which is situated in the southeast of Japan. To make a detailed observation in this region, we deployed more edge nodes in the Kumamoto region, which is shown in top-left corner of Fig. 9(a). Node 3' is a cloned instance of Node 3, with different region of coverage; ranging from bottom-left coordinate [30,129] to top-right coordinate [36,137]. The entire Kumamoto region is sub-divided into 50×40 grids, having a total of 6428 tweets. We also observe that some tweets were originated from outside the land (in ocean, maybe on-board ships); we thus consider them as noise.

The distributed spatial clustering scheme consists of three modules: (a) a clustering module at each edge node based on the local tweets, (b) a data aggregation module to integrate the tweets obtained from the child nodes, and (c) an control module to coordinate the two functions. The overall architecture for the distributed analysis is shown in Fig. 9(b). Fig. 10(a)-(b) show the comparison of the centralized clustering scheme with the distributed clustering scheme, where the clustering is performed at each edge node, and then aggregated by its parent. From this figure we can observe that the distributed spatial clustering outcomes are pretty similar to that of the centralized counterpart. In Fig. 10 the hotspots are the clustered regions where the spatial density of the earthquake related tweets are higher than that of outside (measured from the Kulldorff's spatial scan statistics [117]).

VII. Conclusions

During disasters, the data relevant for situational assessment is generated from many different sources including social media use by the affected people (usually Twitter), direct communications with other, possibly unaffected users who put the information on the social media, and observations by the deployed monitoring infrastructure, etc. Different types of disaster related data may come from various sources: such as from the user's end, from the social medias or from the network operators. The data collected from these sources contains a lot of irrelevant or weakly relevant information, and it becomes necessary to use big data techniques to extract intelligence from them. Spatial information and context is crucial for this, and the paper focuses on several opportunities

and challenges of spatial big-data analytics with partial spatial information to extract situational awareness of the disaster. We hope this article will spur further research and results in solutions to many of these issues.

Acknowledgments

This research was supported by JST-NSF joint funding, Strategic International Collaborative Research Program, SICORP on Japan side, and CNS-1461932 award from NSF.

References

[1] F. Alam, F. Ofli, M. Imran, and M. Aupetit, "A twitter tale of three hurricanes: Harvey, irma, and maria," in *Proc. 15th Intl. Conf. on Information Systems for Crisis Response and Management, ISCRAM 2018* (B. Tomaszewski and K. Boersma, eds.), vol. 2018-May, pp. 553–572, 1 2018.

[2] V. Lorini, C. Castillo, F. Dottori, M. Kalas, D. Nappo, and P. Salamon, "Integrating social media into a pan-european flood awareness system: A multilingual approach," *ArXiv*, vol. abs/1904.10876, 2019.

[3] "Midwest farmers take to twitter to document flood disaster," May 2019.

[4] J. Yee, "Hurricane sandy: A chance to identify vulnerabilities, learn from the past, and increase future resilience." https://research.library.fordham.edu/environ_theses/5, 2013.

[5] Asian Disaster Reduction Center, "2016 Kumamoto Earthquake Survey Report (Preliminary)." http://www.adrc.asia/publications/201604_KumamotoEQ/ADRC_2016KumamotoEQ_Report_1.pdf. Accessed: March, 2017.

[6] "Hurricane Sandy 2012." https://www.huffingtonpost.com/2012/10/31/hurricane-sandy-new-york-city-power-outage-map_n_2050380.html. Accessed: January, 2018.

[7] S. Shekhar, P. Zhang, Y. Huang, and R. R. Vatsavai, "Trends in Spatial Data Mining," *Data mining: Next generation challenges and future directions*, pp. 357–380, 2003.

[8] J. Wang, Y. Wu, N. Yen, S. Guo, and Z. Cheng, "Big data analytics for emergency communication networks: A survey," *IEEE Communications Surveys & Tutorials*, vol. 18, no. 3, pp. 1758–1778, 2016.

[9] Z. Jiang, "A survey on spatial prediction methods," *IEEE Transactions on Knowledge and Data Engineering*, vol. 31, no. 9, pp. 1645–1664, 2018.

[10] A. McGovern, N. Troutman, R. A. Brown, J. K. Williams, and J. Abernethy, "Enhanced spatiotemporal relational probability trees and forests," *Data Mining and Knowledge Discovery*, vol. 26, no. 2, pp. 398–433, 2013.

[11] F. Wu, Z. Li, W.-C. Lee, H. Wang, and Z. Huang, "Semantic annotation of mobility data using social media," in *Proceedings of the 24th International Conference on World Wide Web*, pp. 1253–1263, 2015.

[12] F. Wu and Z. Li, "Where did you go: Personalized annotation of mobility records," in *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, pp. 589–598, 2016.

[13] S. Chawla, S. Shekhar, W. Wu, and U. Ozesmi, "Modeling spatial dependencies for mining geospatial data," in *Proceedings of the 2001 SIAM International Conference on Data Mining*, pp. 1–17, SIAM, 2001.

[14] M. L. Stein, *Interpolation of spatial data: some theory for kriging*. Springer Science & Business Media, 2012.

[15] Z. Jiang, "A survey on spatial prediction methods," *IEEE Transactions on Knowledge and Data Engineering*, vol. PP, pp. 1–1, 08 2018.

[16] S. Shekhar, C. Lu, and P. Zhang, "Detecting graph-based spatial outliers.," *Intelligent Data Analysis*, vol. 6, no. 5, pp. 451–468, 2002.

[17] S. Shekhar and S. Chawla, *Spatial databases: a tour*, vol. 2003. prentice hall Upper Saddle River, NJ, 2003.

[18] C. Yujun, P. Juhua, D. Jiahong, W. Yue, and X. Zhang, "Spatial–temporal traffic outlier detection by coupling road level of service," *IET Intelligent Transport Systems*, vol. 13, no. 6, pp. 1016–1022, 2019.

[19] J. Haslett, R. Bradley, P. Craig, A. Unwin, and G. Wills, "Dynamic graphics for exploring spatial data with application to locating global and local anomalies," *The American Statistician*, vol. 45, no. 3, pp. 234–242, 1991.

[20] N. Cressie, "Statistics for spatial data: Wiley series in probability and statistics," 1993.

[21] R. Webster and M. Oliver, *Software for spatial data analysis in 2D.*, vol. 48. European Journal of Soil Science, 1997.

[22] R. Haining, *Spatial data analysis in the social and environmental sciences*. Cambridge University Press, 1993.

[23] L. Anselin, "Exploratory spatial data analysis and geographic information systems," *New tools for spatial analysis*, vol. 54, 1994.

[24] L. Anselin, "Local indicators of spatial association-lisa," *Geographical analysis*, vol. 27, no. 2, pp. 93–115, 1995.

[25] S. Shekhar, M. R. Evans, J. M. Kang, and P. Mohan, "Identifying patterns in spatial information: A survey of methods," *WIREs Data Mining and Knowledge Discovery*, vol. 1, no. 3, pp. 193–214, 2011.

[26] G. Zheng, S. L. Brantley, T. Lauvaux, and Z. Li, "Contextual spatial outlier detection with metric learning," in *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 2161–2170, 2017.

[27] M. Zala, L. Rushirajsinh, M. Mehta, B. Brijesh, M. Zala, and R. Mahipalsinh, "A Survey on Spatial Co-location Patterns Discovery from Spatial Datasets," *International Journal of Computer Trends and Technology*, vol. 7, no. 3, pp. 137–142, 2014.

[28] Y. Huang, S. Shekhar, and H. Xiong, "Discovering colocation patterns from spatial data sets: a general approach," *IEEE Transactions on Knowledge and data engineering*, vol. 16, no. 12, pp. 1472–1485, 2004.

[29] V. Estivill-Castro and I. Lee, "Data mining techniques for autonomous exploration of large volumes of geo-referenced crime data," in *Proc. of the 6th International Conference on Geocomputation*, pp. 24–26, 2001.

[30] A. S. Fotheringham, C. Brunsdon, and M. Charlton, *Geographically weighted regression: the analysis of spatially varying relationships*. John Wiley & Sons, 2003.

[31] R. Caruana, "Multitask learning," *Machine learning*, vol. 28, no. 1, pp. 41–75, 1997.

[32] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise.," in *Knowledge Discovery in Databases*, vol. 96, pp. 226–231, 1996.

[33] Z. Cai, J. Wang, and K. He, "Adaptive density-based spatial clustering for massive data analysis," *IEEE Access*, vol. 8, pp. 23346–23358, 2020.

[34] X. Zhou, H. Zhang, G. Ji, and G. Tang, "A multi-density clustering algorithm based on similarity for dataset with density variation," *IEEE Access*, vol. 7, pp. 186004–186016, 2019.

[35] B. Wu and B. M. Wilamowski, "A fast density and grid based clustering method for data with arbitrary shapes and noise," *IEEE Transactions on Industrial Informatics*, vol. 13, no. 4, pp. 1620–1628, 2016.

[36] Y.-m. Cheung and Y. Zhang, "Fast and accurate hierarchical clustering based on growing multilayer topology training," *IEEE transactions on neural networks and learning systems*, vol. 30, no. 3, pp. 876–890, 2018.

[37] F. Ferstl, M. Kanzler, M. Rautenhaus, and R. Westermann, "Time-hierarchical clustering and visualization of weather forecast ensembles," *IEEE transactions on visualization and computer graphics*, vol. 23, no. 1, pp. 831–840, 2016.

[38] M. Bendechache, N.-A. Le-Khac, and M.-T. Kechadi, "Hierarchical aggregation approach for distributed clustering of spatial datasets," in *2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW)*, pp. 1098–1103, IEEE, 2016.

[39] A. Woodley, L.-X. Tang, S. Geva, R. Nayak, and T. Chappell, "Using parallel hierarchical clustering to address spatial big data challenges," in *2016 IEEE International Conference on Big Data (Big Data)*, pp. 2692–2698, IEEE, 2016.

[40] J. Wang, C. Zhu, Y. Zhou, X. Zhu, Y. Wang, and W. Zhang, "From partition-based clustering to density-based clustering: Fast find clusters with diverse shapes and densities in spatial databases," *IEEE Access*, vol. 6, pp. 1718–1729, 2017.

[41] J. Yang and S. Puri, "Efficient parallel and adaptive partitioning for load-balancing in spatial join," in *2020 IEEE International Parallel and Distributed Processing Symposium (IPDPS)*, pp. 810–820, 2020.

[42] C. Lu, D. Chen, and Y. Kou, "Detecting spatial outliers with multiple attributes," in *proceedings of 15th IEEE International Conference on Tools with Artificial Intelligence*, pp. 122–128, 2003.

[43] N. Hubballi, B. K. Patra, and S. Nandi, "Ndot: Nearest neighbor distance based outlier detection technique," in *Pattern Recognition and Machine Intelligence*, pp. 36–42, Springer, 2011.

[44] Y. Chen, D. Miao, and H. Zhang, "Neighborhood outlier detection," *Expert Systems with Applications*, vol. 37, no. 12, pp. 8745–8749, 2010.

[45] M. Zhou, T. Ai, G. Zhou, and W. Hu, "A visualization method for mining colocation patterns constrained by a road network," *IEEE Access*, vol. 8, pp. 51933–51944, 2020.

[46] A. M. Sainju, D. Aghajarian, Z. Jiang, and S. K. Prasad, "Parallel grid-based colocation mining algorithms on gpus for big spatial event data," *IEEE Transactions on Big Data*, 2018.

[47] X. Hu, G. Wang, and J. Duan, "Mining maximal dynamic spatial colocation patterns," *IEEE Transactions on Neural Networks and Learning Systems*, 2020.

[48] B. D. Ripley, "The second-order analysis of stationary point processes," *Journal of applied probability*, vol. 13, no. 2, pp. 255–266, 1976.

[49] A. Okabe and F. Miki, "A conditional nearest-neighbor spatial-association measure for the analysis of conditional locational interdependence," *Environment and Planning A*, vol. 16, no. 2, pp. 163–171, 1984.

[50] M. Ruiz, F. López, and A. Páez, "Testing for spatial association of qualitative data using symbolic dynamics," *Journal of Geographical Systems*, vol. 12, pp. 281–309, September 2010.

[51] T. Sakaki, M. Okazaki, and Y. Matsuo, "Earthquake shakes twitter users: Real-time event detection by social sensors," in *ACM WWW*, pp. 851–860, 2010.

[52] M. Mathioudakis and N. Koudas, "Twittermonitor: trend detection over the twitter stream," in *ACM SIGMOD*, pp. 1155–1158, 2010.

[53] C. Li, A. Sun, and A. Datta, "Twevent: segment-based event detection from tweets.," in *CIKM*, pp. 155–164, 2012.

[54] S. Petrović, M. Osborne, and V. Lavrenko, "Streaming first story detection with application to twitter," in *Human Language Technologies*, pp. 181–189, 2010.

[55] X. Dong, D. Mavroeidis, F. Calabrese, and P. Frossard, "Multiscale event detection in social media," *Data Min. Knowl. Discov.*, vol. 29, no. 5, pp. 1374–1405, 2015.

[56] D. C. B. P. S. Earle and M. Guy, "Twitter earthquake detection: earthquake monitoring in a social world," *Annals of Geophysics*, vol. 54, no. 6, pp. 708–715, 2012.

[57] K. Xie, C. Xia, N. Grinberg, R. Schwartz, and M. Naaman, "Robust detection of hyper-local events from geotagged social media data," in *ACM MDMKDD*, 2013.

[58] A. M. MacEachren, A. R. Jaiswal, A. C. Robinson, S. Pezanowski, A. Savelyev, P. Mitra, X. Zhang, and J. Blanford, "Senseplace2: Geotwitter analytics support for situational awareness.," pp. 181–190, 2011.

[59] A. Marcus, M. Bernstein, O. Badar, D. Karger, S. Madden, R. Miller, and O. Arenz, "Twitinfo: Aggregating and visualizing microblogs for event exploration," in *Annual conference on Human factors in computing systems*, pp. 227–236, 2011.

[60] F. Abel, C. Hauff, G.-J. Houben, R. Stronkman, and K. Tao, "Twitcident: Fighting fire with information from social web streams," in *International Conference Companion on World Wide Web*, pp. 305–308, 2012.

[61] J. Yin, S. Karimi, B. Robinson, and M. Cameron, "Esa: Emergency situation awareness via microbloggers," in *ACM CIKM*, pp. 2701–2703, 2012.

[62] M. Imran, S. Elbassuoni, C. Castillo, F. Diaz, and P. Meier, "Extracting information nuggets from disaster- related messages in social media," in *ISCRAM*, ISCRAM Association, 2013.

[63] N. Morrow, N. Mock, A. Papendieck, and N. Kocmich, "Independent evaluation of the ushahidi haiti project," 2011.

[64] Y. Qu, P. F. Wu, and X. Wang, "Online community response to major disaster: A study of tianya forum in the 2008 sichuan earthquake," in *42st Hawaii International International Conference on Systems Science (HICSS-42 2009), Proceedings (CD-ROM and online), 5-8 January 2009, Waikoloa, Big Island, HI, USA*, pp. 1–11, IEEE Computer Society, 2009.

[65] L. Palen, S. Vieweg, S. B. Liu, and A. L. Hughes, "Crisis in a networked world: Features of computer-mediated communication in the april 16, 2007, virginia tech event," *Social Science Computer Review*, vol. 27, no. 4, pp. 467–480, 2009.

[66] I. Shklovski, L. Palen, and J. N. Sutton, "Finding community through information and communication technology in disaster response," in *ACM CSCW* (B. Begole and D. W. McDonald, eds.), pp. 127–136, ACM, 2008.

[67] L. A. S. Denis, L. Palen, and K. M. Anderson, "Mastering social media: An analysis of jefferson county's communications during the 2013 colorado floods," in *ISCRAM* (S. R. Hiltz, L. Plotnick, M. Pfaf, and P. C. Shih, eds.), ISCRAM Association, 2014.

[68] S. Dashti, L. Palen, M. P. Heris, K. M. Anderson, T. J. Anderson, and S. Anderson, "Supporting disaster reconnaissance with social media data: A design-oriented case study of the 2013 colorado floods," in *ISCRAM* (S. R. Hiltz, L. Plotnick, M. Pfaf, and P. C. Shih, eds.), ISCRAM Association, 2014.

[69] T. Shelton, A. Poorthuis, M. Graham, and M. Zook, "Mapping the data shadows of hurricane sandy: Uncovering the sociospatial dimensions of 'big data'," *Geoforum*, vol. 52, pp. 167 – 179, 2014.

[70] A. Tiwari, C. V. D. Weth, and M. S. Kankanhalli, "Multimodal multiplatform social media event summarization," *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 14, pp. 38:1–38:23, Apr. 2018.

[71] S. Poria, E. Cambria, N. Howard, G.-B. Huang, and A. Hussain, "Fusing audio, visual and textual clues for sentiment analysis from multimodal content," *Neurocomputing*, vol. 174, pp. 50–59, 2016.

[72] K. Liu, Y. Li, N. Xu, and P. Natarajan, "Learn to combine modalities in multimodal deep learning," *arXiv preprint arXiv:1805.11730*, 2018.

[73] Y. Wu, A. Pal, J. Wang, and K. Kant, "Incremental spatial clustering for spatial big crowd data in evolving disaster scenario," in *IEEE CCNC*, pp. 1–8, 2019.

[74] F. Laylavi, A. Rajabifard, and M. Kalantari, "A multi-element approach to location inference of twitter: A case for emergency response," *ISPRS Int. J. Geo-Information*, vol. 5, no. 5, p. 56, 2016.

[75] Z. Cheng, J. Caverlee, and K. Lee, "You are where you tweet: a content-based approach to geo-locating twitter users," in *ACM CIKM*, pp. 759–768, 2010.

[76] J. A. de Bruijn, H. de Moel, B. Jongman, J. Wagemaker, and J. C. J. H. Aerts, "Taggs: Grouping tweets to improve global geoparsing for disaster response," *Journal of Geovisualization and Spatial Analysis*, vol. 2, no. 1, p. 2, 2017.

[77] A. Schulz, A. Hadjakos, H. Paulheim, J. Nachtwey, and M. Mühlhäuser, "A multi-indicator approach for geolocalization of tweets," in *ICWSM*, 2013.

[78] W. Zhang and J. Gelernter, "Geocoding location expressions in twitter messages: A preference learning method," *J. Spatial Information Science*, vol. 9, no. 1, pp. 37–70, 2014.

[79] M. Kulldorff, "A spatial scan statistic," *Communications in Statistics - Theory and Methods*, vol. 26, no. 6, pp. 1481–1496, 1997.

[80] D. B. Neill, A. W. Moore, M. Sabhnani, and K. Daniel, "Detection of emerging space-time clusters," in *ACM SIGKDD*, pp. 218–227, 2005.

[81] M. Kulldorff, F. Mostashari, L. Duczmal, W. Katherine Yih, K. Kleinman, and R. Platt, "Multivariate scan statistics for disease surveillance," *Statistics in Medicine*, vol. 26, no. 8, pp. 1824–1833, 2007.

[82] L. Lan, V. Malbasa, and S. Vucetic, "Spatial scan for disease mapping on a mobile population," in *AAAI*, pp. 431–437, 2014.

[83] T. Rattenbury, N. Good, and M. Naaman, "Towards automatic extraction of event and place semantics from flickr tags," in *ACM SIGIR*, pp. 103–110, 2007.

[84] C. Maru, M. Enoki, A. Nakao, S. Yamamoto, S. Yamaguchi, , and M. Oguchi, "Development of failure detection system for network control using collective intelligence of social networking service in large-scale disaster," *Proc. the 27th ACM Conference on Hypertext and Social Media (HT2016), Halifax, Canada*, pp. pp.267–272, July 2016.

[85] C. Maru, M. Enoki, A. Nakao, S. Yamamoto, S. Yamaguchi, and M. Oguchi, "Network failure detection system for traffic control using social information in large-scale disasters," in *ITU Kaleidoscope*, pp. 1–7, 2015.

[86] H. Yanagida, A. Nakao, S. Yamamoto, S. Yamaguchi, and M. Oguchi, "Traffic control system based on sns information in a deeply programmable network," in *IEEE CSCN*, pp. 1–6, 2016.

[87] W. Junbo, K. Sato, S. Guo, W. Chen, and J. Wu, "Big data processing with minimal delay and guaranteed data resolution in disaster areas," *IEEE Transactions on Vehicular Technology*, 2018.

[88] S. S. Khan and J. Wei, "Real-time power outage detection system using social sensing and neural networks," in *IEEE GlobalSIP*, pp. 927–931, 2018.

[89] H. Sun, Z. Wang, J. Wang, Z. Huang, N. Carrington, and J. Liao, "Data-driven power outage detection by social sensors," *IEEE Trans. Smart Grid*, vol. 7, no. 5, pp. 2516–2524, 2016.

[90] K. Lee, J. young Shin, and R. Zadeh, "Twitter's effectiveness on blackout detection during hurricane sandy," 2013.

[91] C. Hultquist, M. B. Simpson, G. Cervone, and Q. Huang, "Using nightlight remote sensing imagery and twitter data to study power outages," in *ACM SIGSPATIAL* (Y. Huang, J. Thill, and H. Zhang, eds.), pp. 6:1–6:6, 2015.

[92] K. Bauman, A. Tuzhilin, and R. Zaczynski, "Using social sensors for detecting emergency events: A case of power outages in the electrical utility industry," *ACM Trans. Management Inf. Syst.*, vol. 8, no. 2-3, pp. 7:1–7:20, 2017.

[93] U. Paul, A. Ermakov, M. Nekrasov, V. Adarsh, and E. Belding, "outage: Detecting power and communication outages from social networks," in *WWW*, 2020.

[94] M. Basu, A. Shandilya, P. Khosla, K. Ghosh, and S. Ghosh, "Extracting resource needs and availabilities from microblogs for aiding post-disaster relief operations," *IEEE Trans. Comput. Social Systems*, vol. 6, no. 3, pp. 604–618, 2019.
[95] D. T. Nguyen, S. R. Joty, M. Imran, H. Sajjad, and P. Mitra, "Applications of online deep learning for crisis response using social media information," *CoRR*, vol. abs/1610.01030, 2016.
[96] H. To, S. Agrawal, S. H. Kim, and C. Shahabi, "On identifying disaster-related tweets: Matching-based or learning-based?," in *IEEE*, pp. 330–337, 2017.
[97] K. Rudra, P. Goyal, N. Ganguly, P. Mitra, and M. Imran, "Identifying sub-events and summarizing disaster-related information from microblogs," in *ACM SIGIR* (K. Collins-Thompson, Q. Mei, B. D. Davison, Y. Liu, and E. Yilmaz, eds.), pp. 265–274, ACM, 2018.
[98] S. Zhang, A. Pal, K. Kant, and S. Vucetic, "Enhancing disaster situational awareness via automated summary dissemination of social media content," in *IEEE GLOBECOM*, pp. 1–7, IEEE, 2018.
[99] T. Sakaki, M. Okazaki, and Y. Matsuo, "Tweet analysis for real-time event detection and earthquake reporting system development," *IEEE Trans. Knowl. Data Eng.*, vol. 25, no. 4, pp. 919–931, 2013.
[100] S. G. Shanthakumar, A. Seetharam, and A. Ramesh, "Understanding the socio-economic disruption in the united states during covid-19's early days," 2020.
[101] M. Saad, M. Hassan, and F. Zaffar, "Towards characterizing COVID-19 awareness on twitter," *CoRR*, vol. abs/2005.08379, 2020.
[102] A. Kruspe, M. Haberle, I. Kuhn, and X. X. Zhu, "Cross-language sentiment analysis of european twitter messages duringthe covid-19 pandemic." https://openreview.net/pdf?id=VvRbhkiAwR, 2020.
[103] M. Cinelli, W. Quattrociocchi, A. Galeazzi, C. M. Valensise, E. Brugnoli, A. L. Schmidt, P. Zola, F. Zollo, and A. Scala, "The COVID-19 social media infodemic," *CoRR*, vol. abs/2003.05004, 2020.
[104] I. J. Cruickshank and K. M. Carley, "Characterizing communities of hashtag usage on twitter during the 2020 COVID-19 pandemic by multi-view clustering," *CoRR*, vol. abs/2008.01139, 2020.
[105] L. Palen and A. L. Hughes, *Social Media in Disaster Communication*, pp. 497–518. Springer International Publishing, 2018.
[106] C. Reuter, A. L. Hughes, and M. Kaufhold, "Social media in crisis management: An evaluation and analysis of crisis informatics research," *Int. J. Hum. Comput. Interaction*, vol. 34, no. 4, pp. 280–294, 2018.
[107] "Using social media for enhanced situational awareness and decision support," 2014.
[108] A. Olteanu, C. Castillo, F. Diaz, and S. Vieweg, "Crisislex: A lexicon for collecting and filtering microblogged communications in crises," in *ICWSM* (E. Adar, P. Resnick, M. D. Choudhury, B. Hogan, and A. H. Oh, eds.), 2014.
[109] D. D. Vu, H. To, W. Shin, and C. Shahabi, "Geosocialbound: an efficient framework for estimating social POI boundaries using spatio-textual information," in *ACM SIGMOD* (A. Züfle, B. Adams, and D. Wu, eds.), pp. 3:1–3:6, 2016.
[110] S. Zhang and S. Vucetic, "Semi-supervised discovery of informative tweets during the emerging disasters," *CoRR*, vol. abs/1610.03750, 2016.
[111] A. Sharma, K. Rudra, and N. Ganguly, "*Going Beyond Content Richness*: Verified information aware summarization of crisis-related microblogs," in *ACM CIKM* (W. Zhu, D. Tao, X. Cheng, P. Cui, E. A. Rundensteiner, D. Carmel, Q. He, and J. X. Yu, eds.), pp. 921–930, ACM, 2019.
[112] "Covid-19 dashboard by the center for systems science and engineering (csse) at johns hopkins university (jhu)." https://coronavirus.jhu.edu/map.html.
[113] T. Alamo, D. G. Reina, and P. Millán, "Data-driven methods to monitor, model, forecast and control covid-19 pandemic: Leveraging data science, epidemiology and control theory," 2020.
[114] M. Kulldorff, "A spatial scan statistic," *Communications in Statistics-Theory and Methods*, vol. 26, no. 6, pp. 1481–1496, 1997.
[115] D. G. Marquez, "Guassion motion data." https://citius.usc.es/investigacion/datasets/gaussianmotiondata. Accessed: January, 2018.
[116] J. Wang, M. C. Meyer, Y. Wu, and Y. Wang, "Maximum data-resolution efficiency for fog-computing supported spatial big data processing in disaster scenarios," *IEEE Transactions on Parallel and Distributed Systems*, 2019.
[117] D. B. Neill and A. W. Moore, "Rapid detection of significant spatial clusters," in *ACM SIGKDD*, pp. 256–265, 2004.