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

Amitangshu Pal, Junbo Wang, Yilang Wu, Krishna Kant, *Fellow, IEEE*, Zhi Liu, *Senior Member, IEEE*, and Kento Sato

Abstract—Situational awareness tries to grasp the important events and circumstances in the physical world through sensing, communication, and reasoning. Tracking the evolution of changing situations is an essential part of this awareness and is crucial for providing appropriate resources and help during disasters. Social media, particularly Twitter, is playing an increasing role in this process in recent years. However, extracting intelligence from the available data involves several challenges, including (a) filtering out large amounts of 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. Spatio-temporal analysis of the data plays a key role in understanding the situation, but is available only sparsely because only a small fraction of people post relevant text and of those very few enable location tracking. In this paper, we provide a comprehensive survey on data analytics to assess situational awareness from social media big data.

Index Terms—Spatial big data analytics, crowd big data, disaster management, situation awareness

1 Introduction

Situational awareness is crucial in a disaster scenario and is often difficult to come by due to the challenges in obtaining the necessary information in a coherent manner and organizing it. Part of the difficulty arises due to patchy availability and overloading of the communications networks; however, it is often unclear what information is most relevant and how it should be gathered. Since disasters can continue to evolve over many 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. However, only a small fraction of people may post about their needs and of those only a tiny fraction usually enables location tracking due to privacy concerns.

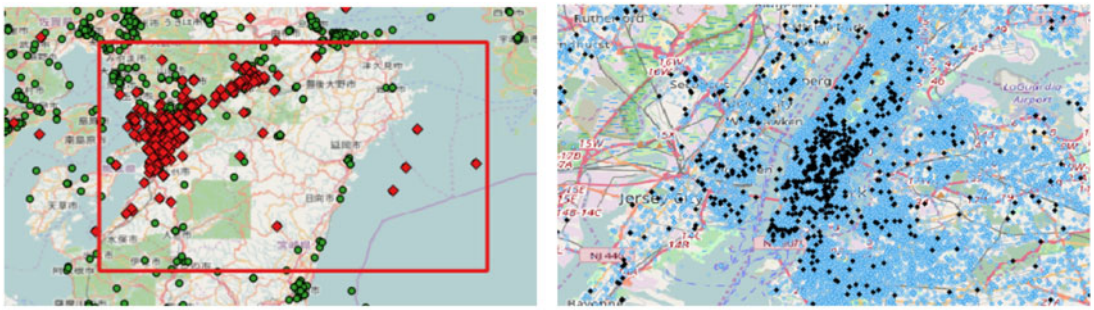
Twitter has established itself as the disaster communication vehicle of choice due to its modest networking requirements, ease of use, and brevity. For example, after the 2011 Japanese earthquake there were more than 5,500 tweets per second after the disaster. Twitter has been used for a wide variety of disaster scenarios, including the three major Hurricanes in 2017, namely Harvey, Maria and Irma that affected Caribbean and US East coast [1], 2019 Pan-European Floods [2], and 2019 US midwestern floods [3], and COVID19 [4] [5].

Fig. 1a 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 official shake map of the region. On the other hand, Fig. 1b shows the power outage related geo-tagged tweets from New York city during Hurricane Sandy in 2012. The regions of Lower Manhattan from Madison Square to the tip of the island was hit the hardest. 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 to social media posts, many other types of data is often also available and can be exploited to gain further insights into both the impacts of the disaster on the physical infrastructure (e.g., damaged transportation routes and assets, damage to power lines/substations, damage to wired/wireless network assets, etc.) and the needs of people affected by it. The sources of such data include various utility companies and service providers. Extracting intelligence from such heterogeneous data involves a lot of challenges including (a) filtering out irrelevant data, (b) fusion of heterogeneous data, (c) dealing with partially geo-tagged social media data, (d) lightweight data analysis mechanisms for near real-time response, and (e) working with

- *Amitangshu Pal is with Computer Science and Engineering, Indian Institute of Technology Kanpur, Kanpur, Uttar Pradesh 208016, India. E-mail: amitangshu@cse.iitk.ac.in.*
- *Junbo Wang is with the Guangdong Provincial Key Laboratory of Intelligent Transportation System, School of Intelligent Systems Engineering, Sun Yat-sen University, Guangdong 510006, China. E-mail: wangjb33@mail.sysu.edu.cn.*
- *Yilang Wu is with PKUtech Company, Ltd., Tokyo 1010037, Japan. E-mail: y-wu@ieee.org.*
- *Krishna Kant is with Computer and Information Science, Temple University, Philadelphia, PA 19122 USA. E-mail: kkant@temple.edu.*
- *Zhi Liu is with the School of Informatics and Engineering, The University of Electro-Communications, Tokyo 182-8585, Japan. E-mail: Liu@ieee.org.*
- *Kento Sato is with the RIKEN Center for Computational Science (R-CCS), Kobe, Hyogo 650-0047, Japan. E-mail: kento.sato@riken.jp.*

Manuscript received 28 May 2021; revised 10 Feb. 2022; accepted 13 Feb. 2022. Date of publication 10 Mar. 2022; date of current version 16 Jan. 2023.
This work was supported in part by JST-NSF joint funding, Strategic International Collaborative Research Program, SICORP on Japan side, and CNS-1461932 award from NSF (USA) and in part by National Natural Science Foundation of China under Grant 62072485.
Sun Yan-sen University is the first affiliation and has major contribution to his work.
(Corresponding author: Junbo Wang.)
Recommended for acceptance by S. Ma.
Digital Object Identifier no. 10.1109/TBDATA.2022.3158431



(a) (b)

Fig. 1. Kumamoto Earthquake Tweets (red: relevant, green: others). (b) Power outage tweets in Hurricane Sandy (dark blue: relevant, light blue: others)

evolving situations. In this paper, we provide a comprehensive survey of these and related issues.

While these challenges apply in general to the processing of multimodal data, there are many aspects that are unique to the disaster applications. **First**, it is important to quickly and continuously process the enormous amount of data being generated during disasters, so that the help can be dispatched expeditiously. A related issue is the continuous evolution of the disaster, which must be reflected in the analysis. *Second*, we can expect that the relevant and useful posts are likely to form a very tiny fraction of all the posts, and even fewer will have location information. However, we expect a temporal and spatial "stickiness" to the application specific posts, particularly those concerning human condition. For example, if a person reports need for food, water, medical care, etc., it is almost certain that (a) the need will persist for some time (even if there is no further post about it), and (b) people in the same or nearby areas have the same need. A suitable modeling of this stickiness can enable reliable conclusions in spite of the sparseness of the data. *Finally*, there are numerous disaster scenarios, each with unique situation awareness needs, however, these can be divided into a small number of categories. One goal of this paper is to provide such a categorization of the disaster applications and review relevant literature on situation awareness for each.

The paper is organized as follows. Section 2 discusses using social media for emergency situational awareness. Section 3 describes the spatial big data analysis. Section 4 summarizes the applications in disaster scenarios. Due to the unique features of disaster scenarios, data mining methods may fail and need to be complemented in several cases, and thus Section 5 discusses challenges and possible solutions. For example, the information decay based spatial clustering (ref. to Section V-D) enlarges the size of data samples allocating the point data into multiple rather than single temporal partitions chronologically. We have demonstrated an application of big data analysis in section 6 through a case study. We introduce evolving clustering to deal with streaming data for detection of evolution of disaster. Some future directions are summarized in section 7. The paper is concluded in section 8.

2 Using Social Media for Disaster 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' used in the tweets [6]. Extensions of this approach include general-purpose detection systems that track a large number of keywords [7], phrases [8] or detect emergence of clusters of similar tweets [9].

Once an event is detected, 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 examples are Sense-place2 [10], 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 [11], a tool that allows an operator to browse a large collection of tweets using a timeline-based 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 [12]. There are also summarization systems that have capability to classify tweets into some of the predefined categories [13]. As a representative system of this type, in [14] 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. In addition to these there are systems that integrate data from multiple sources, such as Ushahidi (www. ushahidi.com) [15], 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 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 [16]. Reference [17] have studied the peer-to-peer communication

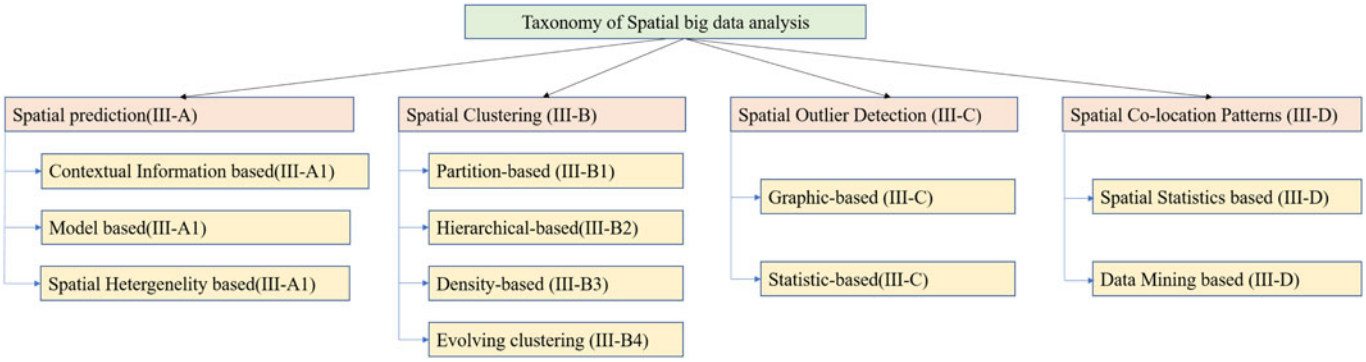


Fig. 2. Characterization of big data analysis for situation awareness.

from a variety of other platforms especially Facebook after the Virginia Shooting in 2007, and southern California wildfires in 2007 [18]. During the 2013 Colorado Floods, different flood-related communications in Facebook and Twitter are examined in [19], [20].

More recently, researchers started paying more attention to the spatial aspect of events [21]. For example, [22] 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 clusters are events and which are not [23]. Big data analysis from the temporal-spatial point of view can assist governments or rescue teams grasp the distribution of the situation in the disaster area, predict the evolution of situations in temporal-spatial space, and find the correlated features behind the data.

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 [24]. Only a small fraction of tweets typically have 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.

3 Spatial Big Data Analytics for Situational Awareness

Spatial analytics studies the relationships between the data 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 [25]. In this section we discuss four major aspects in spatial analytics [26], namely spatial prediction, spatial clustering, spatial outlier detection, and spatial co-location pattern discovery. The taxonomy and structure of this section is briefly shown in Fig. 2; in the following we discuss each category in details.

3.1 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 $s_1, s_2, \ldots, s_n$, and a set of explanatory features $X = [x(s_1), x(s_2), \ldots, x(s_i), \ldots, x(s_n)]^T$ at these locations. Let $Y = [y(s_1), y(s_2), \ldots, y(s_i), \ldots, y(s_n)]^T$ 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 for others. This is illustrated in Fig. 3. 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 can be sub-divided into two categories, i.e., spatial auto-correlation (dependency) and spatial heterogeneity (non-dependency) models.

3.1.1 *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 to have 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 two kinds of approaches, i.e., based on spatial contextual information and based on prediction models [27].

The first approach is through augmentation of the training data with additional spatial contextual information that refers to spatial relationships such as neighborhood of the input data. The relationships can be learned based on traditional machine learning models, e.g., SVM or decision tree. The spatial contextual information can be grasped directly from location information [28], such as distance or direction, or can be collected from multi-source data [29] [30]. The big-

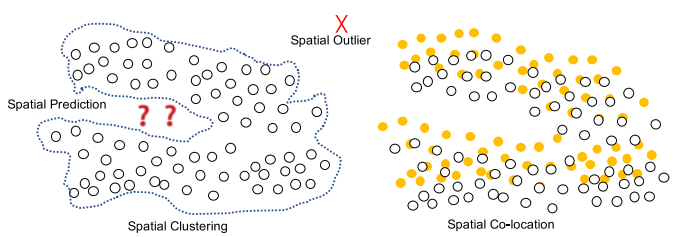


Fig. 3. Four major types of spatial analytics.

gest benefit of this approach is that many traditional non-spatial prediction/learning models can be used, which is much more convenient for researchers. However, the generation of proper spatial contextual features and the integration of spatial and non-spatial features into machine learning models can be nontrivial. Instead of generating spatial contextual information, some approaches directly integrate the spatial relationships in the prediction model; two such approaches are Markov random field based models [31] and Gaussian process based models [32].

Markov random field (MRF) represents as an undirected graph model of random variables, which have a Markov property. The formal definition is given as follows. Given an undirected graph $G = (V, E)$ and a random variable X_u associated with node $u \in V$, the random variables form a MRF with respect to graph G, if X_u is conditionally independent of all other non-neighbouring variables. Generally, Markov random field can be factorized as the cliques of the graph, and the joint probability can be described as follows:

$$P(X = x) = \prod_{c \in cl(G)} \phi(c) \tag{1}$$

Then by optimizing the maximal likelihood with collected data, a learning algorithm can find a best model to fit the data, and finally output conditional probability for spatial prediction of a new data.

Kriging (Gaussian process regression) [32] is another typical method for spatial prediction, which utilizes an observed spatial relation to do spatial prediction for unknown areas. In the Kriging method, it is assumed that each point i in a space is associated with a value z_i. Let u denotes 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 $\hat{z}_u$ at point u is estimated as a weighted linear combination of the known values in $V(n)$ as shown in equation(2). To minimize the estimation error Kriging calculates a set of optimal weights. There are several types of Kriging with different assumptions. 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., $E(\hat{z}_u - z_u) = 0$, where

$$\hat{z}_u = \sum_{i \in V(u)} w_i z_i, \text{ and } \sum_{i \in V(u)} w_i = 1 \tag{2}$$

Effectively, MRF works more like a supervised learning model, which learns model parameters (i.e., transition probability of one state to another) from the original data, and outputs one or zero to fit the unknown area. In contrast, Kriging works similar to a regression model; it learns a set of optimal weights and generates some values (between 0 and 1) to fit the unknown area.

3.1.2 Spatial Heterogeneity

Spatial heterogeneity is another challenging issue in spatial prediction. It refers to the variation in the sample distribution across the study area [33]. 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, geographically weighted models, and multi-task learning.

An example of integrating spatial information into data mining is geographically weighted models (GWR) [34]. Here the integration of spatial information into linear regression model transforms the equation $y = \mathbf{w^T x} + B$ into $y = \mathbf{w^T x \beta} + B$, where $\boldsymbol{\beta}$ represents the vector of location information of the sample data. The advantage of GWR is that the location independent value $\mathbf{x}$ can be integrated as a location dependent value y smoothly and clearly. However, $\boldsymbol{\beta}$ is a matrix needed to be estimated for each point of interest, and the computation cost becomes high for such kind of estimation.

Another approach is based on multi-task learning. It is a common machine learning solution for heterogeneous data, and can group learning samples into several different learning tasks. To solve the spatial heterogeneity problem, it is possible to decompose the entire approach into several subtasks to learn different models for different regions/locations. Then the learnt sub-models are aggregated together, similar to the ensemble learning. As compared to GWR, the advantage of multi-task learning is its flexibility of different shapes of sub-regions, however determining sub-regions can be non-trivial [27]. Meanwhile, how to select base machine learning models is another question for it.

3.2 Spatial Clustering

Spatial clustering groups similar objects based on various measures such as distance, connectivity, or their relative density in space. As a part of *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 further categorized into partitional clustering, hierarchical clustering, density-based clustering, and grid-based clustering.

3.2.1 Partition-Based Clustering

Partition-based clustering such as *k*-means [36] method separates n objects into k clusters to optimize a given criterion (such as the squared error function). The partitioning around medoid (PAM) algorithm [37] effectively finds the most centrally located objects as representatives of each cluster in an iterative way. To improve the efficiency, the sampling-based clustering large applications (CLARA) algorithm [37] was proposed to accelerate PAM on larger datasets. An enhanced version, named CLARA based on randomized search (CLARANS) [38] algorithm outperforms CLARA and PAM both in efficiency and effectiveness by using a randomized search [63] constrained by the maximum number of neighbors. However, the outputs of partitional clustering algorithms are mostly hyper-ellipsoidal and of similar sizes. Therefore, it is not easy for these algorithms to find clusters with different sizes or shapes [64]. In a disaster scenario, these methods are good for clustering

data and finding the center of each cluster, which can be used for the application such as finding an optimal route to deliver supplies, but not suitable to reflect an arbitrary shape of clusters.

In the most recent research works [39], [40], partitional clustering is further enhanced in several ways. In [39], KMDD (clustering by combining K-means with density and distance-based method) was proposed to first cluster data with ball-shape based on K-means, and then in the second stage subclusters are further processed based on DBSCAN to gain an arbitrary shape of clusters. The integration can achieve a fast clustering and allows arbitrary shapes. Reference [40] proposes a parallel adaptive partitioning algorithm (ParADP) for spatial join operation which can achieve a more balanced partition during spatial join operation.

3.2.2 Hierarchical Clustering

The agglomerative hierarchical clustering is a "bottom up" approach by constructing a tree of clusters. The tree will dynamically grow when a new data point comes. Typical algorithms include BIRCH[41], CURE [42] and ROCK[43]. BIRCH measures closeness similarity by using *centroid-* or *medoid*, and it outperforms the CLARANS algorithm for large datasets. The *single-link* hierarchical methods such as CURE [42] find clusters of arbitrary shapes and different sizes by measuring the similarity of the closest pair of data points belonging to different clusters. But hierarchical methods are susceptible to noise, outliers, and artifacts. The aggregate similarity based methods such as ROCK [43] consider new measures, e.g., *inter-connectivity*, and Chameleon [64] further overcomes its limitation by measuring both *inter-connectivity* and *closeness* for identifying the most similar pair of clusters. In a disaster scenario, the advantage can be arbitrary shapes of clusters, however the disadvantage is much more clear, including computation complexity by gradually adding each point to the cluster, and noise data can affect the clustering results a lot in the early stage.

3.2.3 Density-Based Clustering

The most popular density-based clustering method is DBSCAN [48] 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 satisfy density-reachable conditions. Several studies have been proceeded to improve DBSCAN, from parameter setting [65], efficiency optimization [66], and parallelization [57], [59], [60]. In [49], an adaptive DBSCAN was proposed to deal with the data points between two clusters. The research in [50] considered non-uniform distribution of density parameters during clustering, proposed algorithm DENSS to identify the clusters of different densities, shapes and sizes. 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 as long as the grid cell is not too small. In [51], grid based algorithm is integrated with DBSCAN, which can promote a faster clustering while keeping arbitrary shapes of clustering.

3.2.4 Evolving Clustering Techniques

With the development of big data applications, clustering technologies also evolve in order to deal with some emerging challenges, such as handling streaming big data [52], [53], [54], [55]. Evolving cluster was originally proposed in [67], which separates clustering procedure into online and offline stages. During online stage, micro clusters are generated temporally when the streaming data arrives, and once an aggregation command arrives, micro clusters aggregate together to produce a global cluster. This can be done in an offline stage. In [53], ant colony stream clustering (ACSC) algorithm was proposed, in which a tumbling window model is used to read a stream and micro clusters are incrementally formed during a single pass of a window. Micro clusters are then refined by using an ant-inspired method, which emulates an ant's pick-up and drop actions. But how to select representative points in micro clusters and how to handle the rapidly evolving patterns still are critical problems, which have been tackled in [54][55]. They were proceeded by extending the Affinity Propagation (AP) algorithm and an online version STRAP [56]. AP is a message passing-based clustering method proposed in [68], which does not need to decide the number of clusters in advance and the original points can be set as cluster centers directly. STRAP [56] is an enhanced version of AP to process data clustering by incrementally updating the current model.

3.3 Spatial Outlier Detection

Spatial outlier detection [92] discovers the data which are spatially distinct from their surrounding neighbors, such as the red cross mark in Fig. 3. In many real applications using geographic information, such as transportation, public, safety, and location based services [93][94], 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. For example, the size and type of business determines the amount of road traffic that this business will create.

Outlier detection is a typical approach in machine learning and data mining field and can be implemented based on clustering, classification, or regression techniques in machine learning. Spatial outlier detection is similar but more concentrated on discovering some unexpected, interesting, and useful spatial outlier pattern for further analysis. Here spatial objects can be seen as spatial points with attribute values (non-spatial value such as temperature).

There are various statistical tools or methods available for spatial outlier detection. The spatial statistics literature [70] provides two kinds of bi-partite multi-dimensional tests, namely, graphical tests and quantitative tests. One is graphic-based approach, such as variogram clouds in [69][70] and pocket plots in [69] [95]. These visualize the data first and then find the corresponding spatial outliers. However, graphical tests need a precise criteria to distinguish the spatial outliers. The other is a quantitative test including Scatterplots [96][97] that show attribute values on the X-axis and the average of the attribute values in the neighborhood on the Y-axis. A regression line is drawn to identify spatial outliers. Moran scatterplots [74], [76] is another type of quantitative test method, which shows the spatial association or non-association of

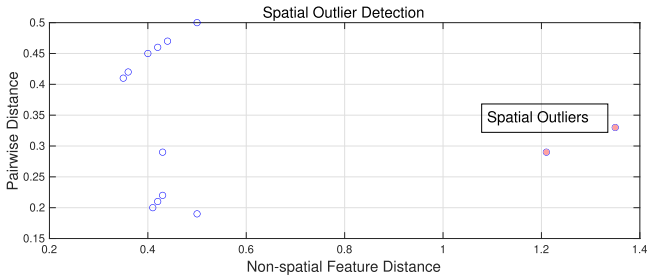


Fig. 4. Spatial outlier detection based on a variogram cloud.

spatially close objects. Although quantitative tests share common technologies with the graphical tests, they outperform graphical tests by providing a more precise result.

As an example, bipartite tests are typical multi-dimensional spatial outlier detection methods, which use the spatial attributes to characterize location, neighborhood, and distance. Then they further find a spatially referenced object in the neighborhood based on non-spatial attributes such as temperature. A Variogram Cloud can be used for spatial outlier detection [98]. In a variogram, the x-axis represents the non-spatial features and y-axis represents the spatial distance of pairs of points. As an illustration, Fig. 4 shows two pairs above the main group of pairs; these are possibly related to spatial outliers. The two pairs marked as spatial outliers have short pairwise distance in y axis, but big difference in x axis. Several other methods used for spatial outlier detection[99] include kNN, and also different statistic measures are used for representing spatial distance, e.g., z-value. The z-value is used to detect spatial outliers for an attribute value, e.g., $f(x)$, which follows a specific distribution, by calculating standard deviation of the value in the location x. For the spatial data at location x, the outlier is detected if its z-value is larger than a predefined threshold.

3.4 Spatial Co-Location

Spatial co-location discovery [100] finds the subsets of features that are frequently located together in the same geographic area as shown in Fig. 3 (white and yellow circles). Spatial co-location mining problem can be formalized as follows [81]: Given a set F of K types of spatial features $F = \{f_1, f_2, \ldots, f_K\}$, and their instances $I = \{i_1, i_2, \ldots, i_D\}$, 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 refers to efficiently finding 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[81] include $cross$-K function [77] with Monte Carlo simulation and mean nearest-neighbor distance [78]. The $cross$-K function for binary spatial features is defined as λ_j^{-1}E[number of type j instance within distance h of a randomly chosen type i instance] [77], which

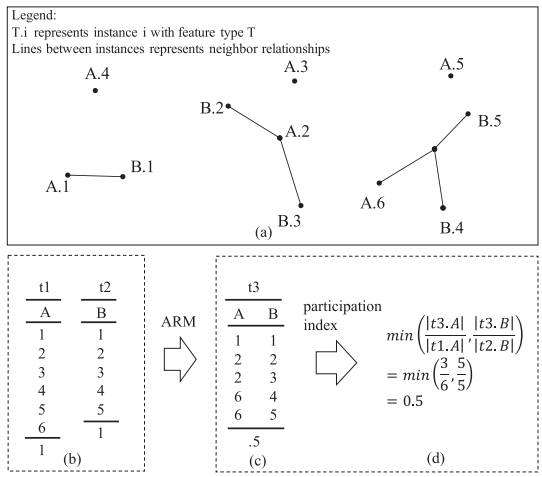


Fig. 5. An example of association rule mining (ARM).

can be estimated by Monte Carlo simulation. It can be used to represent co-location pattern of two features i and j. Mean nearest-neighbor distance calculates average feature distance with other data.

Data mining approaches can be further divided into the clustering-based map approach and association rule-based approaches, or their integration [101]. Association rule mining (ARM) was first introduced in [102] as an efficient approach for finding frequent and meaningful relations among several sets of items in large spatial databases [103]. It outputs participation ratio (between 0 and 1) to represent the co-location relationship of two features: 1 represents almost all the points from two features are co-located, and 0 shows the opposite case.

As inspired by [81], Fig. 5a shows a toy example to compute the participation ratio. There are some points in the figure with two types of features A and B. ARM works by first generating an instance table for each type of feature, i.e., $t1$ for feature A and $t2$ for feature B as shown in Fig. 5b. And then it generates co-location relation table $t3$ in Fig. 5c, in which each pair of instances are located as a neighborhood, e.g., A.1 and B.1. The participation index is further calculated as shown in Fig. 5d.

The output from $cross$-K function is quite clear based on its definition. However, computing $cross$-K function for all possible points can be computationally expensive given a large collection of spatial features [81]. Even the procedure is different, ARM output a similar value, i.e., participation index, to represent co-location pattern of two types of data. Also high computation time is required for spatial join operations in [81], however it has been enhanced by a joinless approach in [83], by using an instance-lookup scheme instead of an expensive spatial join operation. The experiment results show that joinless approach can reduce almost half of execution time when the neighbor distance is setting around 200m in a real data-set. For only clique and star co-location patterns, authors in [84] introduced a more efficient co-location mining method, by defining constraint neighbor for clique or star co-location pattern respectively. For example, to detect a star co-location pattern, a point l_k must be the neighbor of the center point in another star type data.

In recent years, spatial co-location mining has been further developed to deal with big data issues, mainly by

exploiting the parallelism [87], [89], [104]. In [104], parallel co-location mining algorithm is proposed for working with GPU-based platforms based on iCPI tree. iCPI tree is short for (improved Candidate Pattern Instance tree), which is an index to represent the neighborhood relation of instances for different features. Assume there are several instances of feature A, B, and C, denoted as A_i, B_i and C_i. iCPI tree can easily tell us what are the neighbors of a given instance A_i having feature C. A GPU-based version of iCPI tree is proposed in [104], and further enhanced using a grid-based approach [87] to reduce the computation complexity of spatial co-location mining procedure. In [89], spatial co-location mining is redesigned using the MapReduce framework where several reducers accomplish different stages of spatial data processing, including searching neighboring pairs, counting neighboring objects, and finding co-located events.

Another interesting approach is to deal with dynamic relationships among spatial features, e.g., the decrease of "algae" and the increase of "water hyacinth" belongs to spatial co-location patterns. To detect such a dynamic pattern, reference [90] gives various definitions to quantify a dynamic spatial co-location pattern including dynamic feature, dynamic distance threshold, dynamic spatial neighborhood relationships. It then proposes a data mining algorithm to reflect the dynamic relationships among the spatial features. Comparison of different spatial data mining algorithms are summarized in Tables 1 and 2.

3.5 Deep Learning-Based Approach

In recent years, deep learning technologies have been gradually adopted into spatial data analysis. Different from traditional ways, they learn spatial-temporal patterns directly from the data, instead of the predefined rules beforehand. Most popular techniques are CNN to detect spatial patterns, and RNN (e.g., LSTM) to find temporal patterns, or the integration of both. A spatial-temporal event prediction framework was proposed in [105] based on Deep Neural Networks. It consists of 3D convolution networks, where one dimension encodes the temporal information, second dimension represents the spatial information, and the third dimension encodes both. In [106], the authors proposed a deep spatio-temporal residual network (ST-ResNet) to predict inflow and outflow of crowds in each region. ST-ResNet adopts convolution-based residual network to model spatial dependency between any two regions, and then uses the same residual networks in different timeline to detect temporal dependency. Three types of temporal dependency are considered, including distant, near, and recent, to represent three kinds of temporal closeness. Finally the above three residual networks are integrated together to output the final prediction. Authors in [107] have enhanced the research with 3D convolutions for traffic prediction, considering temporal and spatial information. They have similar network architecture with [106], while local temporal patterns and long-term temporal patterns are further adopted in the learning architecture.

LSTM-based traffic flow prediction method has been developed in [108]. They first developed an Attentive Traffic Flow Machine (ATFM), which consists of two Convolutional Long Short-Term Memory (ConvLSTM [109]) units to learn spatial and temporal patterns. It has two LSTM units and connected in a sequential way through a convolution layer in the middle, where the first LSTM unit takes normal traffic features as input and then outputs to the connected convolution layer for spatial feature detection. The second LSTM unit learns more effective spatial and temporal patterns after the convolution layer. The whole learning architecture consists of a sequential representation learning module and a period representation learning module, both of which are constructed based on ATFM. The purpose is to learn different kinds of temporal patterns based on different types of learning architecture.

4 Situation Awareness in Different Disaster Applications

According to World Health Organization, "a disaster is an occurrence disrupting the normal conditions of existence and causing a level of suffering that exceeds the capacity of adjustment of the affected community". Disasters can evolve over very long periods of time; however, the focus of this paper is on events that cover large geographic areas and evolve rather rapidly such that the dissemination of relevant information becomes challenging. There are numerous types of events even under this restricted definition with varying impacts and mitigation challenges; however, they all crucially depend on rapid and accurate situational awareness.

4.1 Practical Disaster Scenarios and Social Media

There are numerous types of disasters, each requiring specialized big data and deep learning techniques depending on the nature of available data. For example, wild-fires typically occur away from populated areas and even monitoring them is a big challenge. Here we only speak of some disasters where social media based analysis plays a significant role.

Earthquake/Landside. Social media big data provides a chance to understand situations after an earthquake such as sentiment/attitude for the government actions, requirements from people and so on. It is quite a typical application for situational awareness based on social media big data.

COVID-19. Big Data/Deep Learning Technologies have been used for COVID-19 to perform semantic situation understanding through content analysis of Twitter posts[4][5]. In [4], the authors have collected a large-scale twitter dataset for COVID19 sentiment analysis, and through the analysis based on various machine learning and deep learning methods, they found in some periods people were losing trust in the government to control the situation, and then the behaviors changed to fear, disgust and sadness later. In [5], authors have researched on the sentiment behaviors of lockdown in India during COVID19. They collected 12741 tweets which includes the keyword "Indialockdown" from April 5 to April 17, 2020, and implemented several supervised machine learning methods to recognize the attitude of tweets. During the analysis, they found that around 49% are positive, 21% are negative and around 30% are neutral.

From the situational awareness perspectives, the existing literature can be categorized as follows: (1) Assessment of communications network outages and mitigation mechanisms, (2) Assessment of power outages that may interfere with communications, (3) Prediction of situational evolution, and (4) Assessment of individual needs (e.g., need for

TABLE 1
Comparisons of Spatial Data Mining Algorithms: Part 1

Types	Categories		Basic Description	Models and Algorithms	Details	Usecases in Disaster
Spatial Prediction	Spatial Autocorrelation	Contextual Information	Extract spatial relationships from contextual information	Reference [28]	Extract spatial relationships directly from location information	Predict situations in an unknown area by prediction models.
				Reference [27]	Extract spatial relationships from raster data	Spatial information can be extracted from multiple data sources.
				Reference [29][30]	Extract spatial contextual information from multiple data sources	
		Model-based	Build a unified model for spatial and non-spatial information	Reference [31]	Build a Markov Random Field based prediction model	
				Reference [32]	Build a Gaussian Process (Kriging) based prediction model	
	Spatial Heterogeneity		Assume spatial data do not follow an identical distribution	Reference [34]	Integrate spatial information into learning model for a location-dependent learning	Wireless signal prediction for disaster resilient base station.
				Reference [35]	Decompose the model into some sub-models based on multi-task learning	
Spatial Clustering	Partition-based Approaches		Separates the whole target into several clusters	Reference [36]	k-means method	Clustering people who need a specific supply, and design a delivery route by visiting cluster centers.
				Reference[37]	PAM which effectively finds the most centrally objects as representatives of cluster	
				Reference [37] [38]	A sampling-based clustering large applications (CLARA) algorithm for large datasets	
				Reference [39]	Partition-Density joint clustering	
				Reference [40]	Adaptive partition for spatial analysis	
	Hierarchical-based Approaches		A "bottom up" approach by constructing a tree of clusters	Reference [41]	BIRCH: measure similarity based on centroid or medoid	Similar to the partition-based approaches, but need to pay attention to noise data.
				Reference [42]	CURE: single-link hierarchical methods	
				Reference [43]	ROCK: Consider new measures, e.g., inter-connectivity	
				Reference [44]	Hierarchical clustering based on topology learning to reduce computation complexity	
				Reference [45]	Time-hierarchical clustering	
				Reference [46]	Hierarchical aggregation for distributed clustering	
				Reference [47]	Parallel hierarchical clustering	
	Density-based Approaches		Take density as critical for clustering	Reference [48]	DBSCAN	Draw an arbitrary shape of area with dense people who need something.
				Reference [49]	Adaptive DBSCAN for massive data	
				Reference [50]	For different densities, shapes and sizes	
				Reference [51]	Grid-based DBSCAN	
	Clustering for big data stream		Evolving of Cluster	Reference [52]	Evolving cluster based on dynamic generating micro clusters	For real-time data stream, and the situation evolves accordingly.
				Reference [53]	Ant colony stream clustering (ACSC) algorithm to incrementally update micro clusters	
				Reference [54][55][56]	To select representative points and handle a better evolving pattern	
				Reference [56]	Evolving clustering based on Affinity Propagation (AP) algorithm	
	Parallel Clustering		Spatial clustering in a parallel way	Reference [57][58]	DBSCAN with MapReduce	Need to speedup the clustering procedure.
				Reference [59]	A parallel-processing model on a multi-core CPU	
				Reference [60]	Distributed spatial clustering by merging local clusters together	
				Reference [61][62]	Parallel clustering in GPU	

food, water, medical help, etc.). The key works in these areas are summarized in Table 3, which are also elaborated in the following subsections.

4.2 Situational Awareness of Network Connectivity

We define network disturbance as any situation that negatively impacts ability of nodes to send and receive data. This

TABLE 2
Comparisons of Spatial Data Mining Algorithms: Part 2

Types	Categories	Basic Description	Models and Algorithms	Details	Usecases in Disaster
Spatial Outlier Detection	Graphic-based Approaches	Visualize data in a graph and find spatial outliers	Reference [69][70]	Variogram clouds: the points with near location but large variance on attribute value indicate spatial outliers.	Abnormal location of people's sentiment: Easy to see in a graph
	Statistics-based Approaches	Detect spatial outliers based various statistics information	Reference [71]	Z-value: compute standardized difference for each point	Spatial outliers from statistics point of view, such as hotspots where need more supplies consider all requirements in the whole area
			Reference [72][73]	kNN-based solution	
			Reference [74] [75]	Scatterplots: a regression line is drawn to identify spatial outliers	
			Reference [76]	Moran scatterplot: outliers are the points surrounded by unusual value of neighbors.	
Spatial Co-location Pattern Discovery	Statistics-based Approaches	Based on staticstics information	Reference [77]	Cross-K function	Co-location pattern such as earthquake magnitude and people's sentiment
			Reference [78]	Cross nearest distance	
			Reference [79]	Q-test	
	Data Mining based Approaches	Association rule based Approaches	Reference [80]	Visualization and data mining	Quick co-location pattern discovery while the computation complexity can be reduced by specific data mining methods
			Reference [81][82]	Spatial join based approach	
			Reference [83]	Joint-less approach	
			Reference [84]	Constraint neighbourhood based approach	
		Clustering based Approaches	Reference [85]	Layer-based approach by finding overlapped areas	
			Reference [86]	Mixed clustering	
	Other Approaches	Parallel based Approaches	Reference [87]	Parallel solution on GPU	For specific requirements, such as understanding evolving situation or needing quick response
			Reference [88] [89]	Parallel based on Map-Reduce framework for big data	
		Dynamic Approaches	Reference [90] [91]	To solve dynamic changing problem of co-location patterns	

might include situations when the network demand exceeds capacity due to bursts of activity and 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 taking adaptive actions to recover from them. Given the high degree of robustness and redundancy of the public communications networks, large scale network failures are very rare, as evidenced by the network damage during the Kumamoto earthquake and hurricane Sandy. 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 large scale network outages.

4.2.1 Detection of Network Disturbances

Detection of network disturbance can be performed by analyzing the spatial scan statistics [110] and its many extensions [111], [112], [113] 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. Of course, the accessibility to the required data often is a constraint.

Spatial scan based algorithms have traditionally been used for disease mapping where the objective is to find regions containing significantly increased incidence of disease symptoms, but many other applications also exist. 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 [114] in order to distinguish between place and event related tags. The key challenge for analyzing such scan statistics is computation 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.

4.2.2 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

TABLE 3
Situational Awareness in Disaster Situations

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 [115], [116]	Characterizing network failures from user complaints about network functionality or slowness
		Reference [117]	Used data plane programmability of the Openflow switches to adopt flexible network control
		Reference [118]	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 [119]	Developed a modified approach of Kleinberg's burst detection algorithm to promptly detect the power outages from the tweets
		Reference [120]	Developed a supervised Latent Dirichlet Allocation to detect power outages
		Reference [121]	Proposed a k-means clustering scheme for the efficient allocation of power resources based on the available tweets
		Reference [122]	Shown that Twitter data fused with satellite imagery can identify power outage information at a street-level resolution
		Reference [123]	Developed a predictive model for identifying Tweets referring to real power outages
		Reference [124]	Separated the tweets into power outage, communication outage and both power-communication outage related events
Disaster Evolution	Analysis of Covid related tweets regarding public awareness,sentiment analysis, and classification of informative tweets from others	Reference [125]	Identified Covid related hashtags, along with the linguistic analysis of the tweets in different hashtag groups
		Reference [126]	Characterized public awareness regarding Covid by analyzing tweets in the affected countries
		Reference [127]	Implemented a neural network for sentiment analysis using multilingual sentence embeddings
		Reference [128]	Discussed the diffusion of Covid related information with a massive data analysis on Twitter
		Reference [129]	Proposed a multi-view clustering for analyzing tweets using clustering hashtags
Resource Need Evolution	Analysis of tweets regarding resource needs, availability; filtering, summarization and classification of informative tweets from others	Reference [130]	Analyzed tweets regarding resource needs and resource availability
		Reference [131]	Developed a DNN to identify and classify informative tweets into topical classes
		Reference [132]	Compared matching-based and learning-based approaches for effectively identifying relevant messages from matching keywords and hashtags in social media data
		Reference [133]	Proposed an ILP to generate summaries of twitter messages
		Reference [134]	Enhanced real-time situational awareness through filtering and summarization of social media data
		Reference [135]	Developed a probabilistic spatio-temporal model to find the center of the target event

characterizing failures in the network [115] – i.e., user complaints about the network functionality or slowness. Examples of such tweets are as follows [116]: *"I cannot get through to Miyagi. I am worried."*, or *"I am 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.

Reference [117] uses the data plane programmability of the Openflow switches to provide a more flexible control of wired networks. 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.

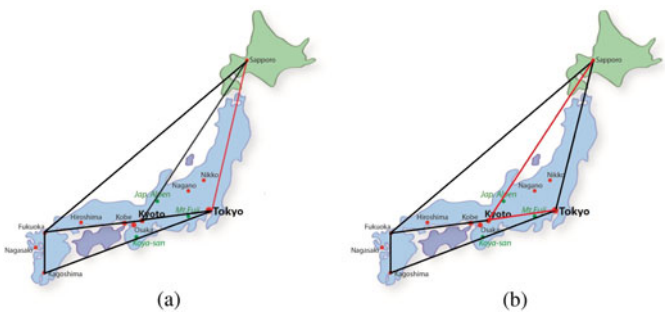


Fig. 6. Adaptation of the routes after sniffing potential congestion between Tokyo and Sapporo. (a) Route-1 is the direct route from Tokyo–Sapporo, whereas (b) Route-2 goes through Kyoto.

4.2.3 Finding Network Isolation and Resource Allocation

Another application of situational awareness is to identify isolated regions that are functional but disjoint 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 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 isolation 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 isolation.

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 [118]. 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.

4.3 Situational Awareness of Power Outages

Real time situational awareness for detecting power outages from social media data has received interest in recent years. Reference [119] have used keyword 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 [120] 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 [121] 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 [122] the authors have analyzed the brightness change in the satellite data along with the density of power outage for identifying the severely impacted areas.

The studies show that Twitter data fused with satellite imagery can identify power outage information at a street-level resolution. In [123] 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 referred to the real outages. In [124] 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 boosting and support vector machine can successfully classify the outage related tweets from 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 90 seconds of training and testing time.

4.4 Situational Awareness concerning Disaster Evolution

In recent years, researchers have started using social media data for deriving evolving disaster events. In fact, it has been studied that the digital footprint of a disaster is typically proportional to its impact in the ground level. For example, the researchers in [136] have studied that the number of photographs uploaded in Flickr during Hurricane Sandy strongly correlates with the atmospheric pressure in New Jersey. In [137] the authors have studied the Twitter activities of 50 metropolitan areas in the United States during hurricane Sandy and have shown strong correlation between the hurricane's path and the hurricane related tweets. The authors have also demonstrated that the per-capita Twitter activity strongly correlates with the per-capita economic damage inflicted by the hurricane. Similar studies are also reported in [138] that shows a close relationship between damages caused by Sandy and Twitter activities. Another similar study has also been reported in [139], where the authors have studied that the disaster related tweets and the distribution of damage, physical extents of floods during the River Elbe Flood in Germany in 2013 follow similar spatio-temporal distribution.

More recently, researchers have begun using social media platforms to derive insights regarding the continued evolution of Covid-19 pandemic over 2020-21, that has placed substantial stress on medical personnel and supplies including hospital-beds, doctors, nurses, paramedics, personal protection equipment (PPE), ventilators, ambulances, police, test kits, testing supplies, common medications currently being prescribed, etc. In [125] the authors have identified Covid-19 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 [126] the authors have characterized public awareness regarding Covid by analyzing tweets in the most affected countries. Specifically, the authors have examined the (a) 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 [127] 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 [128] have addressed the diffusion of Covid related information with a massive data analysis on Twitter, Instagram, YouTube, Reddit and Gab. They have also fit information spreading with epidemic models characterizing the basic reproduction numbers for each platform. The authors in [129] 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 [140], along with their difficulties and challenges.

4.5 Situational Awareness of Human Needs in Disaster

Social media data for situational awareness in crisis scenario are discussed in [141], [142], [143]. In [130] the authors have

analyzed tweets regarding resource needs and availability (e.g., transport, food, water, health-care, etc.) for 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 Italy earthquake 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 [131] 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 [132] has provided a comparison between matching-based [144], [145] and learning-based [131], [146] 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 [133] the authors have proposed an Integer Linear Programming (ILP) technique that summaries a big volume of twitter messages around some identified sub-events, that helps crisis responders to quickly understand the situation. Reference [147] has generated *verified* summaries from the information posted on Twitter during disasters. Enhancing real-time situational awareness through filtering and summarization of social media data is reported in [134]. The authors have reported the study of twitter data during the 2012 Sandy Hurricane from New York, Philadelphia, Boston, and Washington DC. In [135] the authors have devised a classification 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. They 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 of 3 or more.

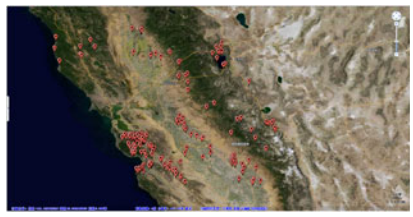
4.6 Open Social Media Datasets for Real Disasters

We now discuss some of the open datasets from some social media during real disasters, as summarized in Table 4. `COVID-19-TweetIDs` [148] is performing an ongoing collection of tweets IDs associated with the COVID-19, which started from January 28, 2020. It gathers historical Tweets from the preceding 7 days, and the coverage is worldwide. `COVID19_twitter` [149] is updated every 2 days since March 2020, and for each day it collects around 4 millions tweets. The dataset is suitable for NLP study, since it also provides top frequent term for each day. The COVID19 datasets for Chinese users can be found in [150]. CrisisNLP provides a dataset [151] including various types of disasters, such as Earthquake, Typhoon, Floods, Landslide and so on. But the time period is only from 2014 to 2015. `GlobalFloodMonitor` [153] collected 88 million tweets, with 10,000 flood events across 176 countries in 11 languages. Among these datasets, `Bdr-tweet` [152] provides geotagged tweets for 15 disasters,

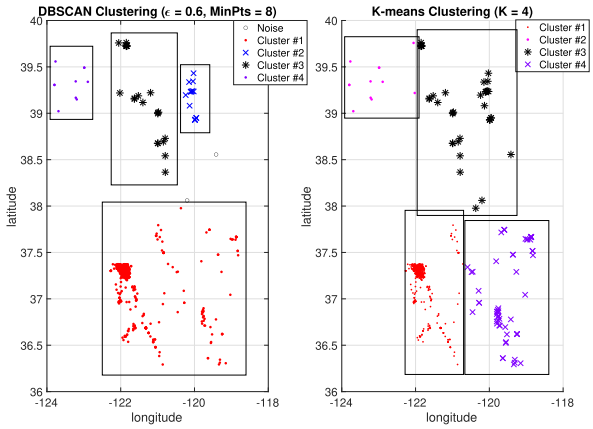
TABLE 4
Summary of Some Open Disaster Related Social Media Datasets

Dataset	Types	Data Size	Time Range	Coverage	Resource
COVID-19-TweetIDs [148]	COVID-19	6.9 GB	Since Jan. 21, 2020	World	1.98 billion tweets
COVID-19 Twitter[149]	COVID-19	About 12GB	Since Mar. 11, 2020	World	1 billion tweets
Weibo-COV V2[150]	COVID-19		From Dec. 1 2019 to Dec 30 2020	China	65.2 million tweets
Disaster-related Tweet[151]	Typhoon, landslide, Ebola virus, etc.	About 3GB	From 2014 to 2015	World	52.8 million tweets
Bdr-tweet[152]	Fire, storm, earthquake, mudslide,etc.	Less than 5GB	From 2014 to 2015	USA	Geotagged tweets for 15 disasters
Global-Flood-Monitor[153]	Flood		From Jul. 2014 to Nov. 2018	176 count-ries	88 million tweets

so that researchers can also see the spatio-temporal spread and patterns of the data. For example in Fig. 7 we show the data of the Napa earthquake; we also perform two types of spatial clustering algorithms, i.e., DBSCAN and k-means on it. The differences between DBSCAN and k-means are quite obvious. DBSCAN finds a cluster based on dense reachable points, and thus it is better at finding a big area in which all of



(a) Raw data of the Napa Earthquake in [152]



(b) Comparison of two clustering methods. In the figure ϵ is the maximum distance between two neighbouring points, *MinPts* is the number of sample points in a neighborhood, and K is the number of clusters.

Fig. 7. Illustration of DBSCAN and k-means clustering methods on dataset from [152].

the people have the same requirements after a disaster. For example, the areas with higher density of people who need water can be gradually connected. As shown in the left part of Fig. 7b, the red tweets which mentioned earthquake can be clustered as a big area. On the other hand k-means is better at locating the center of cluster, which may be important for dispatching supplies.

5 Challenges in Integrating Big Data With Emergency Scenarios

Even if several applications have been studied on situational awareness in disaster scenarios, there are still challenging issues when integrating big data with emergency network. 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. Below we list some of the key challenges regarding deriving situational awareness from disaster related data analysis.

5.1 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.

5.2 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.

5.3 Multimodal Data Fusion

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 multimodal data fusion: feature-level fusion and decision-level fusion, also known as early fusion and late fusion. Feature-level fusion merges features from different types of data resources together before classification. For example, in [154] 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 [155], both of these methods were evaluated with text, video and audio contents, and the results from the both approaches increase around 10% precision as compared to 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 modalities can be automatically selected to achieve high accuracy of situation detection in [156].

5.4 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 [157] 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, 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 the decay model to the clusters, but not for each point data, which will affect the accuracy of situation representation.

5.5 Utilizing Non Geo-Tagged Tweets

Another key challenge of using Twitter data is the scarcity of the number of geo-located tweets, which typically varies between 0.42% to 3.17% [158]. Utilizing the non geo-tagged tweets can also provide useful information if they can be related approximately to their origin. Some works [158], [159]

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 [159]. 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 [158].

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 [160], namely toponym recognition and toponym resolution. Toponym recognition techniques [161] 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 [162] 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.

5.6 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.

6 Temporal Evolution of Spatial Features: A Case Study

Much of the data is very hot when generated, and then its popularity wanes over time. In some cases, the data may become hot again but this is less likely as the data ages. 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 has the highest power when a tweet originates, and then gradually fades over time. Information energy for a specific location can be accumulated with 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. An exponential model is desirable since it corresponds to a fixed additional decay for each additional unit time elapsed. Thus,

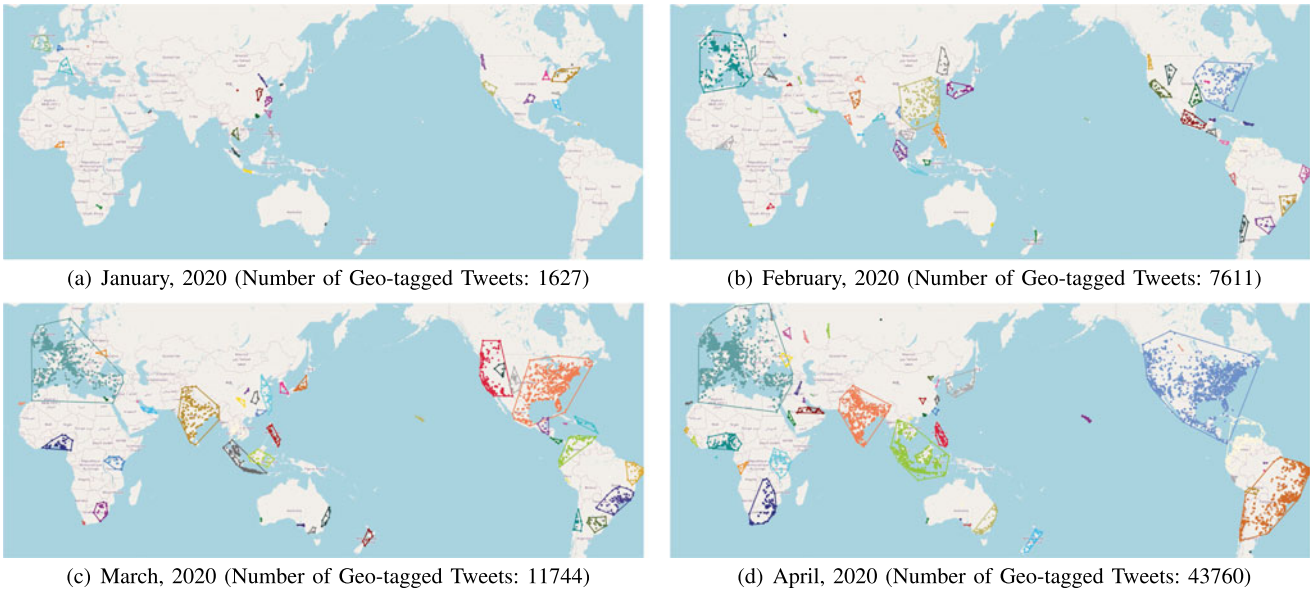
$$E_\epsilon(p, t_c) = E_\epsilon(p, t_p) \cdot \eta^{-\lambda \cdot (t_c - t_p)} \tag{3}$$

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

To find the spatial hotspots during an evolving disaster, we choose a density measure based on *Kulldorff's spatial scan statistic* [164], which is commonly used in finding the significant spatial clusters in case of emerging outbreaks. With this the incremental spatial clustering in an 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 weight over regions. The SDC tracks the boundary and movement of the dense regions of the targeted evolving disasters.

We demonstrate temporal and spatial evolution of tweets related to Covid-19 pandemic [165] as shown in Fig. 8 during January-April 2020 timeline. From this figure we can observe that the temporal density variation of the tweets (geo-points) across different sub-continents roughly match the evolution of Covid-19 over this time. 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 quite sparse which also matches with the small number of cases in those regions. A DBSCAN-based spatial clusters indicates the regions being dense of covid tweets.

We have also implemented an interpolation mechanism on the sparse Covid data while applying the TDIE similar to equation(3). The results are shown in Figs. 9a, 9b, 9c, and 9d, where the green color denotes the Covid-19 related words, whereas the red color denotes the data that mentioned related words about both Covid-19 and personal equipment. The



(a) January, 2020 (Number of Geo-tagged Tweets: 1627)

(b) February, 2020 (Number of Geo-tagged Tweets: 7611)

(c) March, 2020 (Number of Geo-tagged Tweets: 11744)

(d) April, 2020 (Number of Geo-tagged Tweets: 43760)

Fig. 8. Spatial densities of Covid-19 related tweets for Feb to April, 2020. The spatial density here is measured by using the DBSCAN configured by the following three settings: (1) The maximum distance between any two neighbouring points is set as 3, while the unit is the longitude and latitude degree based on the Geographic Coordinate System, (2) the number of samples in a neighborhood for a point to be considered as a core point is set as 10, (3) the nearest-neighbors algorithm is set as "kd_tree". The density is represented by the spatial clusters, and each cluster contains a set of DBSCAN-defined neighbouring points bounded by a convex-hull boundary.

results shows that the red and green points are very much correlated, and show an increasing trend in April-May period as compared to February-march. To compare the results with the ground report, we depict the number of cases, hospitalization and deaths of three US states in Figs. 9e and 9f, from that data obtained from the John Hopkins University's repository [163]. These results also show an increasing trend during April-May timeline, which also validates the outcome of our interpolation mechanism. From these figures we can observe that the trend of Covid is roughly observed by correlating the spatial and temporal distribution of Covid tweets with the pandemic news reports.

However, as the number of geo-tagged, Covid related tweets are extremely low, we did not find a large number of papers studying their detailed spatio-temporal analysis. Rather, we analyze the studies that have geospatial analysis of Covid related data obtained from different sources. In [166] the authors have collected the coronavirus pneumonia data from different official websites during January 30, 2020 to February 18, 2020. In [167] the authors have collected epidemic data till January 30, 2020, and have recorded that confirmed and death cases in Hubei province accounted for 59.91% and 95.77% of the total cases in China respectively. During that time the authors have recorded that the number of cases in some cities was relatively low, although the risk factors appeared to be increasing. In [168] the authors have analyzed the temporal and spatial distribution of the pandemic; an important point that they have noticed is that a large number of people entered into Wenzhou from Hubei Province, which is the main reason for the outbreak in this region. Authors in [169] have studied the spatio-temporal propagation of the first Covid wave in China and compared it to other global locations. They have also studied the spatial propagation of the pandemic from Hubei to other provinces in China in terms of distance, population size, human mobility etc. In [170] the authors have studied the COVID-19 and SARS outbreaks at the provincial levels in mainland China and have concluded that they exhibit distinct spatio-temporal clustering patterns; this may be due to different social and demographic factors, containment strategies or differences in transmission mechanisms. Similar spatio-temporal variations and epidemiological maps of cases in other countries like USA, Iran, Italy, Spain, India etc. are reported in [138], [171], [172], [173], [174].

As mentioned earlier, the number of Covid related tweets with geo-tags are extremely sparse (∼0.036% as observed from our experiments), so, we could not conduct any spatial aggregation and clustering analysis on a daily or weekly basis. We therefore simulate the incremental spatial clustering using a synthetic database obtained from [175]. 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 changes over time. We use the *"3C2D2400Spiral"* dataset, which presents a helix-like 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. 10 to illustrate the position, movement, and coverage of the hot-spots 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.

In the above we have demonstrated a preliminary study of spatial clustering of the relevant data, which imitates disaster related tweets. Through this small experiment, we tried to demonstrate the nature of temporal stickiness of such tweets;, i.e., if somebody from disaster location tweets that he/she needs food, water, medical help, etc., chances are that the same thing applies to most others in that location.

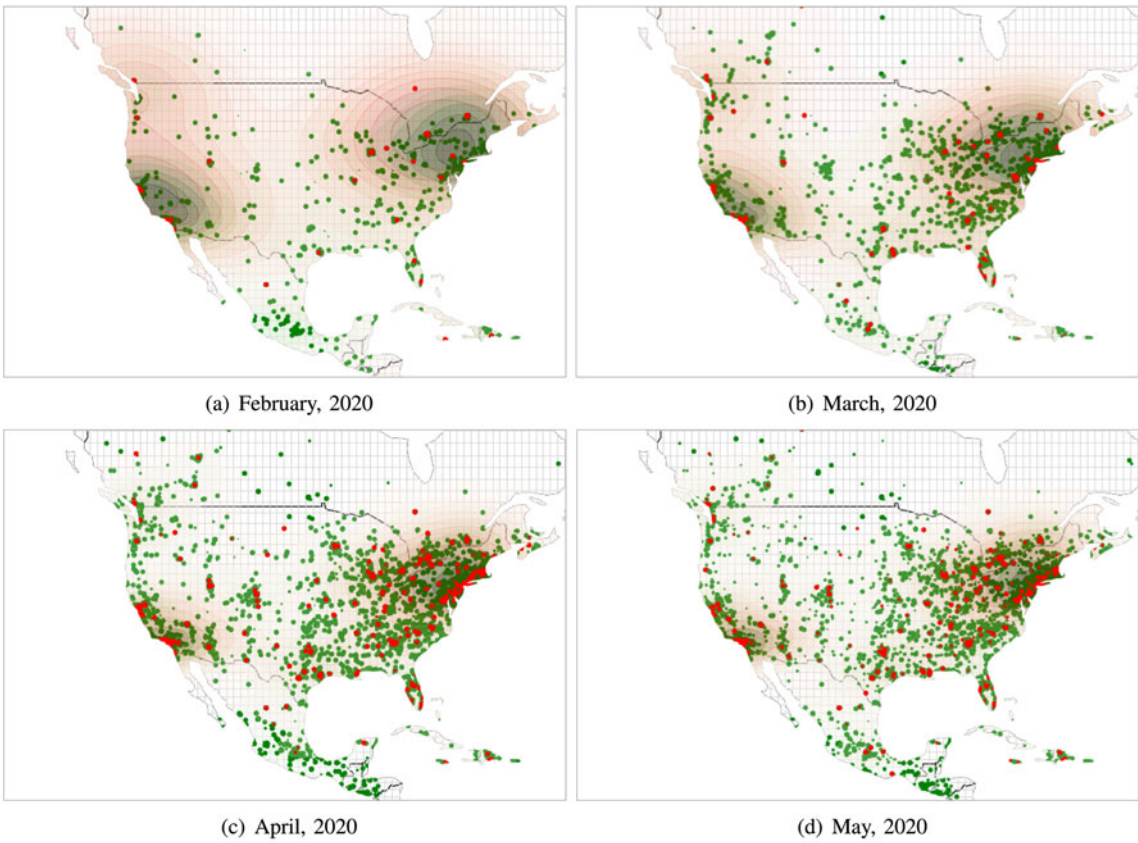


Fig. 9. The spatial-temporal interpolation is applied to the geo-tagged tweets (31,109 collected records), the green and red dots denote the tweets mentioning covid-19 related words, and red dots also mention the emergent words related with personal protection equipment (PPE). The different size of the dots represent the temporal interpolation based on the exponential decay, and the colored region represent the spatial interpolation based on the Kernel density of the temporally decayed point data. The analysis shows a general decrease in the northwest, with the impacted areas moving south an east towards southern California and Arizona. This trend is verified with the actual data from JHU [163] shown in panels (e), (f), (g), in which the dots denote the number of new cases and death cases reported, and the lines are the scatter-plot smoothing based on the dots respectively.

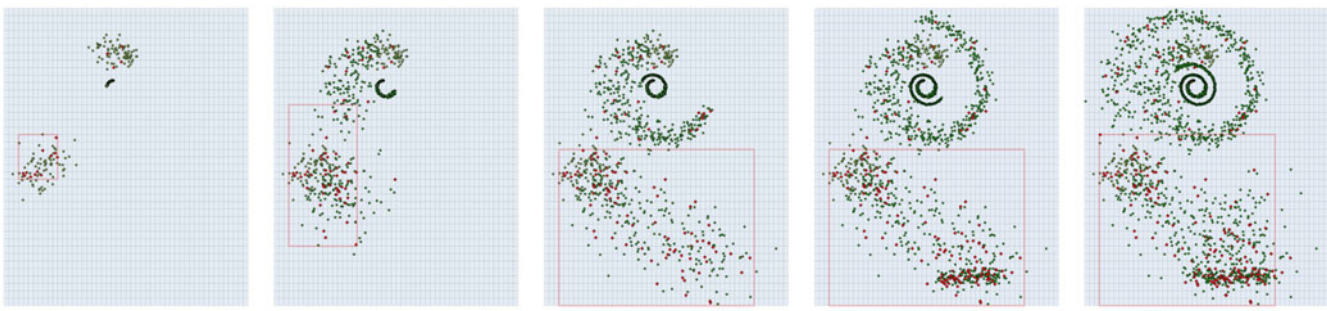


Fig. 10. Position of hot-spot ($\eta^{\lambda} = 2$) for helix-like movement dataset [157] (©[2019]IEEE).

Also such needs may continue to evolve over time and space, and therefore modeling such an evolution is crucial for situational awareness. In fact such stickiness phenomenon is quite a general phenomenon, and so our analysis is quite generic and is applicable to most disaster applications.

7 Future Research Directions

In this section we discuss some possible future directions for social media driven big data analysis.

Choice of Appropriate Mining Algorithm. In this paper, we have summarized the literature on four types of data mining algorithms in section III. However, the success of these methods depends on the volume and quality of data, and how robust the learning model is. Also, supervised learning approaches including deep learning, require large amounts of labeled data, which is not easily available in disaster scenarios. Thus the following issues need to be considered while selecting a learning algorithm in disaster scenarios: (1) different disaster applications require different types of data mining algorithms as summarized in Tables 1 and 2; (2) the data mining algorithm needs to deal with the evolving situation, which has been studied in spatial clustering, but still need further investigation for spatial prediction, spatial outlier detection, and spatial co-location pattern detection; and (3) the human needs expressed by the people trapped by the disaster tend to be sticky, both spatially and temporally, and thus quantifying the impact of this stickiness is another challenging issue.

Edge Computing for Spatial Analytics. Social media-driven big data analytics plays a key role in situational awareness; therefore, a timely analysis is necessary for quick response. Edge computing is an option to enable such analytics services within the Radio Access Network and in close proximity of the affected people. This is particularly important during disasters since the longer-distance communication to reach the cloud may be difficult due to extreme network congestion that is often experienced during disasters.

However, edge computing still has several challenges in disaster scenarios. First, in a disaster scenario, communication and computational resources can be very limited. Designing efficient edge computing requires joint allocation of those resources between edge devices and servers by considering specific limitations in disaster scenarios. Second, some applications in disaster scenarios need to execute tasks of multiple priority levels, corresponding to different emergency levels, different computation workloads and computation results of distinct performances. For example in object detection, considering more detection regions (i.e., region proposals) involves higher computational complexity but can achieve higher detection accuracy. Therefore, optimal allocation of resources in an edge computing scenarios, with different optimization variables and specific objectives, while considering the requirements in disaster scenarios (e.g., energy consumption, completion time, system utility) requires significant future research.

Situation Awareness and Sentiment Analysis. Situational awareness tries to collect the social media data to grasp the important events and circumstances in the physical world through sensing, communication, and reasoning. We have discussed four types of spatial analysis methods in the paper. However, they need support from content analysis such as sentiment analysis, which grasps the feeling of people, e.g., fear after an earthquake or anxiety when there is not enough daily supplies. Integration of spatial and sentiment analysis is very important, but still has some crucial challenges. Some sentiments are quite general and have a spatial dependency, such as worrying about water shortage, whereas others may be tied to very specific needs of the individuals and thus do not have a spatial dependency such as shortage of a specific medicine. Separating such generic requirements from the individualistic ones (to enable its analysis) can be quite complex and needs to be investigated further.

8 Conclusion

During disasters, the data relevant to situational assessment is generated from many different sources including social media used by the affected people (usually Twitter), direct communications with others, possibly unaffected, users who put the information on the social media, and observations by the deployed monitoring infrastructure, etc. 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; therefore the paper focuses on several such opportunities and challenges in extracting situational awareness from disaster related social media data. We hope that this article will spur further research into solutions to many of these issues.

References

[1] F. Alam *et al.*, "A twitter tale of three hurricanes: Harvey, Irma, and Maria," in *Proc. 15th Int. Conf. Informat. Syst. Crisis Response Manage.*, 2018, pp. 553–572.

[2] V. Lorini *et al.*, "Integrating social media into a pan-european flood awareness system: A multilingual approach," 2019, *arXiv:1904.10876*.

[3] "Midwest farmers take to twitter to document flood disaster," May 2019. [Online]. Available: https://grist.org/article/disaster-stricken-farmers-in-the-midwest-take-to-twitter-to-document-destruction/

[4] U. Naseem, I. Razzak, M. Khushi, P. W. Eklund, and J. Kim, "COVIDSenti: A large-scale benchmark twitter data set for COVID-19 sentiment analysis," *IEEE Trans. Computat. Social Syst.*, vol. 8, no. 4, pp. 1003–1015, Aug. 2021.

[5] P. Gupta, S. Kumar, R. R. Suman, and V. Kumar, "Sentiment analysis of lockdown in india during COVID-19: A case study on twitter," *IEEE Trans. Computat. Social Syst.*, vol. 8, no. 4, pp. 992–1002, Aug. 2021.

[6] T. Sakaki *et al.*, "Earthquake shakes twitter users: Real-time event detection by social sensors," in *Proc. 19th Int. Conf. World Wide web*, 2010, pp. 851–860.

[7] M. Mathioudakis *et al.*, "TwitterMonitor: Trend detection over the twitter stream," in *Proc. ACM SIGMOD Int. Conf. Manage. data*, 2010, pp. 1155–1158.

[8] C. Li *et al.*, "Twevent: Segment-based event detection from tweets," in *Proc. ACM Int. Conf. Informat. knowl. Manage.*, 2012, pp. 155–164.

[9] S. Petrović *et al.*, "Streaming first story detection with application to twitter," in *Proc. Annu. Conf. North Amer. Chapter Assoc. Comput. Linguistics Hum. Lang. Technol.*, 2010, pp. 181–189.

[10] A. M. MacEachren *et al.*, "SensePlace2: GeoTwitter analytics support for situational awareness," in *Proc. IEEE Conf. Vis. Anal. Sci. Technol.*, 2011, pp. 181–190.

[11] A. Marcus *et al.*, "Twitinfo: Aggregating and visualizing microblogs for event exploration," in *Proc. Annu. Conf. Hum. Factors Comput. Syst.*, 2011, pp. 227–236.

[12] F. Abel *et al.*, "Twitcident: Fighting fire with information from social web streams," in *Proc. Int. Conf. Companion World Wide Web*, 2012, pp. 305–308.

[13] J. Yin *et al.*, "ESA: Emergency situation awareness via microbloggers," in *Proc. ACM Int. Conf. Informat. Knowl. Manage.*, 2012, pp. 2701–2703.

[14] M. Imran *et al.*, "Extracting information nuggets from disaster-related messages in social media," in *Proc. 10th Int. Conf. Informat. Syst. Crisis Response Manage.*, 2013, pp. 791–801.

[15] N. Morrow *et al.*, "Independent evaluation of the ushahidi haiti project," *Develop. Informat. Syst. Int.*, vol. 8, 2011, Art. no. 111.

[16] Y. Qu *et al.*, "Online community response to major disaster: A study of tianya forum in the 2008 sichuan earthquake," in *Proc. 42st Hawaii Int. Conf. Syst. Sci.*, 2009, pp. 1–11.

[17] L. Palen *et al.*, "Crisis in a networked world: Features of computer-mediated communication in the April 16, 2007, virginia tech event," *Social Sci. Comput. Rev.*, vol. 27, no. 4, pp. 467–480, 2009.

[18] I. Shklovski *et al.*, "Finding community through information and communication technology in disaster response," in *Proc. ACM Conf. Comput. Supported Cooperative Work*, 2008, pp. 127–136.

[19] L. A. S. Denis *et al.*, "Mastering social media: An analysis of jefferson county's communications during the 2013 colorado floods," in *Proc. Int. Conf. Informat. Syst. Crisis Response Manage.*, 2014.

[20] S. Dashti *et al.*, "Supporting disaster reconnaissance with social media data: A design-oriented case study of the 2013 colorado floods," in *Proc. Int. Conf. Informat. Syst. Crisis Response Manage.*, 2014.

[21] X. Dong *et al.*, "Multiscale event detection in social media," *Data Mining Knowl. Discov.*, vol. 29, no. 5, pp. 1374–1405, 2015.

[22] P. S. Earle *et al.*, "Twitter earthquake detection: Earthquake monitoring in a social world," *Ann. Geophys.*, vol. 54, no. 6, pp. 708–715, 2012.

[23] K. Xie *et al.*, "Robust detection of hyper-local events from geotagged social media data," in *Proc. ACM 13th Int. Workshop Multimedia Data Mining*, 2013, pp. 1–9.

[24] T. Shelton *et al.*, "Mapping the data shadows of hurricane sandy: Uncovering the sociospatial dimensions of 'big data'," *Geoforum*, vol. 52, pp. 167–179, 2014.

[25] S. Shekhar *et al.*, "Trends in Spatial Data Mining," *Data Mining: Next Generation Challenges and Future Directions*. Palo Alto, CA, USA: AAAI Press, 2003, pp. 357–380.

[26] J. Wang, Y. Wu, N. Yen, S. Guo, and Z. Cheng, "Big data analytics for emergency communication networks: A survey," *IEEE Commun. Surv. Tuts.*, vol. 18, no. 3, pp. 1758–1778, Jul.–Sep 2016.

[27] Z. Jiang, "A survey on spatial prediction methods," *IEEE Trans. Knowl. Data Eng.*, vol. 31, no. 9, pp. 1645–1664, Sep. 2019.

[28] A. McGovern *et al.*, "Enhanced spatiotemporal relational probability trees and forests," *Data Mining Knowl. Discov.*, vol. 26, no. 2, pp. 398–433, 2013.

[29] F. Wu *et al.*, "Semantic annotation of mobility data using social media," in *Proc. 24th Int. Conf. World Wide Web*, 2015, pp. 1253–1263.

[30] F. Wu *et al.*, "Where did you go: Personalized annotation of mobility records," in *Proc. 25th ACM Int. Conf. Informat. Knowl. Manage.*, 2016, pp. 589–598.

[31] S. Chawla *et al.*, "Modeling spatial dependencies for mining geospatial data," in *Proc. SIAM Int. Conf. Data Mining*, 2001, pp. 1–17.

[32] M. L. Stein, *Interpolation of Spatial Data: Some Theory for Kriging*. Berlin, Germany: Springer, 2012.

[33] Z. Jiang, "A survey on spatial prediction methods," *IEEE Trans. Knowl. Data Eng.*, vol. 31, no. 9, pp. 1645–1664, Sep. 2019.

[34] A. S. Fotheringham *et al.*, *Geographically Weighted Regression: The Analysis of Spatially Varying Relationships*. Hoboken, NJ, USA: Wiley, 2003.

[35] R. Caruana, "Multitask learning," *Mach. Learn.*, vol. 28, no. 1, pp. 41–75, 1997.

[36] S. S. Singh *et al.*, "K-means v/s K-medoids: A comparative study," in *Proc. Nat. Conf. Recent Trends Eng. Technol.*, 2011.

[37] L. Kaufman *et al.*, *Finding Groups in Data: An Introduction to Cluster Analysis*. Hoboken, NJ, USA: Wiley, 2009, vol. 344.

[38] P. S. Bradley *et al.*, "Scaling em (expectation-maximization) clustering to large databases," Microsoft Res. Redmond, Redmond, WA, USA, Tech. Rep. MSR-TR-98-35, 1998.

[39] J. Wang *et al.*, "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.

[40] J. Yang *et al.*, "Efficient parallel and adaptive partitioning for load-balancing in spatial join," in *Proc. IEEE Int. Parallel Distrib. Process. Symp.*, 2020, pp. 810–820.

[41] T. Zhang *et al.*, "BIRCH: An efficient data clustering method for very large databases," in *Proc. ACM SIGMOD Int. Conf. Manage. Data*, vol. 25, no. 2, pp. 103–114, 1996.

[42] S. Guha *et al.*, "Cure: An efficient clustering algorithm for large databases," in *ACM SIGMOD Rec.*, vol. 27, no. 2, 1998, pp. 73–84.

[43] G. Saikat, R. Rajeev, and S. Kyuseok, "Rock: A robust clustering algorithm for categorical attributes," in *Proc. IEEE 15th Int. Conf. Data Eng.*, 1999, pp. 512–521.

[44] Y.-M. Cheung and Y. Zhang, "Fast and accurate hierarchical clustering based on growing multilayer topology training," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 3, pp. 876–890, Mar. 2019.

[45] F. Ferstl, M. Kanzler, M. Rautenhaus, and R. Westermann, "Time-hierarchical clustering and visualization of weather forecast ensembles," *IEEE Trans. Vis. Comput. Graphics*, vol. 23, no. 1, pp. 831–840, Jan. 2017.

[46] M. Bendechache, N. Le-Khac, and M. Kechadi, "Hierarchical aggregation approach for distributed clustering of spatial datasets," in *Proc. IEEE 16th Int. Conf. Data Mining Workshops*, 2016, pp. 1098–1103.

[47] A. Woodley, L. Tang, S. Geva, R. Nayak, and T. Chappell, "Using parallel hierarchical clustering to address spatial Big Data challenges," in *Proc. IEEE Int. Conf. Big Data*, 2016, pp. 2692–2698.

[48] M. Ester *et al.*, "A density-based algorithm for discovering clusters in large spatial databases with noise," *Knowl. Discov. Databases*, vol. 96, no. 34, pp. 226–231, 1996.

[49] Z. Cai, J. Wang, and K. He, "Adaptive density-based spatial clustering for massive data analysis," *IEEE Access*, vol. 8, pp. 23 346–23 358, 2020.

[50] 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. 186 004–186 016, 2019.

[51] B. Wu and B. M. Wilamowski, "A fast density and grid based clustering method for data with arbitrary shapes and noise," *IEEE Trans. Ind. Informat.*, vol. 13, no. 4, pp. 1620–1628, Aug. 2017.

[52] M. Hahsler and M. Bolanos, "Clustering data streams based on shared density between micro-clusters," *IEEE Trans. Knowl. Data Eng.*, vol. 28, no. 6, pp. 1449–1461, Jun. 2016.

[53] C. Fahy, S. Yang, and M. Gongora, "Ant colony stream clustering: A fast density clustering algorithm for dynamic data streams," *IEEE Trans. Cybern.*, vol. 49, no. 6, pp. 2215–2228, Jun. 2019.

[54] X. Zhang, C. Furtlehner, C. Germain-Renaud, and M. Sebag, "Data stream clustering with affinity propagation," *IEEE Trans. Knowl. Data Eng.*, vol. 26, no. 7, pp. 1644–1656, Jul. 2014.

[55] J. Sui, Z. Liu, A. Jung, L. Liu, and X. Li, "Dynamic clustering scheme for evolving data streams based on improved STRAP," *IEEE Access*, vol. 6, pp. 46 157–46 166, 2018.

[56] X. Zhang *et al.*, "Data streaming with affinity propagation," in *Proc. Joint Eur. Conf. Mach. Learn. Knowl. Discov. Databases*, 2008, pp. 628–643.

[57] Y. He *et al.*, "MR-DBSCAN: An efficient parallel density-based clustering algorithm using MapReduce," in *Proc. IEEE 17th Int. Conf. Parallel Distrib. Syst.*, 2011, pp. 473–480.

[58] X. Yue *et al.*, "Parallel k-medoids++ spatial clustering algorithm based on MapReduce," 2016, *arXiv:1608.06861*.

[59] T. Sakai *et al.*, "Parallel processing for density-based spatial clustering algorithm using complex grid partitioning and its performance evaluation," in *Proc. Int. Conf. Parallel Distrib. Process. Techn. Appl.*, 2016, Art. no. 337.

[60] M. Bendechache *et al.*, "Distributed clustering algorithm for spatial data mining," in *Proc. Int. Conf. Spatial Data Mining Geographical Knowl. Serv.*, 2015, pp. 60–65.

[61] C. Böhm *et al.*, "Density-based clustering using graphics processors," in *Proc. ACM Int. Conf. Informat. knowl. Manage.*, 2009, pp. 661–670.

[62] B. Welton *et al.*, "Mr. scan: Extreme scale density-based clustering using a tree-based network of GPGPU nodes," in *Proc. Int. Conf. High Perform. Comput. Netw. Storage Anal.*, 2013, pp. 1–11.

[63] Y. E. Ioannidis *et al.*, "Randomized algorithms for optimizing large join queries," *ACM SIGMOD Rec.*, vol. 19, no. 2, pp. 312–321, 1990.

[64] G. Karypis *et al.*, "Chameleon: Hierarchical clustering using dynamic modeling," *Computer*, vol. 32, no. 8, pp. 68–75, 1999.

[65] M. Ankerst *et al.*, "Optics: Ordering points to identify the clustering structure," *ACM SIGMOD Rec.*, vol. 28, no. 2, pp. 49–60, 1999.

[66] A. Hinneburg *et al.*, "An efficient approach to clustering in large multimedia databases with noise," in *Proc. Int. Conf. Knowl. Discov. Data Mining*, 1998, pp. 58–65.

[67] F. Cao *et al.*, "Density-based clustering over an evolving data stream with noise," in *Proc. SIAM Int. Conf. Data Mining*, 2006, pp. 328–339.

[68] B. J. Frey *et al.*, "Clustering by passing messages between data points," *Science*, vol. 315, no. 5814, pp. 972–976, 2007.

[69] J. Haslett *et al.*, "Dynamic graphics for exploring spatial data with application to locating global and local anomalies," *Amer. Statistician*, vol. 45, no. 3, pp. 234–242, 1991.

[70] N. Cressie, *Statistics for Spatial Data: Wiley Series in Probability and Statistics*. Hoboken, NJ, USA: Wiley, 1993.

[71] S. Shekhar, P. R. Schrater, R. R. Vatsavai, W. Wu, and S. Chawla, "Spatial contextual classification and prediction models for mining geospatial data," *IEEE Trans. Multimedia*, vol. 4, no. 2, pp. 174–188, Jun. 2002.

[72] N. Hubballi *et al.*, "NDOT: Nearest neighbor distance based outlier detection technique," in *Proc. Int. Conf. Pattern Recognit. Mach. Intell.*, 2011, pp. 36–42.

[73] Y. Chen *et al.*, "Neighborhood outlier detection," *Expert Syst. Appl.*, vol. 37, no. 12, pp. 8745–8749, 2010.

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

[75] L. Anselin, "Interactive techniques and exploratory spatial data analysis," *Geographical Information Systems: Principles, Techniques, Management and Applications*. P. Longley, M. Goodchild, D. Maguire, and D. Rhind eds., Hoboken, NJ, USA: Wiley, 1999.

[76] L. Anselin, "The moran scatterplot as an ESDA tool to assess local instability in spatial association," *Spatial Anal. Perspectives GIS*, vol. 111, pp. 111–125, 1996.

[77] B. D. Ripley, "The second-order analysis of stationary point processes," *J. Appl. Probability*, vol. 13, no. 2, pp. 255–266, 1976.

[78] A. Okabe *et al.*, "A conditional nearest-neighbor spatial-association measure for the analysis of conditional locational interdependence," *Environ. Plan. A*, vol. 16, no. 2, pp. 163–171, 1984.

[79] M. Ruiz *et al.*, "Testing for spatial association of qualitative data using symbolic dynamics," *J. Geographical Syst.*, vol. 12, no. 3, pp. 281–309, Sep. 2010. [Online]. Available: https://ideas.repec.org/a/kap/jgeosy/v12y2010i3p281–309.html

[80] 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. 51 933–51 944, 2020.

[81] Y. Huang, S. Shekhar, and H. Xiong, "Discovering colocation patterns from spatial data sets: A general approach," *IEEE Trans. Knowl. Data Eng.*, vol. 16, no. 12, pp. 1472–1485, Dec. 2004.

[82] J. S. Yoo *et al.*, "A partial join approach for mining co-location patterns," in *Proc. 12th Annu. ACM Int. Workshop Geographic Informat. Syst.*, 2004, pp. 241–249.

[83] J. Yoo and S. Shekhar, "A joinless approach for mining spatial colocation patterns," *IEEE Trans. Knowl. Data Eng.*, vol. 18, no. 10, pp. 1323–1337, Oct. 2006.

[84] T. Van Canh and M. Gertz, "A constraint neighborhood based approach for co-location pattern mining," in *Proc. 4th Int. IEEE Conf. Knowl. Syst. Eng.*, 2012, pp. 128–135.

[85] V. Estivill-Castrol *et al.*, "Discovering associations in spatial data—an efficient medoid based approach," in *Proc. Pacific-Asia Conf. Knowl. Discov. Data Mining*, 1998, pp. 110–121.

[86] Y. Huang and P. Zhang, "On the relationships between clustering and spatial co-location pattern mining," in *Proc. 18th IEEE Int. Conf. Tools Artif. Intell.*, 2006, pp. 513–522.

[87] A. M. Sainju, D. Aghajarian, Z. Jiang, and S. Prasad, "Parallel grid-based colocation mining algorithms on GPUs for big spatial event data," *IEEE Trans. Big Data*, vol. 6, no. 1, pp. 107–118, Mar. 2020.

[88] M. Sheshikala, D. R. Rao, and R. V. Prakash, "A map-reduce framework for finding clusters of colocation patterns-a summary of results," in *Proc. IEEE 7th Int. Adv. Comput. Conf.*, 2017, pp. 129–131.

[89] J. S. Yoo, D. Boulware, and D. Kimmey, "A parallel spatial co-location mining algorithm based on MapReduce," in *Proc. IEEE Int. Congr. Big Data*, 2014, pp. 25–31.

[90] X. Hu, G. Wang, and J. Duan, "Mining maximal dynamic spatial colocation patterns," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 3, pp. 1026–1036, Mar. 2021.

[91] J. Duan *et al.*, "Mining spatial dynamic co-location patterns," *Filomat*, vol. 32, no. 5, pp. 1491–1497, 2018.

[92] S. Shekhar *et al.*, "Detecting graph-based spatial outliers," *Intell. Data Anal.*, vol. 6, no. 5, pp. 451–468, 2002.

[93] S. Shekhar *et al.*, *Spatial Databases: A Tour*, vol. 2003, Upper Saddle River, NJ, USA, Prentice-Hall, 2003.

[94] C. Yujun *et al.*, "Spatial–temporal traffic outlier detection by coupling road level of service," *IET Intell. Transport Syst.*, vol. 13, no. 6, pp. 1016–1022, 2019.

[95] R. Webster *et al.*, "Software for spatial data analysis in 2D," *Eur. J. Soil Sci.*, vol. 48, no. 1, pp. 173–175, 1997.

[96] R. Haining, *Spatial Data Analysis in the Social and Environmental Sciences*. Cambridge, U.K.: Cambridge Univ. Press, 1993.

[97] L. Anselin, "Exploratory spatial data analysis and geographic information systems," *New Tools Spatial Anal.*, vol. 54, pp. 45–54, 1994.

[98] S. Shekhar *et al.*, "Identifying patterns in spatial information: A survey of methods," *WIREs Data Mining Knowl. Discov.*, vol. 1, no. 3, pp. 193–214, 2011. [Online]. Available: https://www.onlinelibrary.wiley.com/doi/abs/10.1002/widm.25

[99] G. Zheng *et al.*, "Contextual spatial outlier detection with metric learning," in *Proc. 23rd ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, 2017, pp. 2161–2170.

[100] M. Zala *et al.*, "A survey on spatial co-location patterns discovery from spatial datasets," *Int. J. Comput. Trends Technol.*, vol. 7, no. 3, pp. 137–142, 2014.

[101] V. Estivill-Castro *et al.*, "Data mining techniques for autonomous exploration of large volumes of geo-referenced crime data," in *Proc. 6th Int. Conf. Geocomput.*, 2001, pp. 24–26.

[102] R. Agrawal *et al.*, "Mining association rules between sets of items in large databases," *ACM SIGMOD Rec.*, vol. 22, no. 2, pp. 207–216, 1993.

[103] K. Koperski *et al.*, "Discovery of spatial association rules in geographic information databases," in *Adv. in Spatial Databases*. Berlin, Germany: Springer, 1995, pp. 47–66.

[104] W. Andrzejewski *et al.*, "A parallel algorithm for building ICPI-trees," in *Proc. East Eur. Conf. Adv. Databases Informat. Syst.*, 2014, pp. 276–289.

[105] B. Shen *et al.*, "Stepdeep: A novel spatial-temporal mobility event prediction framework based on deep neural network," in *Proc. 24th ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, 2018, pp. 724–733.

[106] J. Zhang *et al.*, "Deep spatio-temporal residual networks for citywide crowd flows prediction," in *Proc. 31st AAAI Conf. Artif. Intell.*, 2017, pp. 1655–1661.

[107] S. Guo, Y. Lin, S. Li, Z. Chen, and H. Wan, "Deep spatial–temporal 3D convolutional neural networks for traffic data forecasting," *IEEE Trans. Intell. Transp. Syst.*, vol. 20, no. 10, pp. 3913–3926, Oct. 2019.

[108] L. Liu *et al.*, "Dynamic spatial-temporal representation learning for traffic flow prediction," *IEEE Trans. Intell. Transp. Syst.*, vol. 22, no. 11, pp. 7169–7183, Nov. 2021.

[109] S. Xingjian *et al.*, "Convolutional LSTM network: A machine learning approach for precipitation nowcasting," in *Proc. Adv. Neural Informat. Process. Syst.*, 2015, pp. 802–810.

[110] M. Kulldorff, "A spatial scan statistic," *Commun. in Statist. - Theory Methods*, vol. 26, no. 6, pp. 1481–1496, 1997.

[111] D. B. Neill *et al.*, "Detection of emerging space-time clusters," in *Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, 2005, pp. 218–227.

[112] M. Kulldorff *et al.*, "Multivariate scan statistics for disease surveillance," *Statist. Med.*, vol. 26, no. 8, pp. 1824–1833, 2007.

[113] L. Lan *et al.*, "Spatial scan for disease mapping on a mobile population," in *Proc. Conf. Assoc. Advance. Artif. Intell.*, 2014, pp. 431–437.

[114] T. Rattenbury *et al.*, "Towards automatic extraction of event and place semantics from flickr tags," in *Proc. 30th Annu. Int. ACM SIGIR Conf. Res. Develop. Informat. Retrieval*, 2007, pp. 103–110.

[115] C. Maru *et al.*, "Development of failure detection system for network control using collective intelligence of social networking service in large-scale disaster," in *Proc. 27th ACM Conf. Hypertext Social Media*, 2016, pp. 267–272.

[116] C. Maru *et al.*, "Network failure detection system for traffic control using social information in large-scale disasters," in *Proc. ITU Kaleidoscope Trust Informat. Soc.*, 2015, pp. 1–7.

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

[118] 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 Trans. Veh. Technol.*, vol. 68, no. 4, pp. 3833–3842, Apr. 2019.

[119] S. S. Khan and J. Wei, "Real-time power outage detection system using social sensing and neural networks," in *Proc. IEEE Global Conf. Signal Informat. Process.*, 2018, pp. 927–931.

[120] 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, Sep. 2016.

[121] K. Lee *et al.*, "Twitter's effectiveness on blackout detection during hurricane sandy," 2013. [Online]. Available: http://cs229.stanford.edu/proj2013/LeeShin-TwitterEffectivenessOnBlackoutDetection DuringHurricaneSandy_Final.pdf

[122] C. Hultquist *et al.*, "Using nightlight remote sensing imagery and twitter data to study power outages," in *Proc. 1st ACM SIGSPATIAL Int. Workshop Use GIS Emerg. Manage.*, 2015, pp. 6:1–6:6.

[123] K. Bauman *et al.*, "Using social sensors for detecting emergency events: A case of power outages in the electrical utility industry," *ACM Trans. Manage. Inf. Syst.*, vol. 8, no. 2/3, pp. 7:1–7:20, 2017.

[124] U. Paul *et al.*, "Outage: Detecting power and communication outages from social networks," in *Proc. Web Conf.*, 2020, pp. 1819–1829.

[125] S. G. Shanthakumar, A. Seetharam, and A. Ramesh, "Understanding the socio-economic disruption in the united states during COVID-19's early days," 2020, *arXiv:2004.05451*.

[126] M. Saad *et al.*, "Towards characterizing COVID-19 awareness on twitter," *CoRR*, 2020, arXiv:2005.08379.

[127] A. Kruspe *et al.*, "Cross-language sentiment analysis of european twitter messages duringthe COVID-19 pandemic," 2020. [Online]. Available: https://openreview.net/pdf?id=VvRbhkiAwR

[128] M. Cinelli *et al.*, "The COVID-19 social media infodemic," *Sci. Rep.*, vol. 10, no. 1, pp. 1–10, 2020.

[129] I. J. Cruickshank *et al.*, "Characterizing communities of hashtag usage on twitter during the 2020 COVID-19 pandemic by multi-view clustering," *CoRR*, 2020, *arXiv:2008.01139*.

[130] 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. Computat. Social Syst.*, vol. 6, no. 3, pp. 604–618, Jun. 2019.

[131] D. T. Nguyen, S. Joty, M. Imran, H. Sajjad, and P. Mitra, "Applications of online deep learning for crisis response using social media information," 2016, *arXiv:1610.01030*.

[132] H. To, S. Agrawal, S. H. Kim, and C. Shahabi, "On identifying disaster-related tweets: Matching-based or learning-based?," in *Proc. IEEE 3rd Int. Conf. Multimedia Big Data*, 2017, pp. 330–337.

[133] K. Rudra *et al.*, "Identifying sub-events and summarizing disaster-related information from microblogs," in *Proc. 41st Int. ACM SIGIR Conf. Res. Develop. Informat. Retrieval*, 2018, pp. 265–274.

[134] S. Zhang, A. Pal, K. Kant, and S. Vucetic, "Enhancing disaster situational awareness via automated summary dissemination of social media content," in *Proc. IEEE Global Commun. Conf.*, 2018, pp. 1–7.

[135] 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, Apr. 2013.

[136] T. Preis *et al.*, "Quantifying the digital traces of hurricane sandy on flickr," *Sci. Rep.*, vol. 3, no. 3141, 2013, Art. no. 3141.

[137] Y. Kryvasheyeu *et al.*, "Rapid assessment of disaster damage using social media activity," *Sci. Adv.*, vol. 2, no. 3, 2016, Art. no. e1500779.

[138] L. Orea *et al.*, "How effective has been the spanish lockdown to battle COVID-19? A spatial analysis of the coronavirus propagation across provinces," FEDEA, Working Papers 2020–03, 2020. [Online]. Available: https://EconPapers.repec.org/RePEc:fda:fdaddt:2020–03

[139] B. Herfort *et al.*, "Does the spatiotemporal distribution of tweets match the spatiotemporal distribution of flood phenomena? A study about the river elbe flood in june 2013," in *Proc. Int. Conf. Informat. Syst. Crisis Response Manage.*, 2014.

[140] T. Alamo *et al.*, "Data-driven methods to monitor, model, forecast and control COVID-19 pandemic: Leveraging data science, epidemiology and control theory," *Ann. Rev. Control*, vol. 52, pp. 448–464, 2020.

[141] L. Palen *et al.*, *Social Media in Disaster Communication*. Berlin, Germany: Springer, 2018, pp. 497–518.

[142] C. Reuter *et al.*, "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.

[143] Using social media for enhanced situational awareness and decision support, 2014. [Online]. Available: https://www.dhs.gov/publication/using-social-media-enhanced-situational-awareness-decision-support

[144] A. Olteanu *et al.*, "CrisisLex: A lexicon for collecting and filtering microblogged communications in crises," in *Proc. Int. AAAI Conf. Web Social Media*, 2014, pp. 376–385.

[145] D. D. Vu *et al.*, "GeoSocialBound: An efficient framework for estimating social POI boundaries using spatio-textual information," in *Proc. 3rd Int. ACM SIGMOD Workshop Manag. Mining Enriched Geo-Spatial Data*, 2016, pp. 3:1–3:6.

[146] S. Zhang and S. Vucetic, "Semi-supervised discovery of informative tweets during the emerging disasters," 2016, *arXiv:1610.03750*.

[147] A. Sharma *et al.*, "Going Beyond Content Richness: Verified information aware summarization of crisis-related microblogs," in *Proc. ACM Int. Conf. Informat. knowl. Manage.*, 2019, pp. 921–930.

[148] Covid-19-tweetids. [Online]. Available: https://github.com/echen102/COVID-19-TweetIDs

[149] Covid19 twitter. [Online]. Available: https://github.com/thepanacealab/covid19_twitter

[150] weibocov. [Online]. Available: https://github.com/nghuyong/weibo-public-opinion-datasets

[151] Disaster related tweet. [Online]. Available: https://crisisnlp.qcri.org/lrec2016/lrec2016.html

[152] BDR tweet. [Online]. Available: https://github.com/ubriela/bdr-tweet

[153] Global flood monitor. [Online]. Available: https://github.com/jensdebruijn/Global-Flood-Monitor

[154] A. Tiwari *et al.*, "Multimodal multiplatform social media event summarization," *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 14, no. 2s, pp. 38:1–38:23, Apr. 2018.

[155] S. Poria *et al.*, "Fusing audio, visual and textual clues for sentiment analysis from multimodal content," *Neurocomputing*, vol. 174, pp. 50–59, 2016.

[156] K. Liu *et al.*, "Learn to combine modalities in multimodal deep learning," 2018, *arXiv:1805.11730*.

[157] Y. Wu, A. Pal, J. Wang, and K. Kant, "Incremental spatial clustering for spatial big crowd data in evolving disaster scenario," in *Proc. IEEE Annu. Consumer Commun. Netw. Conf.*, 2019, pp. 1–8.

[158] F. Laylavi *et al.*, "A multi-element approach to location inference of twitter: A case for emergency response," *ISPRS Int. J. Geo-Informat.*, vol. 5, no. 5, 2016, Art. no. 56.

[159] Z. Cheng *et al.*, "You are where you tweet: A content-based approach to geo-locating twitter users," in *Proc. ACM Int. Conf. Informat. knowl. Manage.*, 2010, pp. 759–768.

[160] J. A. de Bruijn *et al.*, "TAGGS: Grouping tweets to improve global geoparsing for disaster response," *J. Geovisualization Spatial Anal.*, vol. 2, no. 1, 2017, Art. no. 2.

[161] A. Schulz *et al.*, "A multi-indicator approach for geolocalization of tweets," in *Proc. Int. AAAI Conf. Web Social Media*, 2013, pp. 573–582.

[162] W. Zhang *et al.*, "Geocoding location expressions in twitter messages: A preference learning method," *J. Spatial Informat. Sci.*, vol. 9, no. 1, pp. 37–70, 2014.

[163] J. H. U. Cornavirus Resource Ccenter, "The johns hopkins coronavirus resource center (CRC) is a continuously updated source of COVID-19 data and expert guidance," 2022. Accessed: Feb. 2022. [Online]. Available: https://coronavirus.jhu.edu/region

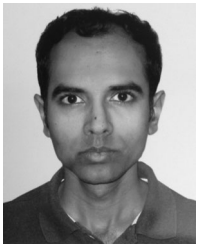
[164] M. Kulldorff, "A spatial scan statistic," *Commun. Statist.-Theory Methods*, vol. 26, no. 6, pp. 1481–1496, 1997.

[165] Covid-19 dashboard by the center for systems science and engineering (CSSE) at johns hopkins university (JHU). [Online]. Available: https://coronavirus.jhu.edu/map.html

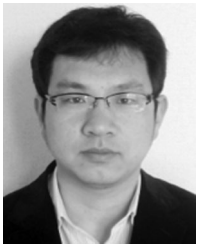
[166] Y. Xiong *et al.*, "Spatial statistics and influencing factors of the COVID-19 epidemic at both prefecture and county levels in hubei province, china," *Int. J. Environ. Res. Public Health*, vol. 17, no. 11, 2020, Art. no. 3903.

[167] Z.-L. Chen *et al.*, "Distribution of the COVID-19 epidemic and correlation with population emigration from wuhan, china," *Chin. Med. J.*, vol. 133, no. 9, pp. 1044–1050, 2020.

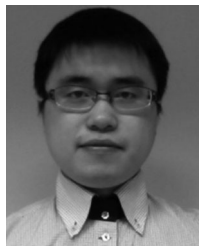
[168] Y. Han *et al.*, "Epidemiological assessment of imported coronavirus disease 2019 (COVID-19) cases in the most affected city outside of hubei province, wenzhou, china," *JAMA Netw. Open*, vol. 3, no. 4, pp. e206 785–e206 785, 2020.
[169] B. Gross *et al.*, "Spatio-temporal propagation of COVID-19 pandemics," *Europhysics Lett.*, vol. 131, no. 5, p. 58003, 2020.
[170] X. Zhang, H. Rao, Y. Wu, Y. Huang, and H. Dai, "Comparison of the spatiotemporal characteristics of the COVID-19 and sars outbreaks in mainland china," *BMC Infectious Diseases*, vol. 20, no. 1, pp. 1–7, 2020.
[171] E. Dong *et al.*, "An interactive web-based dashboard to track COVID-19 in real time," *Lancet Infect. Dis.*, vol. 20, no. 5, pp. 533–534, 2020.
[172] Z. Arab-Mazar *et al.*, "Mapping the incidence of the COVID-19 hotspot in iran – implications for travellers," *Travel Med. Infect. Dis.*, vol. 34, 2020, Art. no. 101630.
[173] D. Giuliani *et al.*, "Modelling and predicting the spatio-temporal spread of COVID-19 in italy," *BMC Infect. Dis.*, vol. 20, no. 700, pp. 1–10, 2020.
[174] A. Saha *et al.*, "Monitoring and epidemiological trends of coronavirus disease (COVID-19) around the world," *Matrix Sci. Medica*, vol. 4, no. 4, pp. 121–126, 2020.
[175] D. G. Marquez, "Guassion motion data," 2018. Accessed: Jan. 2018. [Online]. Available: https://citius.usc.es/investigacion/datasets/gaussianmotiondata



Amitangshu Pal received the PhD degree from the ECE Department, The University of North Carolina at Charlotte, in 2013. Currently he is an assistant professor with Indian Institute of Technology, Kanpur. He has published more than 70 conferences and journal papers. His current research interests include wireless sensor networks, reconfigurable optical networks, smart health-care, cyber-physical systems, mobile and pervasive computing, and cellular networks.



Junbo Wang received the PhD degree from the University of Aizu, Japan in computer science and engineering in 2011. He was a postdoctoral scholar and associate professor with the Univeristy of Aizu, Japan. Currently, he is an associate professor with the School of Intelligent Systems Engineering, Sun Yat-sen University, China. His research interests include collaborative ML, federated learning, fog computing, Big Data and privacy.



Yilang Wu received the BS degree from the School of Geographical Sciences in 2014, the MS degree in computing mechanism from the College of Civil Engineering, Kunming University of Science and Technology in 2017, and the PhD degree from the University of Aizu, Japan in computer science and engineering and 2021. He is a postdoc with the School of Intelligent System Engineering, Sun Yat-sen University, China. His research interests include edge computing, federated learning, and green computing.

Krishna Kant (Fellow, IEEE) is currently a professor with the Computer and Information Science Department, Temple University in Philadelphia, PA. He carries a combined 41 years of experience in academia, industry, and government. His research interests span a wide range including data center storage and networking, communications in challenging environments, and robustness and security in cyber and cyber-physical systems.

Zhi Liu (Senior Member, IEEE) received the PhD degree in informatics from the National Institute of Informatics. He is currently an associate professor with the University of Electro-Communications. His research interest include video network transmission and mobile edge computing. He is also an editorial board member of *Springer Wireless Networks* and *IEEE Open Journal of the Computer Society*.

Kento Sato received the BS degree from the Department of Information Science, Tokyo Tech, in 2008, and the MS and PhD degrees from the Department of Mathematical and Computing Sciences, Tokyo Tech, in 2010 and 2014, respectively. He is currently a team leader of High Performance Big Data Research Team in the Center for Computational Science, RIKEN. His research interests include distributed systems and parallel computing, particularly in high performance computing.