Incremental Spatial Clustering for Spatial Big Crowd Data in Evolving Disaster Scenario

Yilang Wu*, Amitangshu Pal†, Junbo Wang*, Krishna Kant†
*School of Computer Science and Engineering, University of Aizu, Aizuwakamatsu, Japan
†Computer and Information Sciences, Temple University, Philadelphia, USA
E-mail:{y-wu@ieee.org, amitangshu.pal@temple.edu, j-wang@u-aizu.ac.jp, kkant@temple.edu}

***Abstract*—Spatial clustering of the events scattered over a geographical region has many important applications, including the assessment of needs of the people affected by a disaster. In this paper we consider spatial clustering of social media data (e.g., tweets) generated by smart phones in the disaster region. Our goal in this context is to find high density areas within the affected area with abundance of messages concerning specific needs that we call simply as "situations". Unfortunately, a direct spatial clustering is not only unstable or unreliable in the presence of mobility or changing conditions but also fails to recognize the fact that the "situation" expressed by a tweet remains valid for some time beyond the time of its emission. We address this by associating a decay function with each information content and define an *incremental spatial clustering algorithm (ISCA)* based on the decay model. We study the performance of incremental clustering as a function of decay rate to provide insights into how it can be chosen appropriately for different situations.**

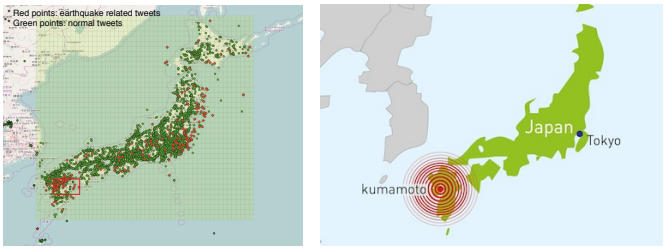
***Index Terms*—Spatial Big Data Analytics; Crowd Big Data; Incremental Spatial Clustering; Information Decay Model**

I. Introduction

Social networks, particularly Twitter, have become a popular platform for communicating information relevant for rescue and recovery during disasters [1,2]. For example, the following information was used during and after the Great East Japan Earthquake [3,4]: (1) The tag '#*j_j_helpme*' was used on Twitter following the earthquake and tsunami as a way for emergency personnel to rapidly identify people in need of rescue; (2) Google tweeted a link on Twitter to its Google Person Finder tool, which enables people to search for missing family members. An automated analysis of the twitter data to extract and understand the "situation" and the needs of the affected people can allow for quick and focused help in order to maximize benefit to the affected people [5]. Our focus in this paper is to identify these "situations" based on the analysis of the tweets in a way that recognizes that the situation expressed by the tweets remains valid over a period of time rather than being of instantaneous value only.

Spatial big data analytics is the science of finding hidden patterns in geospatial data. It has a lot of applications especially in emergency situations to support rescue and recovery activities during or following a disaster by understanding the "situations", such as injuries to the people, need for medical care, shortage food or water, etc. Spatial clustering is the process of grouping a set of spatial objects into clusters so that the objects within a cluster have high similarity than those across clusters. By applying spatial clustering, it becomes possible to discover hot-spot areas, which we define as high density clusters with a given situation (e.g., areas where many people are injured or need some type of help), and track the migration of such areas.

Fig. 1(a) shows the distribution of earthquake related tweets (with keywords 'earthquake', and '地震' which means disaster in Japanese) in the Kumamoto Earthquake that struck at Kumamoto City of Kumamoto Prefecture in Kyushu Region, Japan in 2016. Fig. 1(b) shows the shake map observed to the east of Kumamoto City [6]. This figure clearly shows that the distribution of the disaster related tweets can provide some useful information regrading the extent of damage, the real needs of the locals, etc. In this paper our main objective is to understand situations in a disaster area through spatial clustering of social big data. Such analysis can be useful to discover hot-spot areas (or the regions of interests), where emergency supplies need to be deployed as soon as possible.



(a) (b)

Fig. 1. Distribution of Kumamoto Earthquake (April-May, 2016). (a) Hotspot of Earthquake related Tweets after Kumamoto Earthquake and (b) the region of the epicenter as obtained the ground reality [6].

Spatial clustering is commonly used in the context of epidemiological applications [7,8], to help the epidemiologists identify the region of a possible outbreak. More specifically such applications measure the *overdensities* according to some measure of density measure (such as number of cases reported, units of medication sold, fraction of population affected etc.). However such techniques cannot be directly applied for finding the high-density regions in a disaster scenario by analyzing the tweets. This is because in an evolving disaster scenario, the situation and the needs over a region change over time, sometimes rather rapidly. Also, the incoming tweet data from

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

978-1-5386-5553-5/19/$31.00 ©2019 IEEE

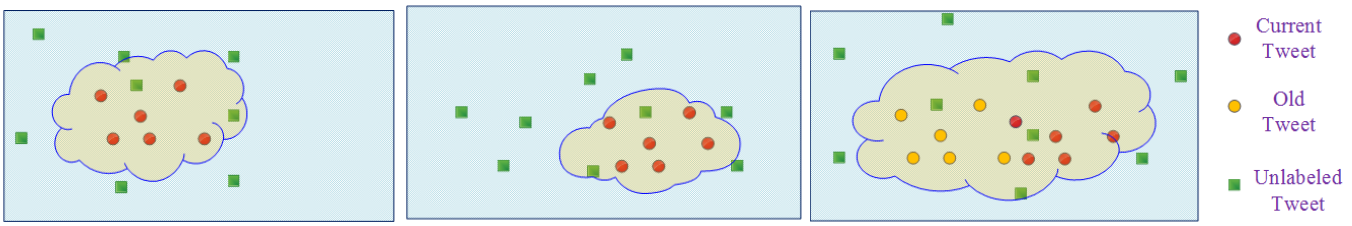


Fig. 2. An illustrative example for the proposed problem statement. Clustering the tweets at time instances (a) t and (b) $t+1$ results in unstable hot-spot identification. This can be alleviated with the (c) proposed energy decay model, where the tweets from the previous time instance are considered with less information energy. In this figure the circles and squares denote the tweets with and without situation respectively.

the crowd is highly dynamic, and the observed situation is intermittent, which becomes a big obstacle when trying to achieve reliable data analysis to support decision-making after a disaster occurs. Thus analyzing the tweets at different contiguous time scales result in a *drastic or haphazard* change in the hot-spot regions, which complicates the caregiver's operations for focusing on the regions of interest.

To address this problem we introduce a *temporal decay factor* along with the density measures. The intuition behind this temporal decay factor is that in a disaster scenario the needs over a region cannot be satisfied immediately or the importance of some tweets do not disappear instantly, rather gradually over time. In other words, the needs are often "sticky". For example, the affected population needs water, the need arises at the instant of the corresponding tweets, and the status sustains for a period. The obsolescence rate of the tweets varies depending on multiple factors such as the topic, how fast a need is served, or how fast the situation evolves. The proposed *incremental spatial clustering algorithm (ISCA)* can smoothen the situation estimation and make it more stable, and thus more useful from the help-provider's perspective. The proposed technique provides an adaptive visual observation of the dynamic change in situation in spatial big crowd data.

The paper is organized as follows. Section II describes the related works. The overall problem statement is discussed in Section III. The proposed ISCA scheme is presented in Section IV. Extensive evaluations are carried out in Section V. Conclusion and future works are highlighted in Section VI.

II. Related Work

Spatial clustering is the process of grouping a set of spatial objects into clusters so that objects within a cluster have a high similarity in comparison to one another, but are dissimilar to objects in other clusters. The traditional clustering methods such as k-means and k-medoids [9] break n objects into k clusters to optimize a given criterion; they can only output clusters of spherical shape and similar size. The density-based clustering methods such as DBSCAN and DBCLASD [10], which define a cluster to be a maximum set of density-connected points, is featured in detection on clusters of arbitrary shape and noise detection, but requires heuristic distance function that can not be applied equivalently to different data space. In comparison, the grid-based clustering methods [11] partition the data space into a certain number of cells, and perform clustering operations on the cells. The performance of grid-based clustering is affected by the choice of spatial-partition models, they can be more efficient when cells' scale is much less than data scale.

Evolving clustering has been studied in recent years to deal with stream data. In [12,13] TEDA (Typicality and Eccentricity Data Analytics) based evolving clustering are investigated. Based on the definitions of "typicality" and "eccentricity" in, affiliation of new data with existing clusters are first found. The affiliation can be one to many by adopting fuzzy theory. Finally mutual joint clusters are merged together by checking the number of common points. As the data comes in a streaming way, small clusters are first temporarily created to organize the received data in the clustering process [14,15]. A decay based function further judges whether the existing small clusters should be removed. Finally small clusters merge together to achieve the final clusters. However it would be more precise if decay function is adopted for each point, since the emergence time for each point is different in one cluster. In this paper, we propose an information energy model to represent the decay for each point, and the proposed spatial clustering algorithm works on the decay function instead of original point, to deal with updates of cluster for stream data.

The research problem in this paper is similar to the uncertainty problem in big data analytics [16]–[19]. Uncertainty-based big data learning is investigated in [16] where the authors first introduce several uncertainty definitions (Shannon entropy, Classification entropy, Fuzziness, etc.), and then discuss uncertainty based machine learning to solve the uncertainty problem. A scalable uncertainty-aware truth discovery mechanism is proposed in [19] which solves the uncertainty problem with source reliability and implements a parallel GPU based algorithm. In [17], the authors investigate a fuzzy theory based big data processing framework for uncertainties in transportation. Finally, reference [18] proposes a formal logical representation framework for video scenario recognition and a logic based uncertainty reasoning mechanism to infer the undergoing events. However our research problem is different in that the superfluous variation in the estimated situation exists in every data unit at every location and needs a new approach.

TABLE I
NOTATIONS AND PARAMETERS

Parameter	Description
$p \in P$	Point objects of Information Element
$s \in S$	Location of point objects
$z \in Z$	Zone or region
$t \in T$	Time instance
$\epsilon \in \mathcal{E}$	Keywords-based expression for situation ϵ
E	Energy Function for Situation Instance
D	Density Function
η^{λ}	Exponential decay rate

III. PROBLEM STATEMENT AND PROPOSED APPROACH

We first discuss the detailed research problem and our proposed approach using Fig. 2. Notations are listed in Table I. In the following a *topical situation* refers to the main situational information that is expressed or discussed in a twitter message. A *spatial or point object* is the location or place that the topic is associated with, here it refers to the geotag in a twitter message. The spatial objects in rectangular shape denote the point data that are not associated with a targeted situation, whereas the ones in circular shape are the points associated with the targeted situation. In a disaster scenario, the targeted situation indicates the damage level after the disaster, so that the rescue services can be provided as soon as possible.

As the tweet messages are generated continuously by the users, various opinions/messages exist for the same situation, which also evolves with time. Although a spatial clustering implemented at different time instants can find high density areas, such an approach will lead to unstable or unreliable identification of the hot-spot areas, as shown in Fig. 2(a)-(b). An unstable clustering is of little use to rescue operations which may take a significant amount of time to plan and execute. We address this issue by associating the messages with a decay function that assigns less weights to the older messages (shown in Fig. 2(c)), and propose an incremental clustering scheme so that the hot-spots regions can evolve steadily and smoothly over time.

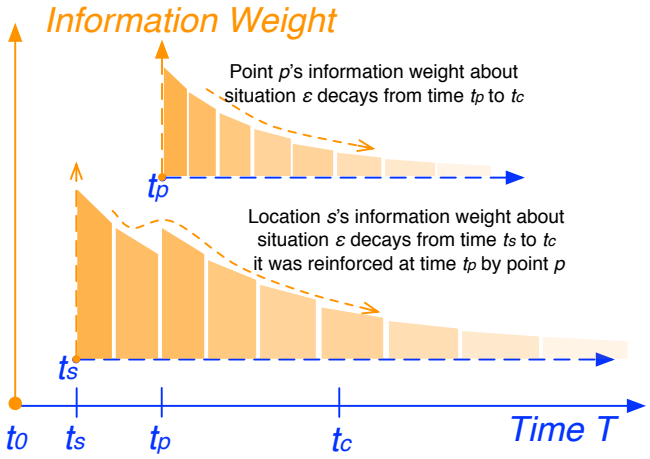


Fig. 3. Decay and Reinforcement of the Information Weight.

A. Information Energy of the Tweets

It's been observed repeatedly that much of the data has a popularity pattern: Very hot when the data is generated, and then the popularity wanes. The data may become hot again. For example, think of the files associated with the paper you are working on. It's popular for some days or weeks, and then there is no activity until you come back to it again. For measurement data coming from the cyberphysical infrastructure or physics experiments, the new data is hot only for some time. It may become hot again, but perhaps with decreasing peak popularity, i.e., short term decay functions superimposed on a long term decay function. In storage and other systems, the popularity is often captured using caching mechanisms such as LRU, but LRU is useful only at short time scales and small access granularities (blocks or objects). At longer time scales and large blobs of data, the stickiness may provide more insights.

In this context, we define the "information energy" of a tweet as the intensity of the tweet that is the highest power when a tweet originates, and then gradually fades over time. Information energy for a specific location can be accumulated with the other messages (or tweets) describing the same situation. An example of information energy is shown in Fig. 3. This ensures that the importance (or needs) of a tweet remains valid beyond the time when it emerges, however, with less energy.

Assume that the information energy for a point object p in spatial big crowd data at time instance t_c is denoted as $E_\epsilon(p, t_c)$. Also assume that the *temporal decay of the information energy* (TDIE) for each spatial data follows an exponential decay. That is,

$$E_\epsilon(p, t_c) = E_\epsilon(p, t_p) \cdot \eta^{-\lambda \cdot (t_c - t_p)} \tag{1}$$

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

Multiple messages, i.e., point objects, can originate at a location/region for a time period to describe a situation from different users. Generally speaking, if more number of messages originate at a location, higher will be the reliability that the situation happens in that location/region at that time instant. Therefore, TDIE can be calculated as an accumulation from multiple point objects of messages at a specific time instant, as shown in Fig. 3. Here $P(s, t_c - t_s)$ is the set of point objects located at location s, all of them are generated during the time span $(t_c - t_s)$ and labeled with situation ϵ. Let $E_\epsilon(s, t_p)$ denotes the information energy of points located at location s at its generating time t_p. Then,

$$E_\epsilon(s, t_c) = \sum\nolimits_{p \in P_\epsilon(s, t_c - t_s)} E_\epsilon(p, t_p) \cdot \eta^{-\lambda \cdot (t_c - t_p)} \tag{2}$$

Thus, the information energy is accumulated for each region Z at time t_c is given by

$$E_\epsilon(Z, t_c) = \sum\nolimits_{s \in Z} E_\epsilon(s, t_c) \tag{3}$$

B. Density Function to Detect Hot-spot Areas

We first define the *density function* D similar to [20]. The density $D(z)$ of a region z is a function of the total information energy corresponding to a situation $E(z)$, and the total population $P(z)$ of the region with following three properties:

1) For a fixed population, the density increases monotonically with information energy.
2) For a fixed information energy, the density decreases monotonically with the population.
3) For a fixed ratio $\frac{E}{P}$, the density increases monotonically with the population.

We choose a density measure based on *Kulldorff's spatial scan statistic* [21], which is commonly used in finding the significant spatial clusters in case of an emerging outbreaks. Assume that Z denotes the entire region. The density value for each region $z \in Z$ on time-stamp t_c for a specific situation ϵ is denoted as $D_\epsilon(z, t_c)$. Using the Kulldorff's spatial scan statistic [20], $D_\epsilon(z, t_c)$ can be calculated as

$$\begin{aligned} D_\epsilon(z,t_c) &= \log\left(\frac{L_\epsilon(z,t_c)}{L_\epsilon(Z,t_c)}\right) \\ &= E_\epsilon(z,t_c)\cdot\log\left(\frac{E_\epsilon(z,t_c)}{P(z,t_c)}\right) - E_\epsilon(z_0,t_c)\cdot\log\left(\frac{E_\epsilon(Z,t_c)}{P(Z,t_c)}\right) \\ &+ (E_\epsilon(Z,t_c) - E_\epsilon(z,t_c))\cdot\log\left(\frac{E_\epsilon(Z,t_c) - E_\epsilon(z,t_c)}{P(Z,t_c) - P(z,t_c)}\right) \end{aligned} \tag{4}$$

if $\frac{E_\epsilon(z,t_c)}{P(z,t_c)} > \frac{E_\epsilon(Z,t_c)}{P(Z,t_c)}$ and 0 otherwise. Here $L_\epsilon(z, t_c)$ is the likelihood function and, $P(z, t_c)$ is the total population in region z at time t_c. With this the spatial clustering algorithm needs to discover the hot-spot area z_C where

$$z_C = \arg\max_{\forall z \in Z}\left(D_\epsilon(z,t_c)\right) \tag{5}$$

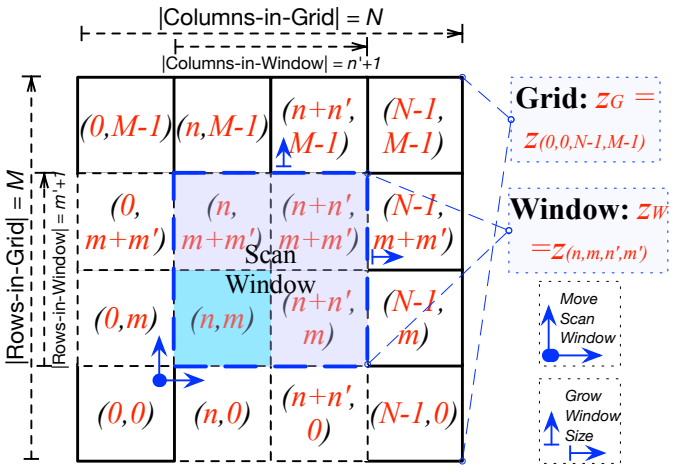


Fig. 4. Grid-based Partition and the Scan Window.

IV. Incremental Spatial Clustering

The proposed ISCA for evolving disaster 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. With these the incremental grid-based spatial scan is implemented as follows.

The commonly used grid-based partition works for the decay model because its partition boundary does not change over time. Thus the plain for the point data is partitioned into an $N \times M$ grid (see the left side of Fig. 4, and there are $N \times M$ cell-regions, and each cell-region is denoted as $z_{(n,m,0,0)}$, where $0 \le n < N$ and $0 \le m < M$), and the point data comes chronologically onto the plain. The spatial clustering component periodically (at time t_c) scans for the hot-spot of targeted situation ϵ, and compare the density value D_ϵ of each candidate region z (denoted as $z_{(n,m,n',m')}$, whose columns $\times$ rows equals to $[n...(n+n')] \times [m...(m+m')]$, and $0 \le n < n+n' < N$ and $0 \le m < m+m' < M$). As shown in Fig. 4, there are $(n+1) \times (m+1)$ number of regions in total with same shape $\left(\text{size}=(n'+1)\times(m'+1)\right)$ as that of $z_{(n,m,n',m')}$.

Algorithm 1 Incremental Grid-based Spatial Scan

```
1:  E_ε=InitMeasure(N, M), P=InitBase(N, M)  ▷ List of value for E_ε Measurement-, and P Basement-point data of all N × M cell-regions
2:  procedure SpatialScan((z_G = z_(0,0,N−1,M−1)))
3:      z_C = z_(0,0,0,0)                                   ▷ z_C: cluster
4:      for n' = 0; n' < N; n'++ do
5:          for m' = 0; m' < M; m'++ do
6:              for n = 0; n < N − n'; n++ do
7:                  for m = 0; m < M − m'; m++ do
8:                      P^z = 0, E^z_ε = 0
9:                      for i = n; i < n + n'; i++ do
10:                         for j = m; j < m + m'; j++ do
11:                             P^z += P[i][j]; E^z_ε += E_ε[i][j]
12:                             if E^z_ε/E_ε < 1 AND P^z/P < 1 AND E^z_ε/P^z < E_ε/P then
13:                                 z_W.d = E^z_ε log(E^z_ε/P^z) + (E_ε − E^z_ε) log((E_ε−E^z_ε)/(P−P^z)) − E_ε log(E_ε/P)
14:                                 if z_W.d > z.d then z_C = z_W
15:     return z_C
```

Here we introduce a naive implementation of the grid-based scan in Alg. 1, which return a rectangle as hot-spot. The performance of this algorithm could be improved for example by indexing the spatial data in tree-based model [20]. In this study, we mainly optimize the implementation of SDA to handle the information weight reinforcement and decay so as to investigate the stickiness of the incremental data.

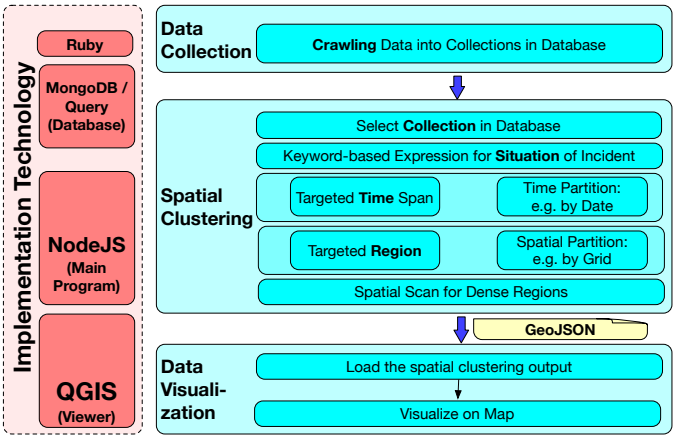


Fig. 5. Data Processing Flow for Incremental Clustering.

Fig. 5 shows the data processing flow in this system. The spatial clustering is implemented in JavaScript. It gets the collection of spatial point data (such as the Twitter data with geo-tag) from database, classify the point data as measurement data (to be labeled with situation ϵ) and basement data. The system will take temporal and spatial partition of the data into several time spans and sub-regions, and then apply spatial clustering to find the dense region of the targeted situation ϵ in each time span. The program then output the hot-spot in GeoJSON format, and visualized in QGIS tool.

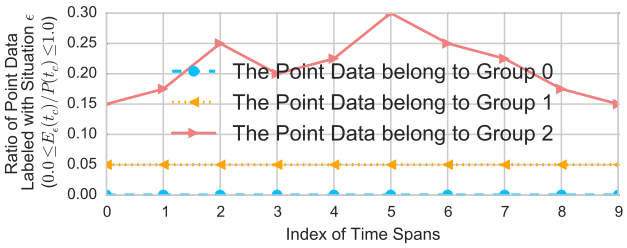


Fig. 6. The Ratio of Situation for 3 groups over time series.

V. Evaluation and Discussion

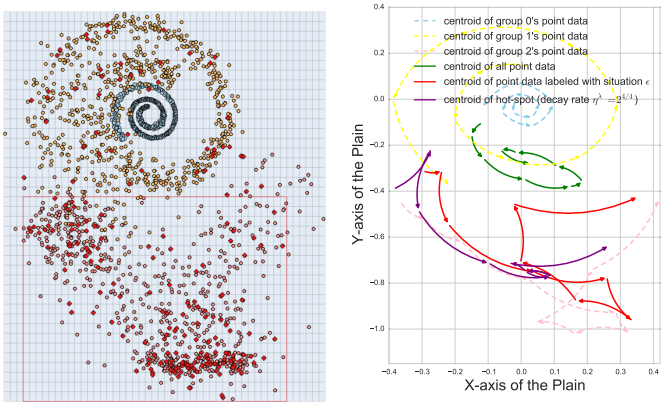
This section presents evaluation results of ISCA that are obtained from a synthetic dataset as well as from a real disaster-related social data.

A. Simulation on Synthetic Data Set

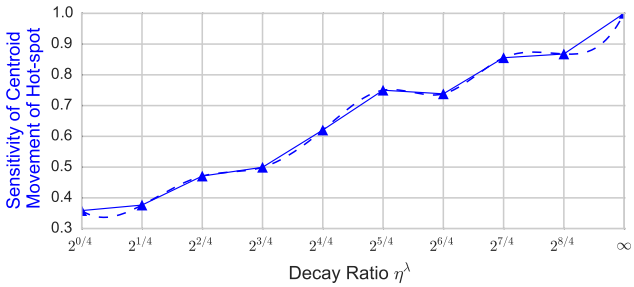
We first simulated ISCA using a synthetic database obtained from [22]. The database is composed of several datasets that model the temporal evolution of the information contents in a two dimensional space. The datasets were generated by Gaussian distributions whose mean and/or variance change over time. We use the *"3C2D2400Spiral"* dataset, which presents a helix alike movement of 3 clusters. These three clusters could be considered as three groups of population with dynamic ratios of the situation ϵ over the time series, as shown in Fig. 6.

The point data distribution of all three groups and the hot-spot over the time series are shown in Fig. and 7(a). We can observe that the hot-spot corresponding to these two data sets is mainly concentrated on Group 2, which has the highest percentage of situation labels among the three groups.

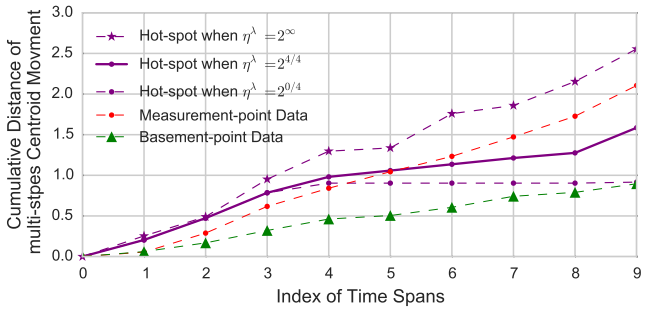
Metrics Used: We believe that the information contained in the crowd data, e.g., a tweet, about the situation remains valid for some time. For how long it remains valid is a property of the situation, and this property is modeled as energy decay in Section III-A. With the exponential decay base $\eta = 2$, different values of λ (the exponential decay exponent) represent different values of 'half-life' for the decay, i.e. the time when the weight of the tweet becomes one-half. In order to estimate the effect for the decay ratio λ on the spatial clustering result under different 'half-life' periods, we introduce a *sensitivity* metric. Assume that the centriod of the point objects at two time instances t and $t+1$ are μ_t and μ_{t+1}, and that of the hotspots are C_t and C_{t+1} respectively.



(a) Overall Hot-spot. (b) Centroid Movement.



(c) Sensitivity with different decay rates. Dotted (solid) line shows estimated curve (and its best fit curve) passing through the points.

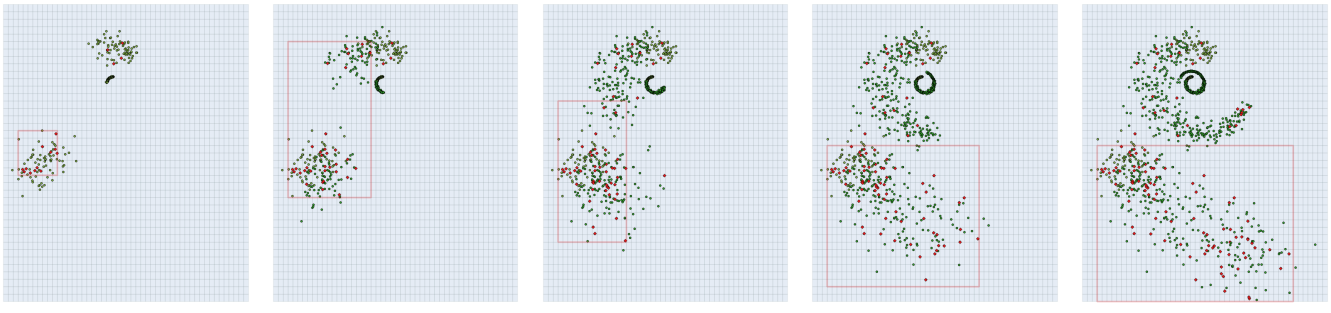


(d) Centroid Movement of Hot-spot for Helix Movement Data.

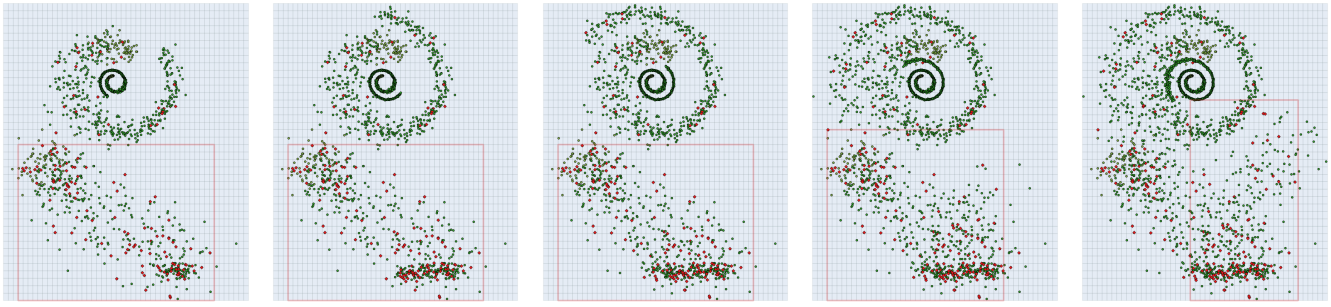
Fig. 7. Centroid Movement of Hot-spots for Helix Movement Data Set. In (a) the gray, yellow and red circles are the point objects of group 1, 2, 3.

Then the *unnormalized sensitivity* at those two instances is given by $\frac{\|C_t - C_{t+1}\|}{\|\mu_t - \mu_{t+1}\|}$, where $\|A - B\|$ denotes the absolute value of the Euclidean distance between two points A and B. The hot-spot movement is most sensitive when the decay ratio is infinite, so we normalize the value of *unnormalized sensitivity* with its maximum value (when $\lambda = \infty$). The *sensitivity* signifies whether the hot-spot movement is proportionate to the movement of the point objects.

Selection of Decay Rate η^λ for Adaptive Observation: Fig. 7(c) shows the changes in the sensitivity function for different value of λ with $\eta = 2$. A slower decay rate results in a smoother and more stable clustering, but at the cost of larger inaccuracies in characterizing the dynamics of the situation as the point data evolves. Here we consider λ as an appropriate

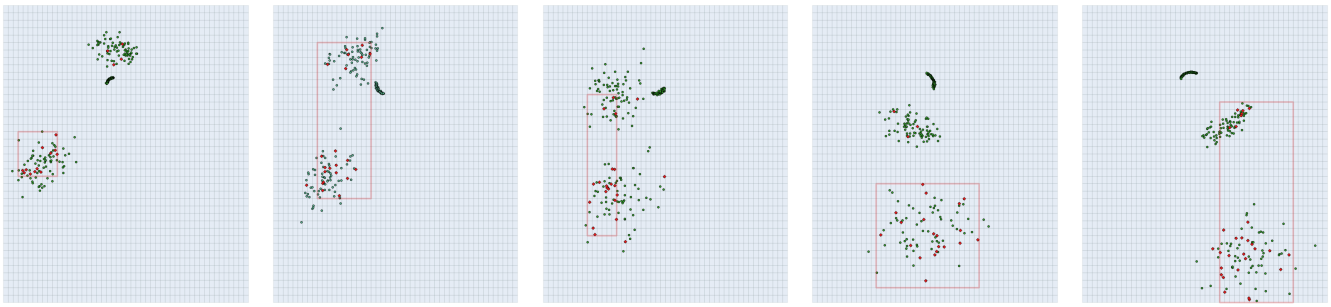


(a) Time 0 (b) Time 1 (c) Time 2 (d) Time 3 (e) Time 4

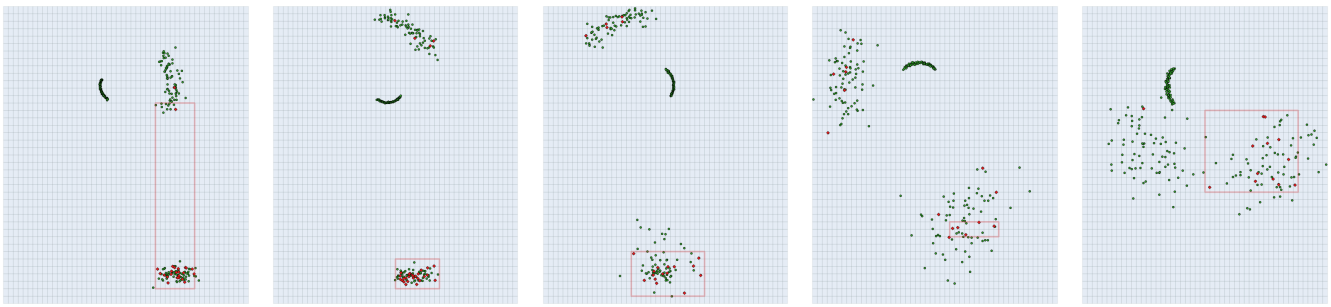


(f) Time 5 (g) Time 6 (h) Time 7 (i) Time 8 (j) Time 9

Fig. 8. Position of Hot-spot ($\eta^{\lambda} = 2^{4/4}$) for Helix Movement Data Set.



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



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

Fig. 9. Position of Hot-spot ($\eta^{\lambda} = 2^{\infty/4}$) for Helix Movement Data Set.

selection, where the sensitivity value is approximately close to its middle value for both data sets. Given $\eta^{\lambda} = 2$ for the point data, the situation is expected to experience its half-life during each time span, and the information weight of the situation becomes $1/2$ after each time span if there is no reinforcement.

We next investigate the spatial clustering result when $\eta^{\lambda} = 2^{4/4}$. Fig. 7(b) illustrates the centroid movement of the hot-spot (when $\eta^{\lambda} = 2^{4/4}$) in comparison with that of the point data. Fig. 7(d) show the sum of the Euclidean distances between centriods in between the time spans. The figures depict the case with $\eta^{\lambda} = 2^{4/4}$ as opposed to two extreme scenarios $\eta^{\lambda} = 2^{0/4}$ and $\eta^{\lambda} = 2^{\infty/4}$.

We visually illustrate the effect of our proposed decay model on the helix movement dataset using Fig. 8-9. These figures

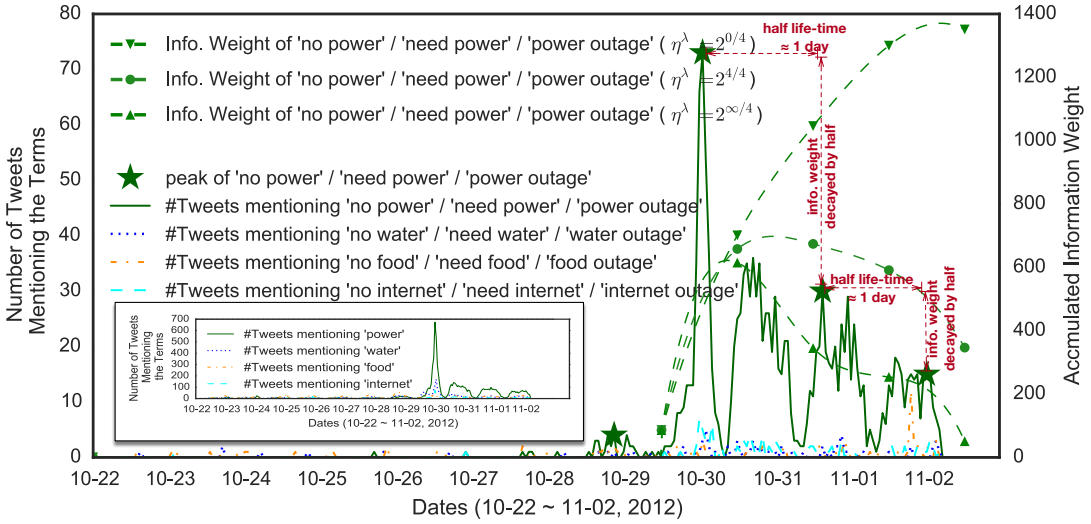


Fig. 10. Time Series of Tweets with Different Mentions nearby New York city, during Hurricane Sandy 2012.

illustrate the position, movement and coverage of the hot-spots when $\eta^{\lambda} = 2$ and 2^{∞} respectively. As expected the former is more sensitive than the latter, but the centroid movement of the hot-spot is still smooth. A slower decay rate results in a smoother and more stable clustering, but at the cost of larger inaccuracies in characterizing the dynamics of the situation as the topics evolve.

B. Observation to Sandy Hurricane 2012 Data Set

We used the geo-tagged twitter data located in New York city during the occurrence of Sandy Hurricane 2012. According to report [23], it hit New York city hard on Oct. 29th night, leaving hundreds of thousands without power. Regions of Lower Manhattan from Madison Square to the tip of the island was hit the hardest, with more than 0.24 million people without power as of noon on Nov. 1st.

We observed 440K Geo-tagged tweets located in New York city during the Hurricane. The power outage was the most concerned topic as shown in the time series in Fig. 10. The daily spike of tweets frequency related with 'power outage' is detected by using Python library '*peakutils*'. The tweets frequency experienced exponential decay, with its half-life time being close to 24 hours. We also plot the accumulated information weight of tweets by setting different decay rate, the $\eta^{\lambda} = 2^{1}$ looks most reasonable.

The Geo-tagged tweets distribution of power outage, as shown in Fig. 11(a), is crowded in Lower Manhattan. We use the grid-based spatial scan to track the daily hot-spots of power outage. The effects of different decay rates are shown in Fig. 11, 12, and 13. As expected the slower decay rate (or smaller λ) results in a smoother/stable clustering and vice versa.

VI. Conclusions

In this paper, we proposed an incremental spatial clustering algorithm based on information weight decay, in order to achieve a stable and reliable decision-making based on dynamic big crowd data. We have evaluated the proposed method by using a disaster related social data set as well as a synthetic data-set. The incremental spatial clustering using the decay model provides an adaptive observation of dynamic changes to the crowd situation. We introduce the metric of sensitivity to assist the user of this system to select proper decay rate to observe the spatial clustering result. In the future work, we will consider how to select decay rate for more real data sets. We will also work on improving the computational performance of ISCA to meet the emergency requirement in a disaster.

References

[1] Gao *et al.*, "Harnessing the crowdsourcing power of social media for disaster relief," *IEEE Intelligent Systems*, vol. 26, pp. 10–14, 2011.

[2] Yin *et al.*, "Using social media to enhance emergency situation awareness," *IEEE Intelligent Systems*, vol. 27, pp. 52–59, 2012.

[3] Slater *et al.*, "Social media, information, and political activism in japan's 3.11 crisis," *The Asia-Pacific Journal: Japan Focus*, 2012.

[4] Wang *et al.*, "Big data analytics for emergency communication networks: A survey," *IEEE Communications Surveys & Tutorials*, vol. 18, pp. 1758–1778, 2016.

[5] Wu *et al.*, "Disaster network evolution using dynamic clustering of twitter data," in *IEEE ICDCS Workshops*, 2017, pp. 348–353.

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

[7] Kulldorff, "Spatial scan statistics: models, calculations, and applications," in *Scan statistics and applications*. Springer, 1999, pp. 303–322.

[8] Auchincloss *et al.*, "A review of spatial methods in epidemiology, 2000–2010," *Annual review of public health*, vol. 33, pp. 107–122, 2012.

[9] Singh *et al.*, "K-means v/s k-medoids: A comparative study," in *National Conference on Recent Trends in Engineering & Technology*, 2011.

[10] Xu *et al.*, "A distribution-based clustering algorithm for mining in large spatial databases," in *IEEE ICDE*, 1998, pp. 324–331.

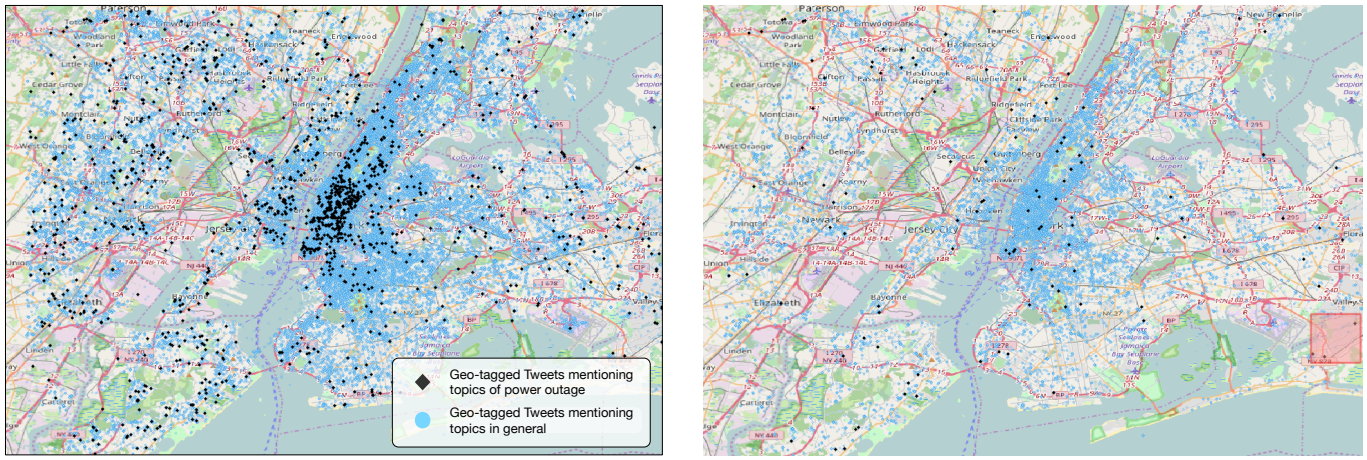
[11] Lin *et al.*, "A deflected grid-based algorithm for clustering analysis," *WSEAS Transactions on Computers*, vol. 7, pp. 125–132, 2008.

[12] Bezerra *et al.*, "A new evolving clustering algorithm for online data streams," in *IEEE EAIS*, 2016, pp. 162–168.

[13] Angelov, "Anomaly detection based on eccentricity analysis," in *IEEE EALS*, 2014, pp. 1–8.

[14] Baruah *et al.*, "Dynamically evolving clustering for data streams," *IEEE EAIS*, 2014.

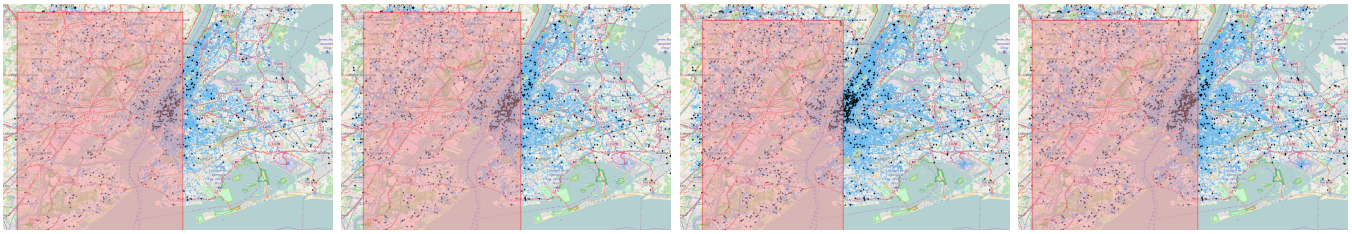
[15] Cao *et al.*, "Density-based clustering over an evolving data stream with noise," in *SIAM international conference on data mining*, 2006, pp. 328–339.



(a) Geo-tagged tweets Distribution of Power Outage (b) 10-29

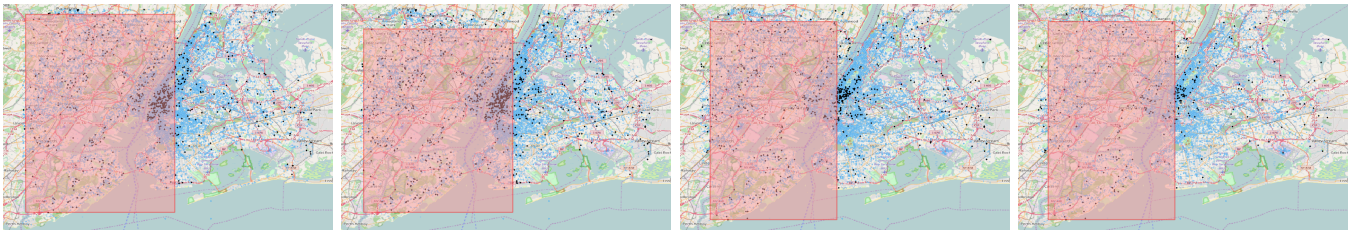
(c) 10-30 (d) 10-31 (e) 11-01 (f) 11-02

Fig. 11. (a) Spatial Distribution of Geo-tagged Tweets nearby New York, during Hurricane Sandy 2012. (b)-(f) Power Outage Hot-spot ($\eta^{\lambda} = 2^{\infty}$) Tracking.



(a) 10-30 (b) 10-31 (c) 11-01 (d) 11-02

Fig. 12. Power Outage Hot-spot ($\eta^{\lambda} = 2^{0}$) Tracking.



(a) 10-30 (b) 10-31 (c) 11-01 (d) 11-02

Fig. 13. Power Outage Hot-spot ($\eta^{\lambda} = 2^{1}$) Tracking.

[16] Wang *et al.*, "Learning from uncertainty for big data: Future analytical challenges and strategies," *IEEE Systems, Man, and Cybernetics Magazine*, vol. 2, pp. 26–31, 2016.

[17] Yang *et al.*, "A big-data processing framework for uncertainties in transportation data," in *IEEE FUZZ*, 2015, pp. 1–6.

[18] Chen *et al.*, "Uncertainty reasoning based formal framework for big video data understanding," in *IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT)*, 2014, pp. 487–494.

[19] Huang *et al.*, "Scalable uncertainty-aware truth discovery in big data social sensing applications for cyber-physical systems," *IEEE Transactions on Big Data*, 2017.

[20] Neill *et al.*, "Rapid detection of significant spatial clusters," in *ACM SIGKDD*, 2004, pp. 256–265.

[21] Kulldorff, "A spatial scan statistic," *Communications in Statistics-Theory and Methods*, vol. 26, pp. 1481–1496, 1997.

[22] Marquez, "Guassion motion data," https://citius.usc.es/investigacion/datasets/gaussianmotiondata, accessed: January, 2018.

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