Mitigating the Impact of Data Sampling on Social Media Analysis and Mining

Kuai Xu, *Senior Member, IEEE*, Feng Wang , *Member, IEEE*, Haiyan Wang,
Yufang Wang, and Ying Zhang

***Abstract*— The last decade has witnessed the explosive growth of online social media in users and contents. Due to the unprecedented scale and the cascading power of the underlying social networks, social media has created a new paradigm for sharing information, broadcasting breaking news, and reporting real-time events by any user from anywhere at any time. Many popular social media sites including Twitter provide streaming data services by standard APIs to the broad researcher and developer communities. Given the sheer data volume, rapid velocity, and feature variety of online social media, these sites often supply only a sampled set of streaming data, rather than the full data set to reduce the resource cost of computations, storage, and network bandwidth. In light of the substantial impact of sampling in Twitter data stream, this article explores a combination of spectral clustering, locality-sensitive hashing (LSH), latent Dirichlet allocation (LDA) topic modeling, and differential equation modeling to mitigate the impact of sampling on social media data analysis, in particular on detecting real-world events and predicting information diffusion. Our extensive experiments demonstrate that our proposed method is able to detect effectively the real-time emerging events and predict accurately the cascading pattern of these events from the 1% sampled Twitter data stream. To the best of our knowledge, this article is the first effort to introduce a systematic methodology to study and mitigate the impact of data sampling on social media analysis and mining.**

***Index Terms*— Big data, data sampling, social media analysis.**

I. Introduction

THE last decade has witnessed the explosive growth and disruptive utilities of online social media such as Twitter for information dissemination and content distribution. A rich literature has explored the benefit of real-time data streams from social media to detect emergency events, natural disasters, and trending topics [1]–[4]. Many popular social media sites including Twitter provide streaming data services by standard APIs to the broad researcher and developer communities. Given the sheer data volume, rapid velocity, and feature variety of online social media, these sites often supply only a sampled set of data streams, rather than the full data sets to reduce the resource cost of computations, storage, and network bandwidth [5], [6]. However, how much we can trust the observations and analysis from the sampled data sets remains a critical and challenging problem.

Our study on real-time Twitter sample stream reveals the dramatic reduction in the data volume due to the 1% sampling process, thus creating challenges and obstacles for effectively detecting the breaking events and accurately predicting the cascading process of information diffusion. To mitigate the impact of data reduction caused by the sampling process, this article proposes a systematical framework to combine spectral clustering and locality-sensitive hashing (LSH) to group effectively the related tweets triggered by the same real-world events into coherent tweet clusters.

To understand the topics and themes of tweet clusters, we adopt the widely used latent Dirichlet allocation (LDA) [7] topic modeling to discover a mixture of latent topics for the clusters. As each latent topic is expressed as a probability distribution over words observed in the tweet corpus, we further identify the most relevant words for each of the latent topics for the clusters.

To demonstrate the benefits of our proposed clustering algorithm in mitigating the impact of data sampling, we leverage tweet clusters for detecting the events from Twitter sample streams over a four-month time span. Our experimental results show that our proposed methodology is able to capture successfully all eight earthquakes that happened in California during our data-collection period between March 2018 and June 2018 and were reported by the Earthquake Hazards Program of the United States Geological Survey. An in-depth analysis shows that our proposed algorithm effectively clusters a set of sampled tweets related to the same earthquake and detect the events by LDA topic modeling and emerging word identification.

In addition, we combine genetic programming and the least square method to build the ordinary differential equation (ODE) models for describing and predicting the dynamic trend of the detected events. Our experimental results show that our ODE model is able to characterize and predict accurately the process of information diffusion for real-time events from sampled data streams. For example, the scale-free normalized mean-square error (NMSE) values on clustered tweets for two randomly selected events are 0.25 and 0.623, while the values on the most popular tweets

Manuscript received August 13, 2019; revised December 30, 2019; accepted January 23, 2020. Date of publication February 18, 2020; date of current version April 3, 2020. This work was supported in part by the National Science Foundation under Grant ATD-1737861, in part by the Humanities and Social Sciences Research, Ministry of Education of China, under Grant 18YJCZH184, and in part by the Tianjin Natural Science Foundation under Grant 19JCQNJC14800. *(Corresponding author: Kuai Xu.)*

Kuai Xu, Feng Wang, and Haiyan Wang are with the School of Mathematical and Natural Sciences, Arizona State University, Glendale, AZ 85306 USA (e-mail: kuai.xu@asu.edu).

Yufang Wang is with the Tianjin University of Finance and Economics, Tianjin 300221, China.

Ying Zhang is with the University of Technology Sydney, Ultimo, NSW 2007, Australia.

Digital Object Identifier 10.1109/TCSS.2020.2970602

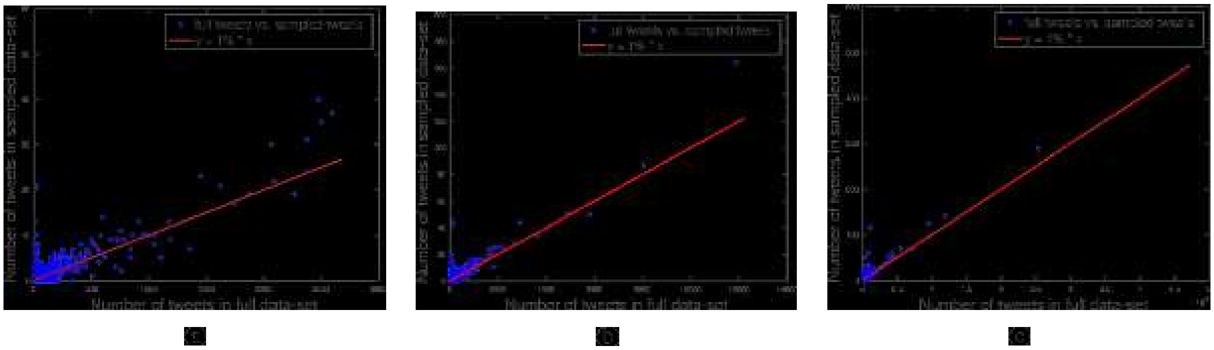


Fig. 1. Sampling ratio converges to approximately 1% as the aggregation time window increases. (a) 1-min time window. (b) 5-min time window. (c) 1-h time window.

are 0.749 and 28.601, respectively. Thus, predicting the diffusion of clustered tweets has much better accuracy than predicting the cascading patterns of the most popular tweets thanks to the data and content aggregations by the tweet clusters.

The contributions of this article are threefold, which are as follows.

1) This article systematically studies the impact of sampling in social media data streams and introduces spectral clustering algorithms for discovering the tweet clusters from the sampled data streams.
2) This article explores LSH for effectively constructing a similarity matrix for clustering analysis, thus significantly reducing the overall running time of tweet clustering.
3) This article demonstrates that our proposed system is able to detect effectively all the real-time events from the 1% sampled Twitter data streams and to predict accurately the process of information diffusion over online social media with simple yet effective data-driven ODE models.

The remainder of this article is organized as follows. Section II discusses the rationale and impact of data sampling in social media analysis and outlines our proposed systematic framework for mitigating such an impact. Section III presents our proposed clustering algorithms to group the related tweets into distinctive tweet clusters from the sampled data streams, while Section IV introduces LSH for effectively constructing the similarity matrix for the clustering algorithm. Section V demonstrates the benefits of tweet clustering for effective event detection from the Twitter sample streams, while Section VI explores the genetic programming and ODE models for predicting the process of information diffusion for the tweet clusters. Section VII discusses the related work on data sampling, event detections in social media, and information diffusion over online social media. Finally, Section VIII concludes this article and outlines our future work.

II. Real-Time Sampled Data Streams

In this section, we first describe the rationale of sampling in social media data sharing, and subsequently shed light on the impact of sampling on social media analysis using real-time event detection and information diffusion as case studies. We conclude this section with an overview of our proposed framework for mitigating the impact of data sampling on social media analysis and mining.

A. Benefits of Sampling

Considering the sheer data volume and velocity, sharing the full data set is often expensive in terms of storage and network bandwidth; thus, sampling becomes a popular choice for online social media to share data to the researcher and developer community. For example, the sample Tweets API available at the Twitter developer platform returns random samples of all tweets in real time. A few research studies [8]–[11] have confirmed that the sampled data stream is approximately 1% of the full data set and, more importantly, have pointed out the challenges and opportunities of analyzing the sampled data sets.

Since early 2018, we have been continuously collecting real-time Twitter data streams by the sample Tweets APIs as well as the complete set of tweets for a number of selected topics by the filter Tweets APIs, which track a predefined list of keywords associated with unpredictable natural disasters and extreme events such as earthquake, typhoons, floods, epidemics, and infectious diseases. For the simplicity of presentation, we refer to the random sampled tweets as the sampled data set, while referring to the complete set of tweets for the selected topics as the full data set.

As shown in Fig. 1(a)–(c), the number of the sampled tweets containing earthquake is approximately 1% of the full set of tweets containing the same keyword. In addition, as we increase the time window for data aggregation from 1 to 5 min or even to 1 h, the observations on the actual sampling ratio of approximately 1% become much clearer due to the law of large numbers.

B. Impact of Sampling

1) Impact of Sampling on Detecting Real-Time Events: The sampling approach for social media data sharing is very effective for reducing the data size of collection and computations; however, data reduction due to the sampling process creates challenges for accurately detecting the emerging events embedded in the Twitter data streams. As Twitter is a major

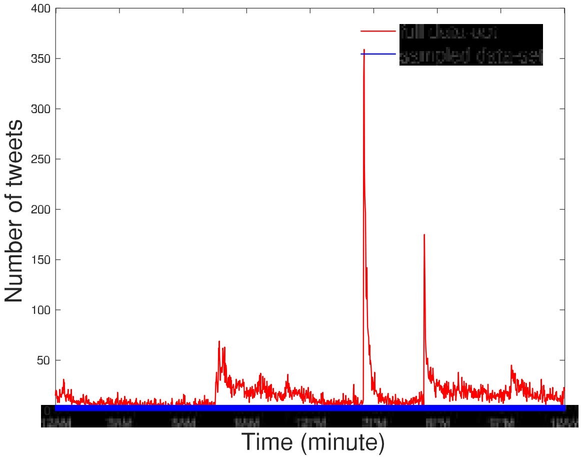


Fig. 2. Tweet time-series from full and sampled data sets on Monday, April 9, 2018 when 27 earthquakes were reported worldwide.

platform for Internet users to report natural disasters or the latest news, it is not surprising to observe a number of the spikes on the tweets related to earthquakes on April 9, 2018 when several earthquakes were reported worldwide [12], from the full data set tracking the earthquake keyword, as evidenced in the top red line of Fig. 2. However, such spikes are insignificant and infrequent in the sampled data set, i.e., the bottom blue line in Fig. 2 due to the sampling process. The early detection on such earthquake events is very crucial to inform the general public on the latest status and to coordinate the disaster recovery.

2) *Impact of Sampling on Characterizing Information Diffusion:* The sampling process with 1% sampling ratio also creates difficulty in characterizing and predicting information diffusion over Twitter social media due to the loss of nearly 99% retweet messages. Fig. 3(a) and (c) shows the diffusion pattern in the first 2 h for the popular tweet messages from three local Los Angles news organizations that report earthquake news on April 5, 2018. The top line in each figure shows the count of cumulative retweet messages of each tweet from the full data set, which reflects the actual diffusion pattern in the first 2 h, while the bottom line represents the count of the cumulative retweets from the sampled data set that apparently is unable to capture the rapid diffusion of the tweets from these credible media organizations reporting the earthquake news.

C. Mitigating Sampling Impact

The sampling process undoubtedly reduces the data size for storage and computations; however, it also creates substantial challenges for early event detections for unexpected natural disasters. In light of the impact of the sampling process on real-world event detections and information diffusion characterizations, we propose a systematic framework to combine cluster analysis, LSH, and LDA topic modeling to mitigate such an impact.

As illustrated in Fig. 4, we start with data collection by real-time Twitter sample stream, and preprocess each tweet or retweet to extract words and tokens from the main text. Subsequently, we explore LSH and cluster analysis to construct efficiently the similarity matrix and group the related tweets from the Twitter sample stream into tweet clusters. For each cluster, we run LDA topic modeling on its text corpus for an in-depth understanding of tweet contents, topics, and themes in the cluster. Our experimental results show that the availability of tweet clusters and their topics allows us to detect effectively the real-world events and characterize and predict accurately the cascading process of information diffusion over online social media.

III. Discovering Tweet Clusters by Spectral Clustering Algorithm

As shown in Section II, the data reduction in the sampling process has smoothed the data volume spikes in the original full data set and has dissolved the spreading pattern of broadcasting tweets on emergency events from the influential news organizations. Thus, it is necessary to develop effective techniques to detect such events from the sampled data streams.

Cluster analysis is one of the widely used methods for grouping similar data or content into coherent clusters [13]; thus, we explore the spectral clustering technique [14]–[16] in this article due to its implementation simplicity and computation efficiency to identify tweets that are scattered due to the sampling process but are related to the same events. These discovered tweet clusters help rapidly and effectively detect real-world events in real time.

Algorithm 1 summarizes the major steps of our algorithm for clustering and characterizing the group of tweets that reflect similar events. After standard data preprocessing on each tweet text such as tokenization, stemming, and lemmatization, the prerequisite step of applying the spectral clustering algorithm for discovering tweet clusters is to find the similarity matrix for all the real-time tweets collected during a given time window from τ_0 to $\tau_0 + \delta$. For each pair of tweets, we adopt Jaccard similarity, a popular document similarity measure, reflecting the proportion of the number of shared common words to the total number of unique words in these two tweets, also referred to as documents in the literature to calculate the content similarity between two original tweets created between $\tau - \delta$ and $\tau + \delta$. For example, given two tweets t_1 and t_2, their Jaccard similarity $j(t_1, t_2)$ is calculated as

$$J(t_1, t_2) = \frac{d_1 \cap d_2}{d_1 \cup d_2} \tag{1}$$

where d_1 and d_2 represent the set of unique words in t_1 and t_2, respectively. The above Jaccard similarity captures the similarity of contents in different tweets that are posted in the same time window; thus, the Jaccard similarity indirectly captures the temporal similarity of the tweets as well. We have experimented additional features of tweeting and retweeting behaviors, such as retweet count and comment count, which exhibit little improvement on top of content-based and temporal-based similarity. We conjecture that the tweet similarity is largely driven by when and what are posted, rather than who post the actual tweets.

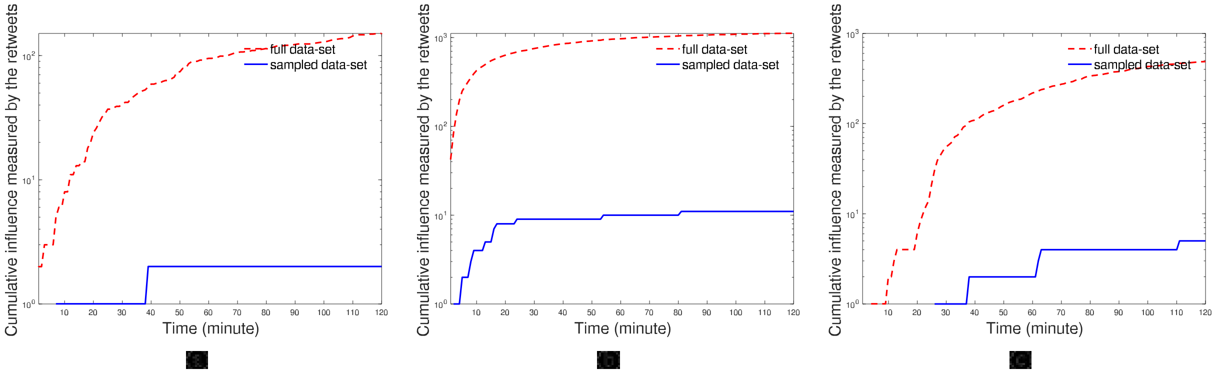


Fig. 3. Impact of sampling on the diffusion pattern and the influence intensity of the popular tweets reporting earthquake news by three different news organizations. (a) Tweet 1. (b) Tweet 2. (c) Tweet 3.

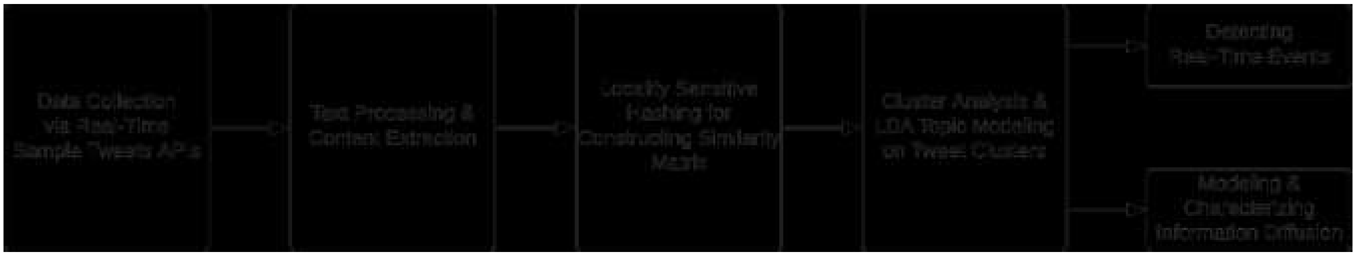

Data Collection via Real-Time Sample Tweets APIs
Text Processing & Context Extraction
Locality Sensitive Hashing for Constructing Similarity Matrix
Cluster Analysis & LDA Topic Modeling on Tweet Clusters
Detecting Real-Time Events
Modeling & Characterizing Information Diffusion

Fig. 4. Schematic of our proposed methodology for mitigating the impact of data sampling on social media analysis and mining.

The availability of a similarity matrix $S \in \mathbb{R}^{n\times n}$ between the tweets leads us to the next step of constructing the Laplacian matrix L as $L = A^{-1/2}SA^{-1/2}$, where A is a diagonal matrix and $A(i,i) = \sum_{j=1}^{n} s_{i,j}$, where $1 \leq i \leq n$. To identify the optimal number of clusters, we employ the elbow principle of searching for the largest k eigenvalues [17]. Subsequently, we identify the top k eigenvalues and their corresponding tweet clusters.

The last step of the clustering algorithm is to perform LDA [7] topic modeling to unveil the latent topics of each tweet cluster. The basic idea of applying LDA on the text corpus in each tweet cluster is to represent these tweets as a random mixture of latent topics. Each latent topic is expressed as a probability distribution over the words observed in the tweet corpus. The clustering steps presented in Algorithm 1 essentially explore a two-step clustering approach for discovering distinct tweet clusters from real-time Twitter sample streams. The first clustering step relies on the content similarity to group tweets that share similar contents including event, time, location, and people, while the second clustering step uses LDA topic modeling to capture tweets with similar latent topics due to the same underlying events.

To quantify the topic quality, we rely on The topic coherence measure score [18], which is calculated as

$$\text{coherence}(W_t) = \sum_{\{w_i,w_j\}\in W_t} \text{score}(w_i,w_j) \tag{2}$$

where W_t is the set of words for a given latent topic t, and w_i and w_j are the two words in W_t. The pairwise coherence score score(w_i,w_j) is derived as

$$\text{score}(w_i,w_j) = \log\frac{D(w_i,w_j)+E}{D(w_j)} \tag{3}$$

where $D(w_i,w_j)$ is the number of tweets containing both words w_i and w_j and $D(w_j)$ is the number of tweets containing w_j. E is set to 1 to address the scenario of $D(w_i,w_j) = 0$.

In addition, the word with the highest probability or weight in the latent topic is considered as the most representative token for the topic. Thus, the step of LDA topic modeling generates the latent topics represented with a bag of weighted words for each tweet cluster. In our experiments, we choose the number of latent topic as 3 for each cluster, since our empirical analysis shows that three latent topics are often sufficient to characterize the actual topics and themes from the overall content of the tweet clusters.

IV. Efficient Similarity Matrix Construction by LSH

The similarity matrix construction in the aforementioned spectral clustering algorithm has a running time of $O(n^2)$ due to the pairwise similarity measure calculation, where n is the number of the tweets or retweets captured during a given time window. Thus, optimizing the construction of similarity matrix could essentially improve the overall running time of the clustering algorithm, a very crucial system aspect for real-time event detections, for a large value of n even in the Twitter sample streams.

Given the diversity of tweets reflecting different events over the world, many tweets actually share little or no content or event similarity. Thus, a natural optimization strategy would be focusing on the similarity calculation on the tweets that are likely to be related to the same events. In this article, we explore the benefit of LSH, a widely used algorithm for near-duplicate detection and near-neighbor search [19], to calculate the similarity measures among tweets

Algorithm 1 Algorithm of Discovering Tweet Clusters From Real-Time Twitter Sample Streams

```
Input: a set of sampled tweets, denoted as $T$ with a size of $n$, within a given time window from $\tau_0$ to $\tau_0 + \delta$.
Calculate the Jaccard similarity, $J_{i,j}$, for each pair of tweets, $t_i$ and $t_j$ based on their content, and obtain the similarity matrix $S_T \in \mathbb{R}^{n \times n}$ for all tweets in $T$;
Build a diagonal matrix $A$ with $A(i, i) = \sum_{j=1}^{n} s_{i,j}$, where $1 \leq i \leq n$, and construct the Laplacian matrix $L$ as $L = A^{-1/2} SA^{-1/2}$;
Search the largest $k$ eigenvalues, $\lambda_1, \lambda_2, \cdots, \lambda_k$ such that $\sum_{i=1}^{k} \lambda_i \geq \alpha \times \sum_{j=1}^{n} \lambda_n$ and $(\lambda_k - \lambda_{k+1}) \geq \beta \times (\lambda_{k-1} - \lambda_k)$;
Construct the matrix $E = [e_1\ e_2 \cdots e_k] \in \mathbb{R}^{n \times k}$ with the corresponding $k$ eigenvectors $(ev_1, ev_2, \dots, ev_k)$ associated with the above $k$ eigenvalues, and normalize $E$ to derive the matrix $Z$ such that each row in the normalized matrix $Z$ has a unit length, and consider each row as a point;
Perform $k$-means cluster analysis on $Z$ to identify $k$ clusters $(D_1, D_2, \cdots, D_k)$;
Assign the tweet $t_i$ to the cluster $C_j$ if the row $i$ of $Z$ is assigned to the cluster $D_j$;
Run LDA topic modeling to discover latent topics $L_i$ for all the tweets in each cluster $C_i$ in $C$.
Output: tweet clusters $C_1, C_2, \cdots, C_k$, where $C_i = \{p_j | z_j \in Y_j\}$, and the latent topics $L_i$ for each cluster.
```

whenever necessary, rather than calculating all pairs of similarity measure in a brute-force fashion.

Minhash, one of the first and most popular LSH methods, is an LSH function originally designed for efficiently approximating the Jaccard similarity [20]. Let h denote a hash function that maps words or terms in the tweet documents to distinct integers, also referred to as hash bucket numbers. The minhash on a set X, $h_{\min}(X)$, is defined as the minimal value $h(x)$, where x is one of the elements in the set X and $h(x)$ is the smallest value among all hash values for the elements in X. A key property of minhash lies in that the probability of the minhash function on the two word sets d_1 and d_2 from two tweets t_1 and t_2, producing the same values as the Jaccard similarity of two sets, i.e., Probability$[h_{\min}(d_1) = h_{\min}(d_2)] = J(t_1, t_2)$.

The proof of this property is straightforward. Obtaining the same value for $h_{\min}(d_1)$ and $h_{\min}(d_2)$ indicates that the element with the smallest value of applying the hash function h on $d_1 \cup d_2$ is also in $d_1 \cap d_2$. In other words, the probability of $h_{\min}(d_1) = h_{\min}(d_1)$ essentially becomes $d_1 \cap d_2 / d_1 \cup d_2$. As shown in (1), $J(t_1, t_2) = d_1 \cap d_2 / d_1 \cup d_2$. Thus, Probability$[h_{\min}(d_1) = h_{\min}(d_2)] = J(t_1, t_2)$ holds.

Minhashing maps any tweet to a signature consisting of a set of integers and effectively preserves the content similarity for any two tweets. However, comparing minhashing for all possible pairs of tweets is fairly expensive due to the sheer size of the tweet streams. Thus, our next step is to explore minhash-based LSHs to calculate the similarity between the tweets in T by the hash collision probability distribution over a set of $H = \{h_1, h_2, \dots, h_n\}$ hash functions. The intuition of LSH on reducing similarity calculation is to hash tweets several times such that two tweets with shared content have higher probability to experience hash collision, i.e., be hashed to the same bucket, than the tweets sharing little or no content. Specifically, LSH first divides the hash values from n hash functions into b bands of r rows, and subsequently maps r values in each band with a simple hash function to a hash table. In other words, each band is divided into buckets. If two tweets exhibit the same hash values in one band, they will be mapped to the same bucket, thus becoming a candidate pair of tweets for further similarity calculations.

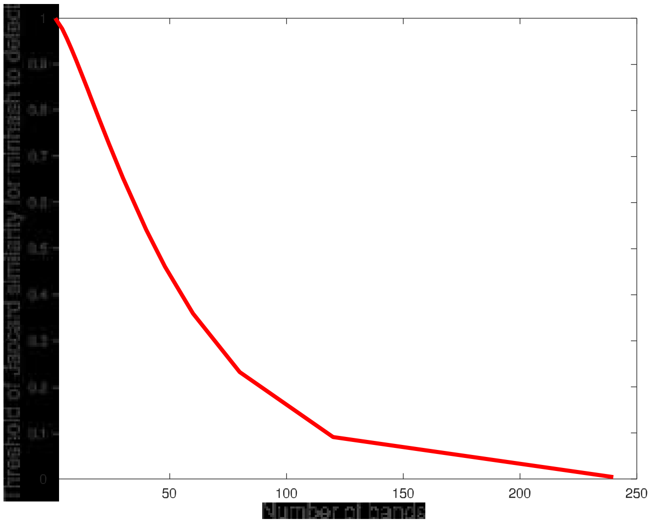


Fig. 5. Impact of the number of bands in minhash design on the threshold of Jaccard similarity, which are likely to be detected by 240 minhashes.

As proved in [21], the probability of detecting a candidate pair, p, via minhash LSH is a function of s, b, and r, i.e., $p = 1 - (1 - s^r)^b$. Thus, the threshold of Jaccard similarity between the two tweets for ensuring the 50%+ probability of becoming a candidate pair actually depends on the parameters of b and r. As shown in Fig. 5, as the number of bands in 240-minhash LSH increases, the threshold of Jaccard similarity decreases.

For a given configuration of b bands with r rows for each band, a higher Jaccard similarity of two tweets increases the probability of detecting the pair as a candidate. As illustrated in Fig. 6, the three S-curves for 60, 80, and 120 bands, respectively, capture the relationship between the Jaccard similarity of two tweets and the probability of detecting the pair as a candidate in minhash-based LSH. For the same Jaccard similarity, the detection probability is higher for 120 bands LSH, which is consistent with the observations in Fig. 5. In this article, we choose 120 for the parameter b to balance the detection coverage and computational overhead.

Our experimental results have revealed the significant benefit of applying minhash-based LSH for calculating similarity matrix S for the tweet set T. As shown in Fig. 7, this method has successfully and consistently reduced over 90% similarity calculations on real-time Twitter sample streams from different time windows when choosing 120 bands for 240-minhash-based LSH for approximately constructing the similarity matrix.

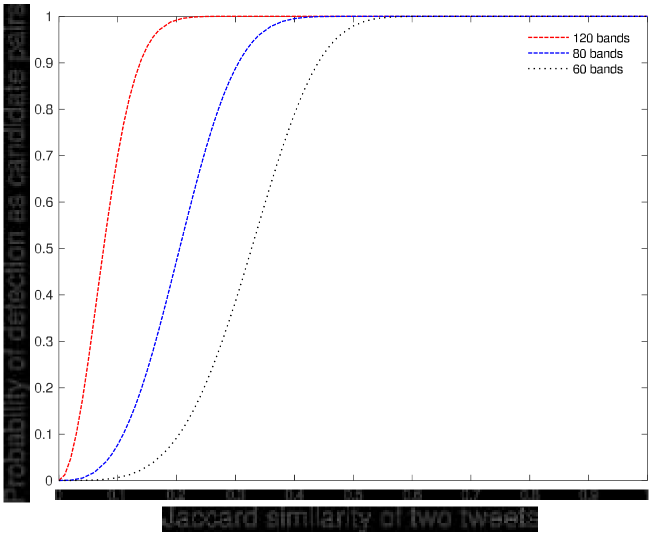


Fig. 6. Probability of detection as candidate pairs for two tweets with different Jaccard similarities.

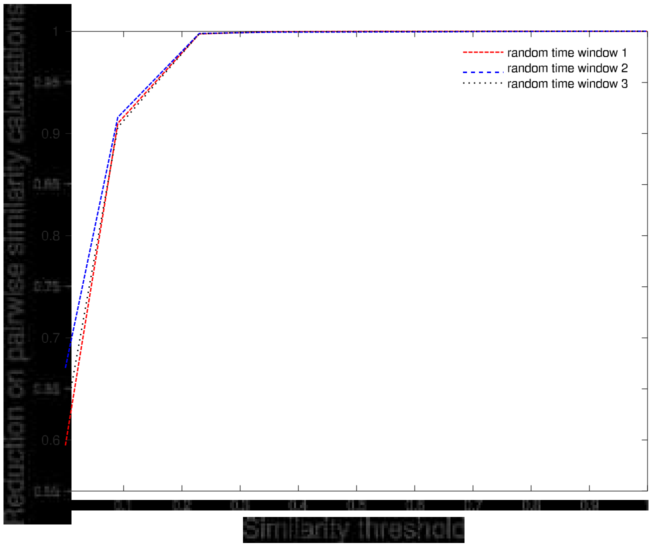


Fig. 7. Reduction of pairwise similarity calculations for varying similarity threshold.

V. Effective Event Detection by Tweet Clusters

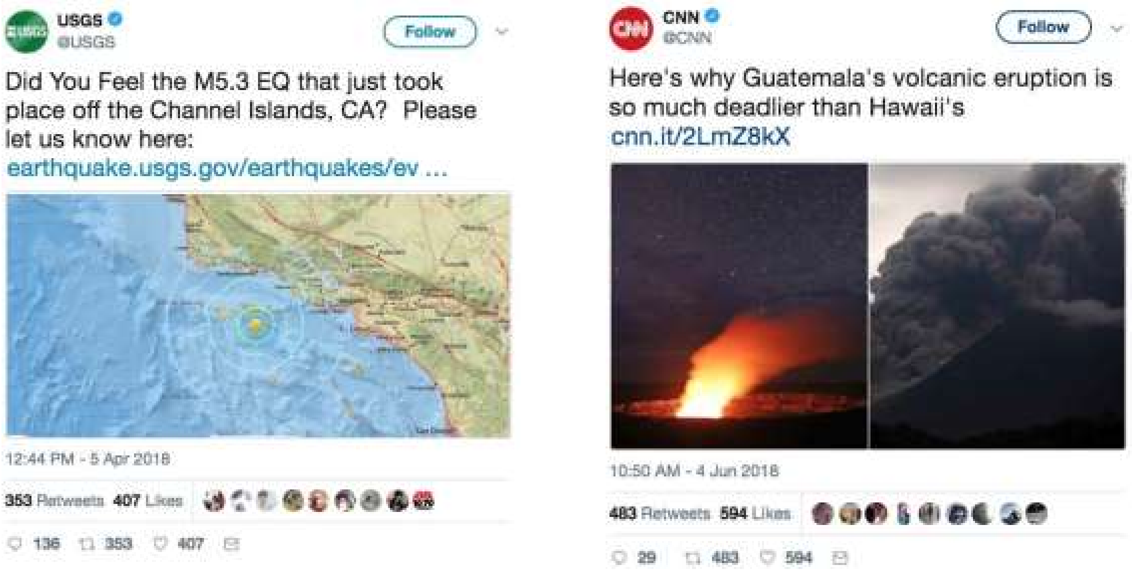
In this article, we define an event as a continuous data stream contributed by different users on the same social media platform who share, discuss, and comment on the same topic during a specific time window. For example, Fig. 8(a) shows an original tweet posted by the official Twitter account of the United States Geological Survey reporting a 5.3 earthquake on the Santa Rosa Island near Los Angles on April 5, 2018. This timely post shared a breaking news on a significant natural disaster, and was retweeted, liked, and commented by 432, 578, and 91 Twitter users, respectively, during a very short time window. Similarly, Fig. 8(b) illustrates a CNN's tweet reporting Guatemala's deadly volcanic eruption on June 4, 2018, which was retweeted, liked, and commented by 483, 594, and 29 users, respectively. These two examples reflect the disruptive utility of online social media for event broadcasting and information sharing.

Detecting the events from the full data set of Twitter streams is relatively simple due to the sudden increase in retweets, comments, and likes of the most popular tweets. However, the 1% sampling process has significantly reduced the data volumes and smoothed the changes and dynamics of the trends for almost all tweets including the most popular ones. Thus, our tweet clustering algorithm addresses the limitations of the sampling process by recreating the social media dynamics of the same underlying real-world events by grouping together all related tweets or retweets for the same event into coherent tweet clusters.

Considering the importance of detecting the events from real-time Twitter sample streams, we build a prototype system that takes the sampled data streams from each 1-min time window as the input and generate tweet clusters and their corresponding latent topics as output. To detect the emerging events from these tweet clusters, we devise a simple change-detection algorithm that constantly searches for emerging tokens, which are the top-weighted words from the latent topics for each cluster and are not observed in the previous time windows, from the latent topics for each cluster.

The earthquake event in Fig. 8(a) happened at 19:29:16 UTC on April 5, 2018. Table I shows the tokens with the highest weight for the top five cohesive clusters along with the emerging statuses between 19:27:00 UTC and 19:32:00 UTC during that day. Our simple yet effective change-detection algorithm is able to uncover the emerging earthquake token within 2 min of the actual earthquake event from a tweet cluster that carries earthquake as an emerging token and consists of only 36 sampled tweets or retweets. Discovered by the our proposed clustering algorithm, these tweet clusters group together similar tweets reporting and commenting the same real-world event, and significantly improve our ability to detect quickly these extreme events or natural disasters in the very beginning. Such early detection is very critical for the first respondents in the disaster recovery and rescue. Similarly, our proposed technique has successfully identified the event of Guatemala's deadly volcanic eruption, as shown in Fig. 8(b), within 2 min.

Table II summarizes the effective event detection of our proposed methodology for capturing all eight significant earthquakes that happened in California during our data collection period between March 2018 and July 2018 and were reported by the Earthquake Hazards Program of the United States Geological Survey [22]. As shown in Table II, our proposed algorithm effectively clusters a set of sampled tweets related to the same earthquakes and detect the events by LDA topic modeling and emerging token identification. In this article, we rely on the case studies to evaluate the effectiveness of detecting the real-time events of our proposed methodology. For each case, we manually identify the corresponding event on the official site of the Earthquake Hazards Program of the United States Geological Survey. Due to the manual validation process, we are unable to run a large-scale event detection in this article. One of our future works is to work with the Earthquake Hazards Program and other official channels to



USGS
@USGS
Follow

Did You Feel the M5.3 EQ that just took place off the Channel Islands, CA? Please let us know here: earthquake.usgs.gov/earthquakes/ev ...

12:44 PM - 5 Apr 2018

353 Retweets 407 Likes

136 353 407

[redacted]

CNN
@CNN
Follow

Here's why Guatemala's volcanic eruption is so much deadlier than Hawaii's cnn.it/2LmZ8kX

10:50 AM - 4 Jun 2018

483 Retweets 594 Likes

29 483 594

[redacted]

Fig. 8. Natural disaster events reported on Twitter in real time. (a) Tweet reporting a 5.3 earthquake on the Santa Rosa Island near Los Angles. (b) Tweet reporting on Guatemala's volcanic eruption.

TABLE I

Top Tokens With the Most Weight From the First Latent Topic for the Top Five Cohesive Tweet Clusters in Each 1-min Time Window

[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Sergio, no	Junto, yes	Trump, no	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	Liverpool, no	Sergio, no	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	Rambo5, no
[redacted]	[redacted]	[redacted]	Rambo5, no	Liverpool, no	[redacted]
Trump, no	Sergio, no	[redacted]	Trump, no	[redacted]	Liverpool, no

TABLE II

Effective Event Detection on All Eight Significant California Earthquakes Between March 2018 and July 2018

Location	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]

explore the API-based automation methods for collecting the ground-truth events for large-scale experimental evaluations.

VI. Modeling and Predicting Information Diffusion of Tweet Clusters

The process of data sampling has substantially reduced the number of the retweets, comments, or likes even for the most popular tweets that report breaking news such as earthquake events. The reduction of tweets creates challenges for accurately characterizing and predicting the cascading process of information diffusion over online social media. However, our clustering approach for aggregating similar tweets into coherent clusters for the same real-world events accumulates sufficient signals to apply mathematical models for characterizing and predicting the diffusion process of tweet clusters.

Mathematical and statistical models are widely used for various predictions. The ODE model is a vital mathematical tool arising in biology, sociology, economics, physics, and other fields. It has been extensively used for describing and predicting various time evolutions. In this article, we combine the genetic programming and least square method [23], [24] to build the ODE models to describe and predict the dynamic trend of the earthquake.

Genetic programming is an effective evolutionary method, which mimics the mechanisms of natural selection and

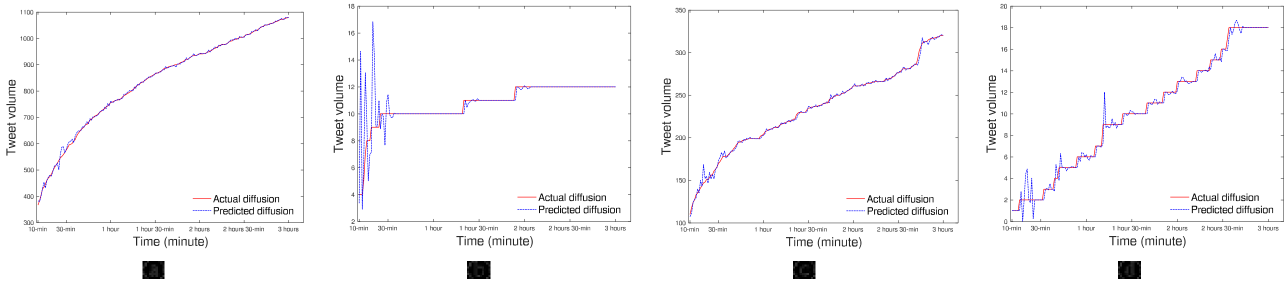


Fig. 9. Predicting diffusion patterns for tweet clusters and the most popular tweets for two earthquake events. (a) Tweet cluster, April 5 earthquake. (b) Most popular tweet, April 5 earthquake. (c) Tweet cluster, May 15 earthquake. (d) Most popular tweet, May 15 earthquake.

genetic variation. Based on suitable coding, genetic programming uses genetic operators and the principle of "survival of the fittest" to search for the optimal solutions. Specifically, we use tree-structure-based evolutionary algorithm to evolve the ODE model. The main procedure of constructing the ODE model is as follows.

1) Define the function set and the operator set.
2) Generate the initial ODE model as the first generation of the population.
3) Perform structure optimizations of the ODE models with various operations such as mutation, crossover, and selection.
4) Run parameter optimizations of the ODE models with the least square method.
5) Repeat steps 3 and 4 in each generation until a predefined number of iterations have reached or an optimal structure is found.

As a result of the procedure, an ODE model is developed in the following form:

$$\frac{dy(t)}{dt} = f(t, y(t)) \tag{4}$$

where f is a function involving multiple elementary functions, e.g., $f(t, y(t)) = \alpha t \sin(t) - \beta y(t)e^{t} + \gamma\, t$, with the structure and constants α, β, and γ of the ODE model determined by the above procedure.

Using the genetic programming and least square methods, we apply the retweets on any of the tweets in the cluster from the first 10 min since discovering the tweet cluster to train the ODE model, and then predict the growth of the retweets in the next 2 h and 50 min. Fig. 9 illustrates the prediction accuracy on the diffusion patterns for the tweet clusters and the most popular tweet for two earthquake events happening on April 5, 2018 and May 15, 2018. As shown in Fig. 9, the prediction quality on the tweet clusters [see Fig. 9(a) and (c)] is much higher than the prediction on the most popular tweets [see Fig. 9(b) and (d)] from the Twitter sample streams.

To measure quantitatively the actual prediction improvement on tweet clusters over the most popular tweets for both events, we use NMSE to calculate the scale-free difference between the actual diffusion over the predicted diffusion. The NMSE values on the tweet clusters for two events are 0.025 and 0.623, while the values on the most popular tweets are 0.749 and 28.601, respectively. Thus, the NMSE metrics also confirm that predicting the diffusion of tweet clusters has much better accuracy than predicting the cascading patterns of the most popular tweets from the Twitter sample streams. Similar to the evaluation of event detection, we also use case studies for evaluating the prediction of information diffusion for the underlying events. The automated ground-truth collection framework developed in our future work will enable us to present quantitative evaluations on a much larger scale.

VII. Related Work

Online social media has recently become a major venue for disseminating breaking news and broadcasting emergency events such as natural disasters, epidemic outbreaks, and even local traffic congestions due to its large-scale user base and rising popularity. In light of the unprecedented scale of online social media, sampling becomes an intuitive and important technique for collecting, exploring, and understanding big social media data. A number of research studies [25], [26] have pointed out the impact of sampling on social media analysis and modeling. For example, [25] shows the nontrivial impact of attribute and topology-based sampling strategies on a variety of metrics of information diffusion with crawled data sets from Twitter, while [26] systematically compares the popularity, topical diversity, trustworthiness, and timeliness of the content generated by the users who are randomly selected with the content generated by a sampled set of expert users on Twitter. Complement to these prior efforts, this article sheds light on the impact of sampling on real-time Twitter data streams on event detection and information-diffusion modeling and prediction.

Event detection is one of the important applications of exploring online social media data. A number of literature studies [27]–[30] have explored Twitter data streams to detect real-world events. For example, [28] develops a real-time system to detect automatically and geo-tag security events from social media data, and demonstrates the system with Westgate shopping mall attack in September 2013 as a case study. Similarly, [29] proposes a location–time-constrained topic model to extract content, time, and location features from the tweet messages for monitoring online social events, while [30] explores the combination of the occurrence information of social media content and the profile of social media users to detect real-time events from microblogging text streams as well as to predict the cascading popularity of the detected events. Our article is complement to these studies, since the combination of spectral clustering, LSH similarity search, and LDA topic modeling in our proposed methodology can serve as the preprocessing steps for these event-detection methods and ultimately improve their detection quality and performance.

Information diffusion is another hot topic of online social media research in the last decade due to its wide applications such as sentiment analysis and fake news detection [31]–[33]. For example, [31] characterizes the temporal and spatial-diffusion patterns of information spreading over social media and proposes a linear diffusive model based on the partial differential equation (PDE) to model and predict information diffusion over the time and underlying social network. This article will enhance the algorithms proposed in these prior studies in the context of sampled data sets. Similarly, [32] characterizes the temporal dynamics and cascading of topic-specific information such as hashtags over Twitter, and [33] proposes a new method of inferring multi-aspect diffusion networks with multi-pattern cascades for characterizing heterogeneous user interactions and diverse cascading patterns in social media. As analyzed in [25], the sampling process creates challenges in modeling and predicting information diffusion over online social media. Thus, our proposed method of spectral clustering will become a critical step for modeling information diffusion of tweet clustered discovered from real-time Twitter sample streams.

VIII. Conclusion and Future Work

The last decade has witnessed the unprecedented growth of online social media. For example, Twitter has become a major channel for reporting breaking news and sharing the latest updates of social events. However, the sampling process of Twitter real-time data streams has created substantial challenges for making sense of social media data such as detecting real-time events and predicting the cascading process of information diffusion. This article proposes a systematic methodology to combine clustering algorithms, LSH, and topic modeling to mitigate effectively the impact of data sampling in social media analysis and mining with earthquake events as case studies. Our extensive experimental results have shown that our proposed system is able to effectively detect all significant earthquake events happening in California between March 2018 and June 2018 from the 1% sampled Twitter data stream, and our system accurately predicts the cascading pattern of information diffusion for earthquake events with sampled data streams. Our future work is centered on understanding and mitigating the impact of social media data sampling on sentiment analysis and content modeling with word embedding [34] and natural language processing (NLP) techniques. In addition, we are planning to integrate the prototype system with the Apache Spark real-time streaming analytics engine [35] to reduce the latency of event detections to less than 1 min.

References

[1] M. Dayarathna and S. Perera, "Recent advancements in event processing," *ACM Comput. Surv.*, vol. 51, no. 2, Jun. 2018.

[2] B. Poblete, J. Guzman, J. Maldonado, and F. Tobar, "Robust detection of extreme events using Twitter: Worldwide earthquake monitoring," *IEEE Trans. Multimedia*, vol. 20, no. 10, pp. 2551–2561, Oct. 2018.

[3] P. Wagenseller, F. Wang, and W. Wu, "Size matters: A comparative analysis of community detection algorithms," *IEEE Trans. Comput. Social Syst.*, vol. 5, no. 4, pp. 951–960, Dec. 2018.

[4] R. Dong, L. Li, Q. Zhang, and G. Cai, "Information diffusion on social media during natural disasters," *IEEE Trans. Comput. Soc. Syst.*, vol. 5, no. 1, pp. 265–276, Mar. 2018.

[5] S. Kulkarni *et al.*, "Twitter heron: Stream processing at scale," in *Proc. ACM SIGMOD Int. Conf. Manage. Data*, Jun. 2015, pp. 239–250.

[6] A. Toshniwal *et al.*, "Storm@ Twitter," in *Proc. ACM SIGMOD Int. Conf. Manage. Data*, Jun. 2014, pp. 147–156.

[7] D. Blei, A. Ng, and M. Jordan, "Latent Dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, Jan. 2003.

[8] K. Xu, F. Wang, X. Jia, and H. Wang, "The impact of sampling on big data analysis of social media: A case study on flu and Ebola," in *Proc. IEEE GLOBECOM*, Dec. 2015, pp. 1–6.

[9] Y. Wang, J. Callan, and B. Zheng, "Should we use the sample? Analyzing datasets sampled from Twitter's stream API," *ACM Trans. Web*, vol. 9, no. 3, pp. 1–23, Jun. 2015.

[10] F. Morstatter, J. Pfeffer, H. Liu, and K. Carley, "Is the sample good enough? Comparing data from Twitter's streaming api with Twitter's firehose," in *Proc. Int. Conf. Weblogs Social Media (ICWSM)*, Jul. 2013.

[11] D. Palguna, V. Joshi, V. Chakaravarthy, R. Kothari, and L. V. Subramaniam, "Analysis of sampling algorithms for Twitter," in *Proc. Int. Conf. Artif. Intell. (IJCAI)*, Jul. 2015, pp. 967–973.

[12] Global Earthquake Monitor. *Map & List of Recent Earthquakes Worldwide*. Accessed: Aug. 1, 2019. [Online]. Available: https://www.volcanodiscovery.com/earthquakes/archive/2018-apr.html

[13] Z. Miller, B. Dickinson, W. Deitrick, W. Hu, and A. H. Wang, "Twitter spammer detection using data stream clustering," *Inf. Sci.*, vol. 260, pp. 64–73, Mar. 2014.

[14] A. Ng, M. Jordan, and Y. Weiss, "On spectral clustering: Analysis and an algorithm," in *Proc. Neural Inf. Process. Syst. (NIPS) Conf.*, 2001, pp. 849–856.

[15] Y. Weiss, "Segmentation using eigenvectors: A unifying view," in *Proc. Int. Conf. Comput. Vis.*, 1999, pp. 975–982.

[16] M. Maila and J. Shi, "Learning segmentation with random walk," in *Proc. Neural Inf. Process. Syst. Conf.*, 2001, pp. 873–879.

[17] W. Krzanowski and Y. Lai, "A criterion for determining the number of groups in a data set using sum-of-squares clustering," *Biometrics*, vol. 44, no. 1, pp. 23–34, Mar. 1988.

[18] K. Stevens, P. Kegelmeyer, D. Andrzejewski, and D. Buttler, "Exploring topic coherence over many models and many topics," in *Proc. Joint Conf. Empirical Methods Natural Lang. Process. Comput. Natural Lang. Learn.*, Jul. 2012, pp. 952–961.

[19] P. Indyk and R. Motwani, "Approximate nearest neighbors: Towards removing the curse of dimensionality," in *Proc. ACM Symp. Theory Comput.*, May 1998, pp. 604–613.

[20] A. Broder, "On the resemblance and containment of documents," in *Proc. Compress. Complex. Sequences (SEQUENCES)*, Jun. 1997, pp. 20–29.

[21] J. Leskovec, A. Rajaraman, and J. Ullman, *Mining of Massive Datasets*, 2nd ed. Cambridge, U.K.: Cambridge Univ. Press, 2014.

[22] United States Geological Survey. *Earthquake Hazards Program*. Accessed: Aug. 1, 2019. [Online]. Available: https://earthquake.usgs.gov/earthquakes/browse/significant.php

[23] J. Madár, J. Abonyi, and F. Szeifert, "Genetic programming for the identification of nonlinear input—Output models," *Ind. Eng. Chem. Res.*, vol. 44, no. 9, pp. 3178–3186, 2005.

[24] H. Cao, L. Kang, Y. Chen, and J. Yu, "Evolutionary modeling of systems of ordinary differential equations with genetic programming," *Genetic Program. Evolvable Mach.*, vol. 1, no. 4, 2000.

[25] M. Choudhury, Y. Lin, H. Sundaram, K. Candan, L. Xie, and A. Kelliher, "How does the data sampling strategy impact the discovery of information diffusion in social media?" in *Proc. 4th Int. AAAI Conf. Weblogs Social Media (ICWSM)*, May 2010.

[26] M. Zafar, P. Bhattacharya, N. Ganguly, K. Gummadi, and S. Ghosh, "Sampling content from online social networks: Comparing random vs. expert sampling of the Twitter stream," *ACM Trans. Web*, vol. 9, no. 3, pp. 1–33, Jun. 2015.

[27] M. Hasan, M. Orgun, and R. Schwitter, "A survey on real-time event detection from the Twitter data stream," *J. Inf. Sci.*, vol. 44, no. 4, pp. 443–463, Mar. 2017.

[28] M. Osborne *et al.*, “Real-time detection, tracking, and monitoring of automatically discovered events in social media,” in *Proc. Annu. Meeting Assoc. Comput. Linguistics, System Demonstrations*, Jun. 2014, pp. 37–42.

[29] X. Zhou and L. Chen, “Event detection over Twitter social media streams,” *Int. J. Very Large Data Bases*, vol. 23, no. 3, pp. 381–400, Jun. 2014.

[30] X. Zhang, X. Chen, Y. Chen, S. Wang, Z. Li, and J. Xia, “Event detection and popularity prediction in microblogging,” *Neurocomputing*, vol. 149, pp. 1469–1480, Feb. 2015.

[31] F. Wang, H. Wang, K. Xu, J. Wu, and X. Jia, “Characterizing information diffusion in online social networks with linear diffusive model,” in *Proc. IEEE Int. Conf. Distrib. Comput. Syst. (ICDCS)*, Philadelphia, PA, USA, Jul. 2013, pp. 307–316.

[32] E. Stai, E. Milaiou, V. Karyotis, and S. Papavassiliou, “Temporal dynamics of information diffusion in Twitter: Modeling and experimentation,” *IEEE Trans. Comput. Social Syst.*, vol. 5, no. 1, pp. 256–264, Mar. 2018.

[33] S. Wang, X. Hu, P. Yu, and Z. Li, “MMRate: Inferring multi-aspect diffusion networks with multi-pattern cascades,” in *Proc. ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2014, pp. 1246–1255.

[34] T. Mikolov, K. Chen, S. Greg Corrado, and J. Dean, “Efficient estimation of word representations in vector space,” in *Proc. Int. Conf. Learn. Represent.*, 2013, pp. 1–12.

[35] M. Zaharia *et al.*, “Apache spark: A unified engine for big data processing,” *Commun. ACM*, vol. 59, no. 11, pp. 56–65, Nov. 2016.

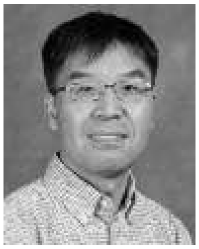
Kuai Xu (Senior Member, IEEE) received the B.S. and M.S. degrees in computer science from Peking University, Beijing, China, in 1998 and 2001, respectively, and the Ph.D. degree in computer science from the University of Minnesota, Minneapolis, MN, USA, in 2006.

He is currently an Associate Professor with Arizona State University, Glendale, AZ, USA. His research interests include network security, internet measurement, big data, data mining, and machine learning.

Dr. Xu is a member of the Association for Computing Machinery (ACM).

Feng Wang (Member, IEEE) received the B.S. degree from Wuhan University, Wuhan, China, in 1996, the M.S. degree from Peking University, Beijing, China, in 1999, and the Ph.D. degree from the University of Minnesota, Minneapolis, MN, USA, in 2005, all in computer science.

She is currently a Professor with the School of Mathematical and Natural Sciences, Arizona State University, Glendale, AZ, USA. Her research interests focus on network science, social media analysis, network optimization, network security, and wireless sensor networks.



Haiyan Wang received the bachelor’s degree in mathematics from Northwest Normal University, Lanzhou, China, in 1985, the master’s degree in applied mathematics from the Ocean University of China, Qingdao, China, in 1988, and the master’s degree in computer science and the Ph.D. degree in mathematics from Michigan State University, East Lansing, MI, USA, in 1997.

After several years spent working as a Software Engineer in industry, he came to Arizona State University, Glendale, AZ, USA, as an Assistant Professor, in 2005, where he is currently a Professor. He was promoted to an Associate Professor and a Professor in 2011 and 2015, respectively. His research interests include applied mathematics, differential equations, mathematical biology, and online social networks, and big data.

Yufang Wang received the Ph.D. degree in computational mathematics from Zhejiang University, Hangzhou, China, in 2016.

She is currently a Lecturer with the School of Statistics, Tianjin University of Finance and Economics, Tianjin, China. Her research interests include applied mathematics and data science.

Ying Zhang received the bachelor’s and master’s degrees in computer science from Peking University, Beijing, China, in 1998 and 2001, respectively, and the Ph.D. degree from the School of Computer Science and Engineering, University of New South Wales (UNSW), Sydney, NSW, Australia, in 2008.

He is currently a Senior Lecturer with the Centre of Quantum Computation and Intelligent Systems, University of Technology Sydney, Ultimo, NSW, Australia. His research interests include efficient query processing on spatial data, stream data, and graphs.