Proceedings of 2016 IEEE 13th International Conference on Networking, Sensing, and Control
Mexico City, Mexico, April 28-30, 2016

# Analyzing the Evolution of Rare Events via Social Media Data and *k*-means Clustering Algorithm

Xiaoyu Sean Lu
Department of Electrical and Computer Engineering.
New Jersey Institute of Technology
Newark, NJ, USA
xl267@njit.edu

MengChu Zhou
Institute of Systems Engineering
Macau University of Science and Technology,
Macau, China & ECE, NJIT, Newark, NJ, USA
zhou@njit.edu

***Abstract* — Recently, many researchers attempt to find relationships between rare events and social media activities. This work proposes a data processing method based on the *k*-means clustering algorithm and analyze the evolution of a rare event via social media data. We use *k*-means twice in the spatial and time domains, respectively. The effectiveness of the method is verified by analyzing the damage of a hurricane named Sandy that occurred in 2012. The data set with respect to Sandy is obtained from a very popular social media, Twitter. The results show that our method can precisely predicate the accurate evolution of the hurricane, i.e., the affected place, time and severity. Besides, two new concepts, growth ratio and DRR rate, are presented to analyze the dataset in the time domain.**

***Keywords* — Big data, rare events, data processing, *k*-means clustering.**

## I. Introduction

Social media provides a platform that broadcasts and exchanges information among users. This information may cover the important events and human being's attitudes at that time span. Social media data contains features that lead a way to analyze rare event-related information such as relationship between happiness and mobility patterns [Frank, 2013], and between tourist origins and attractions [Wood, 2013]. A rare event, such as the occurrence of hurricane Sandy studied in [Guan, 2014], does not only impact the real world, but also effect the distribution of disaster-related messages via social media. In this work, we utilize hurricane Sandy as our background and study the patterns in the duration of hurricane Sandy.

A disaster is regarded as a disruption on the earth and involves environmental or economic loss. A serious disaster greatly threatens human beings' and animals' lives. This type of disaster is a rare event, since it occurs rarely but has really serious destructions. [Fritz, 1961] and [Quarantelli, 1977] start to describe disasters as social events. [Tierney, 2007] presents a deeper concept that physical events alone do not constitute disasters unless they affect human beings and social systems in negative ways. Thus, a disaster is not an isolated event and its crisis arises because of vulnerability in human beings and natural and technological systems [Cutter, 2006]. Chen *et al*. [2013] emphasize that a focus on social disruptions is a key to understand and assess a disaster. Their work connects the physical disasters and human beings' social activities. In [Chen, 2013], a framework is proposed for assessing disasters' crisis in a human system environment and helps one understand the evolution of a disaster. A new concept named the degree of disaster towards human system is given in [Chen, 2013]. Guan [2014] adopts the information from Twitter to analyze and assess the disaster, 2012 hurricane Sandy. When hurricane Sandy landed the United States, a large population of users posted their messages through their Twitters. This type of social media gives a feasible way to obtain disaster-related messages that can be used to assess and evaluate the impacts of a disaster. However, choosing a reliable metric is an important key to estimate the influence of an event. It is hard to just count the total number of event-related messages during a specific time span, since the number of tweets is randomly generated. For example, when there is a high level basketball competition on one day's night in a city, many fans may post their messages via their social media platforms. At the same time, if this day is snowy, some people may post the bad weather-related messages as well. Both of them are able to bring a high message count on that day, but it is hard to distinguish whether people pay more attention to the basketball competition or the bad weather in that area. With this consideration, we cannot simply count the number of messages. In order to assess the impacts of an event through the social media data, a metric, pioneered in [Guan, 2014] disaster-related ratio (DRR), illustrates the relationship between a disaster and social media activities. DRR calculates the ratio between numbers of related and unrelated messages at a same time span in the same area. In their work, the result shows that the DRR is high when the disaster is occurring, otherwise it is low. However, their work only calculates a day's DRR in a specific city. Its curves can only describe the roughly impacted date by the disaster. The errors are large since their work only finds the peak dates and cannot get accurate time points. In order to conquer this issue and increase the accuracy, we use the *k*-means clustering algorithm [Arthur, 2007] to find the more accurate time that is impacted by hurricane Sandy. Furthermore, we present a spatial pattern that shows the impacted area by DRRs during the specific time span. Two new concept, growth ratio and DRR velocity is given to analyze the pattern on time domain and describe an influence of a rare event.

The rest of the paper is organized as follows. Section II briefly describes the *k*-means clustering algorithm. The proposed data processing method is illustrated in Section III. The experimental results are given in Section VI to explain the effectiveness of our method. Section V draws our conclusions and presents future works.

## II. *K*-means Clustering Algorithm

Since its simplicity, efficiency and easy implementation, a *k*-means clustering algorithm [Jain, 2010] and [Arthur, 2007] becomes one of the most popular and simple clustering

978-1-4673-9975-3/16/$31.00 ©2016 IEEE

methods and has been independently used in different scientific fields. For example, it is used in the texture and image segmentation in [Likas, 2003] and [Tatiraju, 2008]. Meanwhile, [Oyelade, 2010] utilizes *k*-means to predict students' academic performance. The *k*-means algorithm is formally described as follows:

Let $X = \{x_i\} \subset R^d$, $i \in \{1, 2, \ldots, n\}$, be the set of n d-dimensional points where $R^d$ denotes a d-dimensional data set, and $C = \{C_k\}$ , $k \in \{1, 2, \ldots, K\}$, be a set of $K$ clusters that partition $X$ where $K$ is a positive integer greater than one. The mean $u_k$ of cluster $C_k$ is defined as:

$$u_k = \frac{1}{|C_k|} \sum_{x \in C_k} x \quad (1)$$

where $|C_k|$ is the cardinality of set $C_k$. The squared error between $u_k$ and the points in $C_k$ is defined as:

$$\emptyset(C_k) = \sum_{x_i \in C_k} \|x_i - u_k\|^2 \quad (2)$$

The objective of the *k*-means algorithm is to find the minimized sum of squared errors over all $K$ clusters and can be computed as:

$$\emptyset_{min} = \min\left(\sum_{k=1}^{K} \sum_{x_i \in C_k} \|x_i - u_k\|^2\right) \quad (3)$$

The main steps of the *k*-means algorithm are shown as follows [Jain, 1988] and [Jain, 2010]:

1. Select an initial partition with $K$ clusters;
2. Compute a new partition by assigning each pattern to its closest cluster center;
3. Compute new cluster centers according to (1); and
4. Repeat Steps 2 and 3 until cluster membership is stabilized.

## III. Data Processing Method

### A. Data Processing Method

Since hurricane Sandy went through our huge selected region over time, we expect to divide up them into several sub-regions and study these small ones. *k*-means provides a method that can cut and combine those nearest tweets. Thus, this section describes a data processing and shows the *k*-means-based method. Because some tweets were posted too far early or too late before or after the Sandy landed, these are filtered and deleted in our filtered dataset. Fig. 1 describes the procedure of data processing that has two main steps.

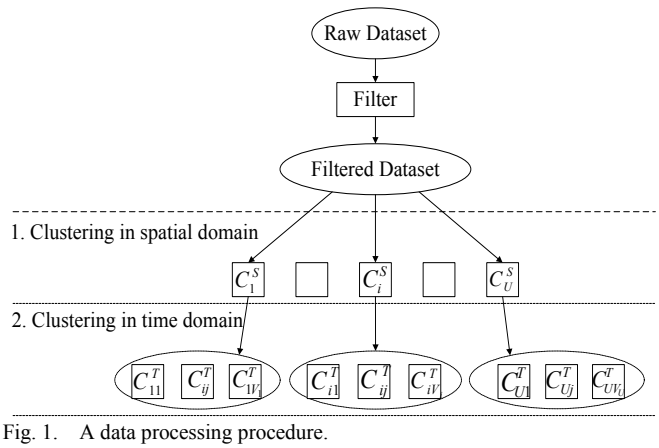


Fig. 1. A data processing procedure.

The first step clusters the close tweets into spatial clusters based on their locations or geo-coordinates and obtains the coordinate mean of each cluster's centroid. Let $C^S = \{C_1^S, C_2^S, \ldots, C_i^S, \ldots, C_U^S\}$, $i \in \{1, 2, \ldots, U\}$, be a set of spatial clusters that partition set $X$ into $U \geq 2$ spatial clusters. $k_1 \in N^+$ is the number of clusters. Each element $C_i^S \in C^S$ represents an individual spatial cluster. The mean of a spatial cluster $C_i^S$ is denoted by $u_i^S$. In a physical meaning, $u_i^S$ also denotes a pair of geo-coordinates. The second step clusters the time of the posted tweets in each spatial cluster. $C_i^T = \{C_{i1}^T, C_{i2}^T, \ldots, C_{ij}^T, \ldots, C_{iV_i}^T\}$, $j \in \{1, 2, \ldots, V_i\}$, represents a set of time-based clusters and partitions set $C_i^S$ into $V_i \geq 2$ time-based clusters. The mean of a time-based cluster $C_{ij}^T$ is denoted by $u_{ij}^T$ and represents a time point. According to our proposed procedure, the Twitter dataset is partitioned into $U$ spatial-based clusters. Then, each spatial cluster is divided into $V_i$ time-based clusters.

### B. Growth ratio and DRR rate

Now we present two concepts that are called growth ratio and DRR rate. The growth ratio computes an incremental ratio between its initial DRR and its peak value in a given area. It is defined as follows:

$$\rho = DRR_{peak}/DRR_{initial} \quad (4)$$

where $DRR_{initial}$ denotes its initial DRR and $DRR_{peak}$ is its peak value. DRR rate computes the speed that a DRR increases or decreases in a given area. It is given as follows:

$$\gamma_{ij} = (DRR_i - DRR_j)/(t_i - t_j) \quad (5)$$

where $t_i$ and $t_j$ represent respectively the $i$-th and $j$-th time points, and $i, j \in \{1, 2, \ldots\}$. $DRR_i$ and $DRR_j$ are the ratios that are associated with $t_i$ and $t_j$, respectively. This rate denotes that the distribution speed of messages in a given area. If $\gamma_{ij}$ is positive, it is called a rate of increase and indicates that *DRR* is increasing. Otherwise, it is decreasing and called a rate of decrease. If $\gamma_{ij}$ is positive and huge, it means that the disaster-related messages are produced and distributed quickly between the two time points. Otherwise, the messages are slowly generated. Similarly, if $\gamma_{ij}$ is negative and small, it means that disaster-related messages are reduced quickly.

## IV. Data Source and Experimental Results

This section first introduces our data set that is from a popular social media website named Twitter. Then the proposed method is used to obtain the relationship between the social media and impacts of hurricane Sandy.

### A. Data Collection and Analysis

Twitter is a well-known online social media platform and enables users to post maximum 140-character messages, also called tweets. Created and launched in 2006, Twitter rapidly gained more than 100 million users posting 340 million tweets per day in 2012. The number of users reached more than 500 million in 2015. Thus, Twitter is selected as the data source in [Guan, 2014]. Twitter opens and provides its resource to developers via Twitter search Application Programming Interface (API). With the help of API, totally more than 9

million records are crawled [Guan, 2014]. Each record has its identifier, geographic coordinates, posted location and time, contents and categories.

We specify the Northeast region of the United States as our concerned region. It contains some states with a large population, such as New Jersey and New York, and some large cities, such as Boston, New York City and Washington DC. This region was badly impacted by the hurricane and brought us a sufficiently large disaster-related dataset. Temporally, our study's time period spans from Oct 27, when the storm warning was issued, to Nov 7, 2012, a week after the hurricane landed the selected region. Meanwhile, spatial geo-coordinates are limited by the latitudes from 37.84 to 42.86 and the longitudes from -70.89 to -78.8. After this filtering, it returns us about 1,281,000 tweets. Then, we use keywords to filter out those disaster-unrelated tweets. These keywords are "Sandy", "hurricane", "storm" and "rigov" as also used in [Guan, 2014]. This step returns about 74,000 tweets that are related to hurricane Sandy.

### *B. Experimental results*

Based on the procedure in Fig. 1, the first step adopts the *k*-means clustering algorithm with parameter $k = 50$. This step partitions the disaster-related tweets into 50 clusters in the spatial domain. If a cluster's DRR is high, it means that the corresponding physical area is highly impacted by hurricane Sandy and vice versa. In Fig. 2, we use a point to represent geo-coordinates of a cluster's center and the digital number next to it is its identifier. If the digital number is small, it represents that the DRR of its corresponding cluster is large; otherwise, it is small. Fig. 3 shows the DRR values of the points in Fig. 2.

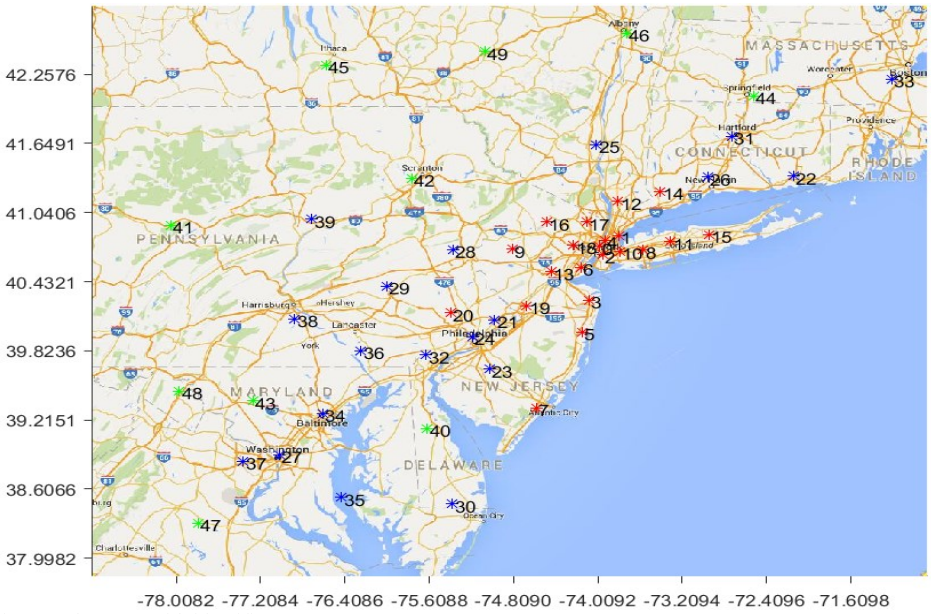


Figure 2. Hurricane Sandy impacted pattern with identifiers.

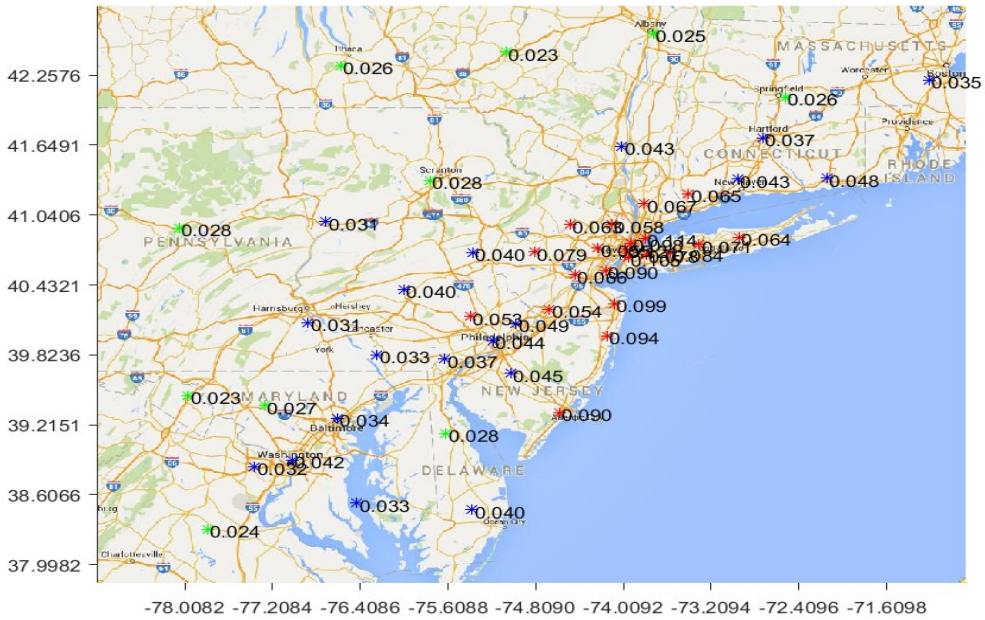


Figure 3. Hurricane Sandy impacted pattern with DRRs.

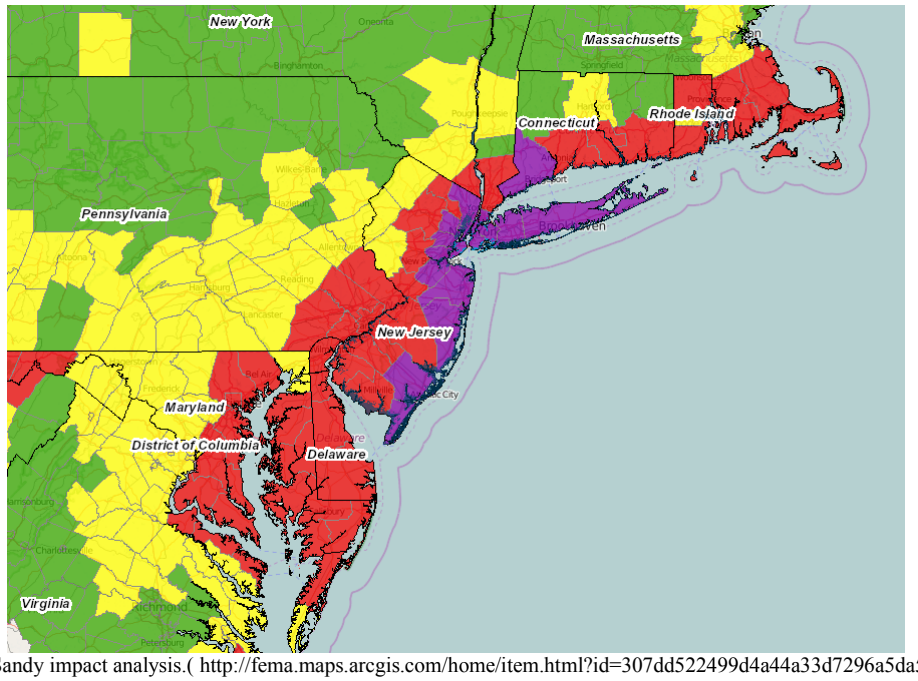

Figure 4. FEMA hurricane Sandy impact analysis.( http://fema.maps.arcgis.com/home/item.html?id=307dd522499d4a44a33d7296a5da5ea0 )

In both figures, points are marked with colors where red, blue, and green, representing that the corresponding clusters' DRRs are greater than 0.05, between 0.03 to 0.05, and less than or equal to 0.03, respectively. The two values, 0.05 and 0.03, are specified as thresholds and partition 50 points into 3 levels. Red points are the highly impacted regions; blue ones are moderately impacted regions; and green ones are slightly impacted ones. This matches a hurricane Sandy's impact pattern with 4 levels: very high (purple), high (red), moderate (yellow) and low (green), given by Federal Emergency Management Agency (FEMA) as shown in Fig. 4.

In Fig. 4, very high (purple) area means that greater than 10,000 of county population is exposed to surge; high (red) one indicates that 500 - 10,000 of county population exposed to surge, or modeled wind damages are greater than $100M, or high precipitation (>8"); moderate (yellow) one represents that 100 - 500 of county population exposed to surge, or modeled wind damages are between $10M and $100M, or medium precipitation (4" to 8"); and low (Green) one indicates that there are no surge impacts.

Red points in Fig. 2 are located in the purple area in Fig. 4. Blue points in Fig. 2 match the red area in Fig. 4. Similarly, green points are located in the yellow area. However, a very few of them may not match it since the thresholds may not be accurate.

The second step clusters the close tweets based on the time. In this step, we set $k$ = 10. Each spatial cluster runs the $k$-means algorithm once and obtains its corresponding curve in the time domain as shown in Figs. 5-7.

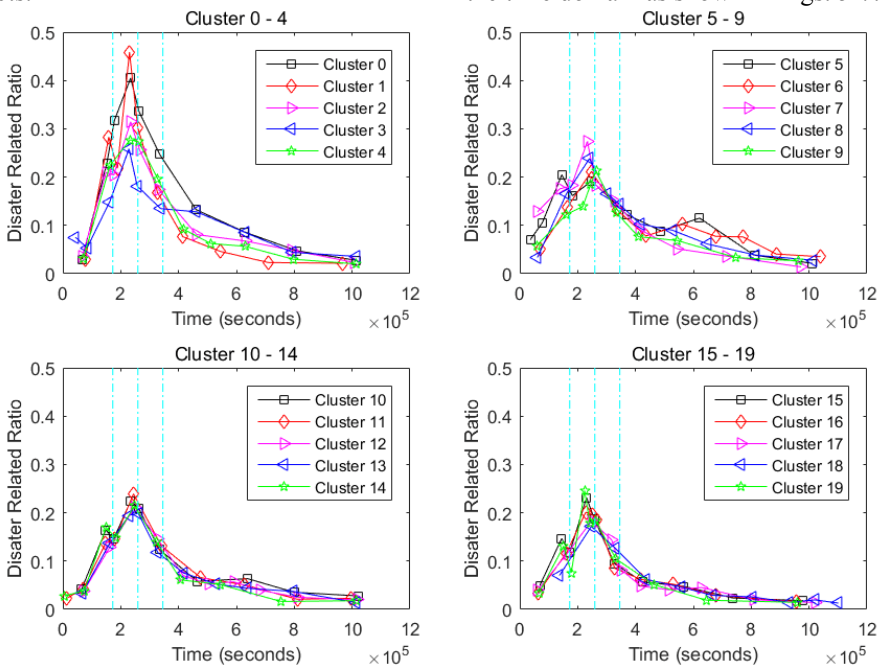

Figure 5. Curves of cluster 0-19 in the time domain.

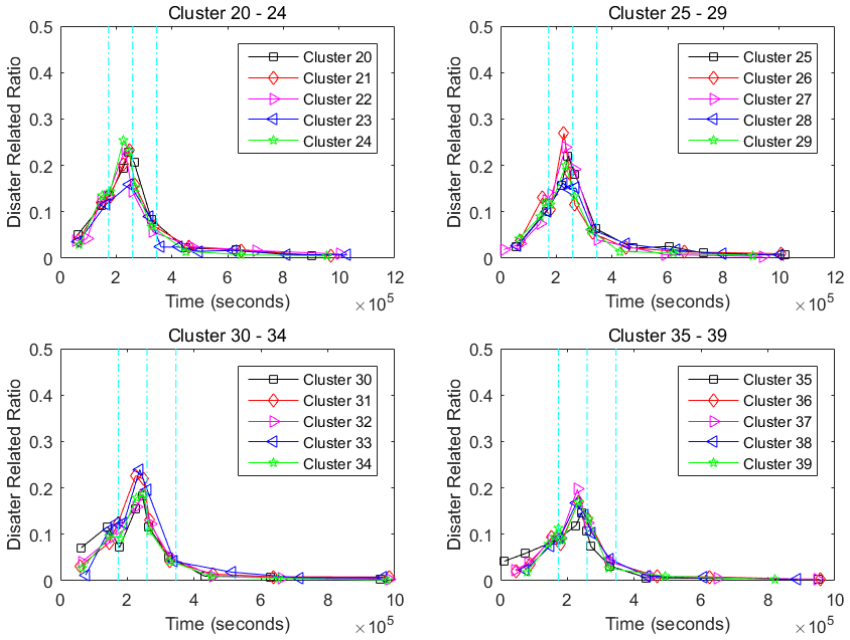


Figure 6. Curves of cluster 20-39 in the time domain.

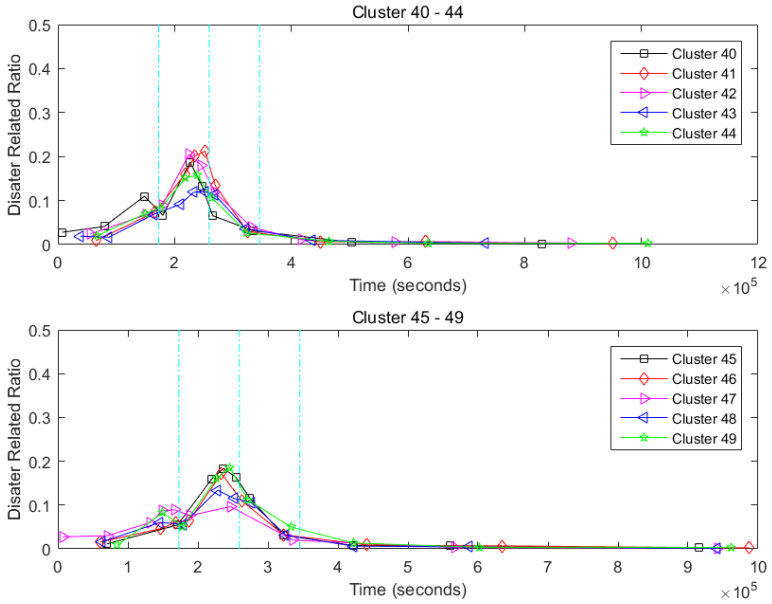


Figure 7. Curves of cluster 40-49 in the time domain.

Figs. 5-7 show the variants of DRRs in the time domain for each spatial cluster. In order to compute time points conveniently, we convert the tweets' posted time into seconds. Since the dataset is filtered and retains the tweets posted from Oct 27 to Nov 7, we set our starting time at 00:00:00, Oct 27 and it is the original point in Figs. 5-7. For example, a tweet was posted at 10:20:00, Oct 29. It is converted into 210,000 seconds. Then our curves are obtained by using seconds as the unit of the horizontal axis. Note that three dashed lines label two special dates, Oct 29th and 30th, and are converted into 172,800, 259,200 and 345,600 seconds, respectively. Except the curve of cluster 5, other curves' peaks are located on Oct 29 which means that it is a highly impacted date in their corresponding areas. The peak of the exception, cluster 5, appears earlier than Oct 29 that may be caused by the early warning messages near the coast. In other words, there are 98% curves that have their peak values on Oct 29. In addition, we obtain not only the peaks on Oct 29, but also the accurate time that hurricane Sandy impacted an area. For example, cluster 17's corresponding physical area was impacted by hurricane Sandy seriously at 20:49:50, Oct 29 (234,290 seconds). Cluster 18's corresponding physical area was impacted at 20:40:10, Oct 29. The physical location of cluster 17 is near Passaic County, NJ and cluster 18 is next to Newark Liberty International Airport, New Jersey, USA. National Weather Service ( http://www.weather.gov/okx/HurricaneSandy ) also gives its rough estimated impacted time in these two areas, i.e., 21:30:00, Oct 29 and 19:51:00, Oct 29. The differences between our measured time points and rough estimated ones are less than one hour. It reflects that our method is efficient and more accurate than Guan's work that has the time accuracy measured by a full day (24 hours).

We compute the means of red, blue, and green points in the time domain, and obtain three types of curves, respectively, as shown in Fig. 8. The bold, dashed, and dotted curves correspond to the highly, moderately and slightly impacted areas, respectively. Obviously, the bold curve has the largest peak value and the dotted one has the lowest peak value. The

growth ratios among the three curves are 4.64, 5.51, and 8.82 times, respectively. The dotted one has the largest growth ratio, but the bold one has the lowest growth ratio. It concludes that slightly impacted area has a higher growth ratio, but a highly impacted area has a lower growth ratio. It is acceptable, since the highly impacted area starts with a high initial DRR. This value may be caused by high attentions and alerts among individuals, news media and government. The rates of increase along bold, dashed, and dotted curves are computed between their first and peak DRRs and are $8.8\times10^{-7}$ $s^{-1}$, $7.65\times10^{-7}$ $s^{-1}$ and $7.66\times10^{-7}$ $s^{-1}$, respectively. The rate of increase along the bold one is much higher than other two, because the red area is badly impacted. Meanwhile, the other two are very close. The rates of decrease along the three curves are $2.64\times10^{-7}$ $s^{-1}$, $2.38\times10^{-7}$ $s^{-1}$ and $2.28\times10^{-7}$ $s^{-1}$, respectively and are very close. Thus, if these three impacted area recover to the same DRR level, the bold curve needs a longer time. This conclusion is obvious based on Fig. 8 as well. It makes sense because the bold curve corresponds to the impacted area that has serious damages and needs the longest time to recover. On the contrary, the dotted one reflects that it needs less time to recover. In addition, no matter which curve it is, its rate increase is almost three times more than the corresponding of decrease. This concludes that when a disaster comes, an area is impacted instantly, but needs more time to recover.

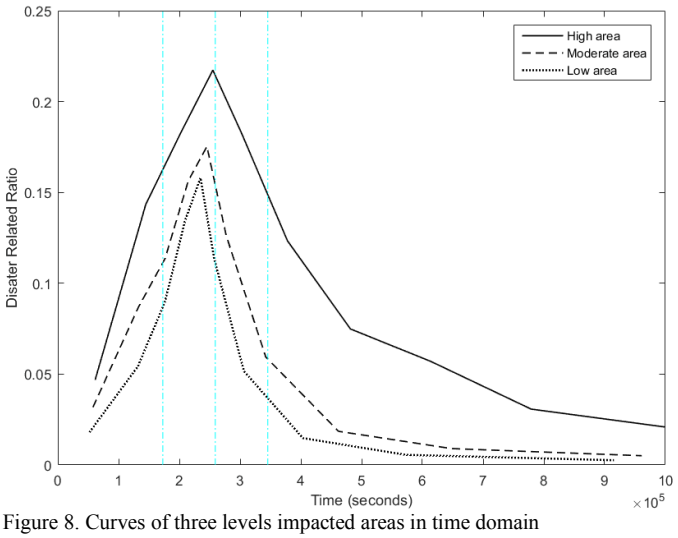

Figure 8. Curves of three levels impacted areas in time domain

## V. Conclusion

This paper proposes a data processing method based on *k*-means clustering algorithm. The relationship between social media activities and a rare event can be obtained. It reveals the impacted areas in the spatial domain and time points in the time domain, i.e., relatively more accurate impacted locations and time points of the hurricane are obtained. Furthermore, our proposed growth ratio and DRR rate are proved useful for analyzing the relationship between social media activities and a rare event.

In the future work, we will attempt to find a better way that can obtain more accurate time points and locations. Meanwhile, the growth ratio and DRR rate need to be used in multiple clustering partitions and prove their efficiency and accuracy when they are used to analyze a rare event. Moreover, a rare event can be treated as an input activates a human system and generate its outputs. We will build models and systems that reflect its impacts in both spatial and time domains. Some recently emerging intelligent optimization methods, e.g., [13]-[20], will be used to perform big media data analysis.

## Acknowledgement

This work is in part supported by the U.S. NSF under Grant CMMI-1162482 and by FDCT (Fundo para o Desenvolvimento das Ciencias e da Tecnologia) under Grant 119/2014/A3.

## References

[1] X. Guan and C. Chen, "Using social media data to understand and assess disasters," *Nat. Hazards*, vol. 74, pp. 837–850, 2014.

[2] E. L. Quarantelliand and R. R. Dynes, "Response to social crisis and disaster," *Ann. Rev. Sociol.*, vol. 3, pp. 23-49, 1977.

[3] Cutter S (2006) Are we asking the right question? In: Perry R, Quarantelli EL (eds) What is a disaster? New answers to old questions. *Xlibris* 2006, Lexington, pp 39–48

[4] K. J. Tierney, "From the margins to the mainstream? Disaster research at the crossroads," *Ann. Rev. Sociol.* vol. 33, pp. 503-525, 2007.

[5] Fritz CE (1961) Disasters. In: Merton RK, Nisbet RA (eds) Contemporary social problems. *University of California Press*, Riverside, pp 97–122

[6] C. Chen, D. Neal, and M. Zhou "Understanding the evolution of a disaster-a framework for assessing crisis in a system environment (FACSE)," *Nat. Hazards* vol. 65, pp. 407-422, 2013.

[7] D. Arthur and S. Vassilvitskii, "k-means++: The advantages of careful seeding," In *Proc. of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, 2007, Philadelphia, PA, USA pp. 1027-1035.

[8] A. K. Jain, "Data clustering: 50 years beyond K-means," *Pattern. recognition letters*, vol. *31*, no. 8, pp. 651-666, 2010.

[9] S. Tatiraju, and A. Mehta, *Image Segmentation using k-means clustering, EM and Normalized Cuts* Department of EECS, 2008, 1: 1-7.

[10] O. J. Oyelade, O. O. Oladipupo, and I. C. Obagbuwa, "Application of k Means Clustering algorithm for prediction of Students Academic Performance," *(IJCSIS) International Journal of Computer Science and Information Security*, vol. 7, no. 1, pp. 292–295, 2010.

[11] A. Likas, N. Vlassis, and J. J. Verbeek, "The global k-means clustering algorithm," *Pattern recognition*, vol. 36, no. 2, pp. 451-461, 2003.

[12] A. K. Jain and R. C. Dubes, Algorithms for clustering data, *Prentice hall*, 1988.

[13] Q. Kang, M. C. Zhou, J. An, and Q. Wu, "Swarm Intelligence Approaches to Optimal Power Flow Problem with Distributed Generator Failures in Power Networks," *IEEE Trans. on Automation Science and Engineering*, 10(2), pp. 343-353, April 2013.

[14] W. Dong and M. C. Zhou, "Gaussian Classifier-based Evolutionary Strategy for Multimodal Optimization," *IEEE Trans. on IEEE Transactions on Neural Networks and Learning Systems*, 25(6), pp. 1200 – 1216, June 2014.

[15] X. Zuo, et al., "Vehicle Scheduling of Urban Bus Line via an Improved Multi-objective Genetic Algorithm," *IEEE Trans. on Intelligent Transportation Systems*, 16(2), pp. 1030-1041, April 2015.

[16] J. Li, et al., "Colored Traveling Salesman Problem," in *IEEE Transactions on Cybernetics,* 45(11), pp. 2390 – 2401, Nov. 2015.

[17] A. Che, P. Wu, F. Chu, M. C. Zhou, "Improved Quantum-Inspired Evolutionary Algorithm for Large-Size Lane Reservation," *IEEE Transactions on Systems, Man, and Cybernetics: Systems,* 45(12), pp. 1535 - 1548, Dec. 2015.

[18] J. Li, J. Zhang, C. Jiang and M. C. Zhou, "Composite Particle Swarm Optimizer with Historical Memory for Function Optimization," *IEEE Trans. on Cybernetics*, 45(10), pp. 2350 – 2363, Oct. 2015.

[19] X. Liang, W. Li, Y. Zhang, and M. Zhou, "An adaptive particle swarm optimization method based on clustering," *Soft Comput*ing, 19(2), pp. 431-448, Feb. 2015.

[20] X. Luo, et al. "Generating Highly Accurate Predictions for Missing QoS Data via Aggregating Nonnegative Latent Factor Models," *IEEE Transactions on Neural Networks and Learning Systems*, 27(3), 524 - 537, Mar. 2016.