Information Processing and Management 56 (2019) 1204–1219

Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/infoproman

Whose posts to read: Finding social sensors for effective information acquisition

Kun Yuan[a], Guannan Liu[a,*], Junjie Wu[a,b,c]

[a] School of Economics and Management, Beihang University, Beijing 100191, China
[b] Beijing Advanced Innovation Center for Big Data and Brain Computing, Beihang University, Beijing 100191, China
[c] Beijing Key Laboratory of Emergency Support Simulation Technologies for City Operations, Beihang University, Beijing 100191, China

ARTICLE INFO

Keywords:
Social sensing maximization
Social media
LeCELF
Participation paradox
Information acquisition

ABSTRACT

In the era of big data, it is extremely challenging to decide what information to receive and filter out in order to effectively acquire high-quality information, particularly in social media where large-scale User Generated Contents (UGC) is widely and quickly disseminated. Considering that each individual user in social network can take actions to drive the process of information diffusion, it is naturally appealing to aggregate spreading information effectively at the individual level by regarding each user as a social sensor. Along this line, in this paper, we propose a framework for effective information acquisition in social media. To be more specific, we introduce a novel measurement, the *preference-based Detection Ability* to evaluate the ability of social sensors to detect diffusing events, and the problem of effective information acquisition is then reduced to achieving *social sensing maximization* through discovering valid social sensors. In pursuit of social sensing maximization, we propose two algorithms to resolve the longstanding problems in traditional greedy methods from the perspectives of efficiency and performance. On the one hand, we propose an efficient algorithm termed LeCELF, which resolves the redundant re-evaluations in the traditional Cost-Effective Lazy Forward (CELF) algorithm. On the other hand, we observe the participation paradox phenomenon in the social sensing network, and proceed to propose a randomized selection-based algorithm called FRIENDOM to choose social sensors to improve the effectiveness of information acquisition. Experiments on a disease spreading network and real-world microblog datasets have validated that LeCELF greatly reduces the running time, whereas FRIENDOM achieves a better detection performance. The proposed framework and corresponding algorithms can be applicable in many other settings in resolving information overload problems.

1. Introduction

Information overload, a critical concern in information acquisition, has become a pervasive issue in the era of big data, particularly in the social media environment, where large amount of User Generated Contents (UGC) are spreading every second. Therefore, users have to make decisions on what information to receive and filter out in order to keep up with the latest news and trendy topics. The abundant UGC and the retweeting mechanism have enabled information to spread at an extremely fast pace through microblogs, making the platform become an information aggregation channel where events may break out as a headline under a spotlight (Kwak, Lee, Park, & Moon, 2010). For example, the death of Osama bin Laden in 2011 was speculated about and

* Corresponding author.
E-mail address: liugn@buaa.edu.cn (G. Liu).

https://doi.org/10.1016/j.ipm.2019.01.009
Received 15 March 2018; Received in revised form 25 October 2018; Accepted 21 January 2019
Available online 09 March 2019
0306-4573/

discussed first on Twitter, even earlier than the official announcement by the White House and reports from the traditional news media. Meanwhile, earthquake (Sakaki, Okazaki, & Matsuo, 2010) information may also break out on social media before reports from official organizations. Thus, the public and professional organizations are always keeping close eyes on text streams in microblog platforms in order to acquire high-quality information in time.

One possible solution for effective information acquisition via social media is to keep watching some continuously updating statistics from aggregated text streams (Laylavi, Rajabifard, & Kalantari, 2017; Schubert, Weiler, & Kriegel, 2014; Tu & Seng, 2012), which, however, is often overwhelmed by large volumes of noisy UGC and severe uncertainty of information diffusion paths. Moreover, the bursts can only be detected when the related texts are accumulated to a certain level, making real-time detection practically infeasible. To address these issues, some studies have resorted to detecting bursts in a specific domain, such as the one for real-time earthquake detection from Twitter (Sakaki et al., 2010), which keeps tracking some task-specific keywords and, meanwhile, filters out vast amounts of irrelevant information. This keyword-sensing solution, however, suffers from high computational intensity due to frequent keyword searches and relies heavily on the scope and accuracy of the keywords.

In this paper, we adopt a user-centric view regarding effective information acquisition from social media. That is, we regard each individual user as a "social sensor" for information threads of interest. The task of acquiring effective information of a specific type in a timely manner therefore reduces to finding a proper set of users that could *sense, i.e.*, review or propagate, the targeted diffusing events. This social-sensing solution has some obvious merits. First, by watching only the group of social sensors online, we can filter out massive amounts of irrelevant information that might "disguise" truly valuable information and thus reduce the computational intensity greatly. Additionally, since the social sensors are deemed to sense the diffusing events in a more efficient manner, high-quality information can be acquired in real time. On the other hand, to set a social sensor could also incur a cost for extra information processing and disambiguation. Therefore, the prerequisite for effective information acquisition in social media has become searching for a set of users who can provide maximum sensing ability to targeted events within budget, which is defined as the *Social Sensing Maximization* (SSM) problem in this paper.

To resolve the SSM problem, we should first consider the measurements for the sensing ability of each user. Naturally, the number of events that can be detected and the time it takes to detect these events are the two crucial aspects in evaluating the sensing ability. However, previous studies usually set independent objectives, such as the detection likelihood (DL), detection time (DT), and population affected (PA) (Mirzasoleiman, Badanidiyuru, Karbasi, Vondrák, & Krause, 2015), to choose sensors without a comprehensive view. Following the basic assumption of "*better late than never*", the sensors should at least detect the diffusion events, and then, we can further take the detection time into consideration. Therefore, in this paper, we propose a novel measurement, called *preference-based Detection Ability* (*pDA*), that combines both aspects, and the aggregate sensing ability of the set of social sensors is regarded as the objective function for the SSM problem. Specifically, the proposed *pDA* sets special priorities for choosing social sensors, who can detect as many events as possible, whereas the corresponding detection time is treated as a reward for further evaluating the effectiveness of each sensor.

With respect to the proposed objective function *pDA*, the SSM problem can be reduced to a constrained search problem analogous to the budgeted max-cover problem (Khuller, Moss, & Naor, 1999), knapsack problem (Kellerer, Pferschy, & Pisinger, 2004), and influence maximization problem (Chen, Wang, & Yang, 2009; Domingos & Richardson, 2001), *etc*. It can be proven that the optimization problem of SSM is NP-hard, and many greedy methods have been proposed to address the challenges of combinatorial explosion in such constrained search problem, among which Cost-Effective Lazy Forward (CELF) (Leskovec et al., 2007) is deemed to be a most efficient one. CELF applies a "lazy-evaluation" strategy by maintaining a ranking list of the candidates according to their marginal rewards in decreasing order. Considering that the change of marginal gains between two adjacent iterations would not be too large, only several candidates on the top need to be re-evaluated and those remain at the top would be chosen. In this manner, re-evaluations for the rewards of the candidates below the top candidates can be reduced, which can dramatically speed up the search process. However, in the practical scenarios of social sensing, many different users may detect the same event, which would compose a *tight coupling* structure in social sensing network and result in redundant detections. In pursuit of social sensing maximization, such ineffective detections should be avoided, and therefore more repeated re-evaluations of the candidates are required. As a result, the "lazy-evaluation" mechanism in CELF may be depreciated in terms of the efficiency.

To that end, in this paper, we propose two algorithms to solve the SSM problem in terms of computational efficiency and solution quality. We firstly present an improved algorithm based on the traditional CELF, called *List-enhanced CELF* (LeCELF), in which an extra event list is maintained to keep track of the corresponding detected events of each user, and the users that have large number of detected events are fed into a *target list*, such that the search space and re-evaluations are greatly reduced. Additionally, we observe a phenomenon that coparticipators of a random user, *i.e.*, users who have participated in the same event, may participate in more events than the random user does, which can be defined as the *participation paradox*. Such a phenomenon is also observed in large-scale networks and utilized as a random schema for the sensor-selection problem (Christakis & Fowler, 2010; Garcia-Herranz, Moro, Cebrian, Christakis, & Fowler, 2014). Inspired by this finding, we further develop a randomized selection algorithm called *Friends Random* (FRIENDOM), which prefers to select the most valuable coparticipator of a random user in a top-user set as sensor.

To validate the performances of the proposed algorithms, we conduct extensive experiments on two datasets. One is based on a real-world event diffusing data from *Sina Weibo*, the most popular microblogging platform in China, and the other concerns a simulation disease spreading network based on a real-world high-resolution human contact network from the work of Salathé et al. (2010). The results are shown that social sensors selected by our proposed objective function *pDA* can do better detection with less delay time. Furthermore, the improved LeCELF achieves the same detection performance as CELF-type algorithms, but the computational time is greatly reduced. Meanwhile, the social sensors selected by the FRIENDOM algorithm indeed have better performance than all of the baselines, including LeCELF, in terms of the average detection results.

The rest of the paper is organized as follows. Section 2 reviews related work of this paper. Section 3 introduces the decision framework and the related terminologies, and then formally defines the social sensing maximization problem in Section 4. Sections 5 and 6 detail the two types of algorithms respectively. We further conduct experiments on both simulation and real-world datasets in Section 7, and finally conclude the paper in Section 8.

2. Related work

Our work is related to the following streams of literature including topic detection, social sensing and social influence in network.

Topic Detection and Tracking (TDT) is one the most direct methods to discover trendy events in large scale text streams, which can generally be classified into three categories including topic-model based, feature-clustering based and document-clustering based method. Topic-model based methods present the emerging topic as a probabilistic distribution over words, in which variants of probabilistic topic models such as Twitter-LDA (Diao, Jiang, Zhu, & Lim, 2012), BBTM (Bursty Biterm Topic Model) (Yan, Guo, Lan, Xu, & Cheng, 2015) are proposed to track the dynamic topics in social networks. Xie, Zhu, Jiang, Lim, and Wang (2016) proposed TopicSketch, a sketch-based topic model to achieve real-time topic detection on Twitter. Huang et al. (2017) built an emerging topics tracking method, which aligns emerging word detection from temporal perspective with coherent topic mining from spatial perspective. Another prominent method in tracking topics is feature-clustering based methods, which view the event as a set of keywords, phrases and hashtags, and then cluster them into topics. Different features extracted from the text streams such as wavelet-based signals (Weng & Lee, 2011) and segment frequency (Li, Sun, & Datta, 2012) over a fixed time window were used to design different clustering methods. Schubert et al. (2014) designed the SigniTrend algorithm to track all the significant trending keyword pairs under a fixed amount of memory using a clustering approach. Document-clustering based methods cluster documents according to their similarity in content, and it can also help detect prevailing events in social media. Allan, Papka, and Lavrenko (1998) proposed a clustering algorithm of related documents, in which each document is represented as a point in a vector space, and each new coming document will be either clustered to the nearest one or labeled as a new event. Subsequently Brants, Chen, and Farahat (2003) extended it with incremental TF-IDF model and improved similarity score normalization. In addition, locality sensitive hashing are introduced to handle the large scale of texts in Twitter (Petrović, Osborne, & Lavrenko, 2010). The limitation of TDT-based methods for the goal of information acqusition lies in the fact that such methods are generally based on aggregate text streams, and thus it takes more time for the related texts to accumulate so as to be detected.

Social sensing is a newly emerged concept in recent years (Sakaki et al., 2010) that provides a micro perspective to detect outbreaks (Takahashi, Tomioka, & Yamanishi, 2011) in different applications, such as water contamination detection (Hart & Murray, 2010; Ostfeld et al., 2008), prediction of behavioral epidemiology and public health (Madan, Cebrian, Lazer, & Pentland, 2010), understanding socioeconomic environments (Liu et al., 2015), and earthquake prediction on Twitter (Sakaki et al., 2010). Leskovec et al. (2007) introduced a sensor placement problem in blog and water networks and proposed the CELF algorithm, whereas several other works have extended CELF to improve the efficiency (Goyal, Lu, & Lakshmanan, 2011b; Mirzasoleiman et al., 2015). Zhao, Lui, Towsley, and Guan (2014) considered efficient followee selection for cascading outbreak detection from the incomplete network. Mahmoody, Riondato, and Upfal (2016) designed a probing schedule to detect valuable information by probing a small set of nodes at each time step. Zhang, Wang, and Vassileva (2013) proposed a user-centric system called SocConnect for aggregating social data and providing personalized recommendation to each user. Furthermore, local information of the social network is also exploited for sensor selection. For example, a local randomness schema is applied for early detection of contagious outbreaks at universities (Christakis & Fowler, 2010), and outbreaks detection on Twitter (Garcia-Herranz et al., 2014). The underlying idea comes from the interesting phenomenon termed the “friend paradox” (Feld, 1991), which is also similar to the concept of immunizing random friends of a randomly chosen node, *i.e., Acquaintance Immunization* (Cohen, Havlin, & ben Avraham, 2003), provided with a theoretical guarantee. Although sensor placement problems have been studied in previous work, the concept of social sensing is merely addressed, and an objective function concerning the sensing ability has not been thoroughly formulated.

In addition, the proposed SSM problem is also related to **Social Influence** in networks. Generally speaking, the research on social influence mainly includes empirical studies on whether social influence indeed exists (Cha, Mislove, & Gummadi, 2009; Hill, Provost, Volinsky et al., 2006), how to infer the link probability between users (Goyal, Bonchi, & Lakshmanan, 2010), and how to select the most influential users in network (Goyal, Bonchi, & Lakshmanan, 2011a; Leskovec et al., 2007) which is known as *Influence Maximization* (IM) problem (Domingos & Richardson, 2001; Kempe, Kleinberg, & Tardos, 2003; Liu, Zhang, & Chen, 2014). IM problem was firstly studied by Domingos and Richardson (2001), and Kempe et al. (2003) further considered it as a combinatorial optimization problem. However, these algorithms may be faced with efficiency issues, and then TIM (Tang, Xiao, & Shi, 2014), NewGreedy and degree discount heuristics (Chen et al., 2009), and Explore-Exploit strategies (Lei, Maniu, Mo, Cheng, & Senellart, 2015) are proposed subsequently. The goal of IM is to find a set of nodes that can influence the largest number of users, which can be applied to many applications such as viral marketing and personalized recommendations. Differently, the SSM problem not only concerns the quantity of detected events, but also takes the delay time before detection into account.

3. Social sensing maximization

3.1. Effective information acquisition from social sensors

Social media users have to make decisions on what to read, whom to follow, whose tweets to retweet, etc. in pursuit of comprehensive information about recent updates and bursty events regarding their topics of interest. For example, users that are

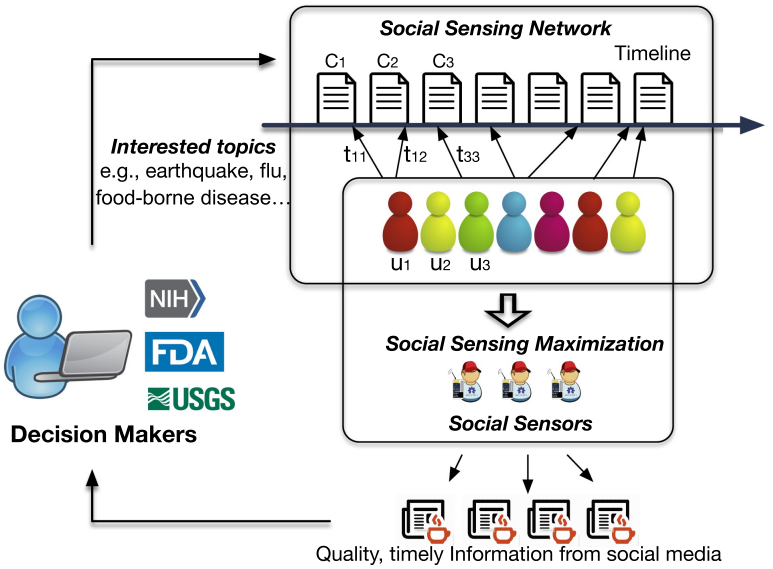

Fig. 1. Framework for effective information acquisition from social sensors.

interested in sports are eager to know the up-to-date match results; those who like online shopping may want to know the latest promotion and coupon information as soon as possible. Additionally, from a more professional perspective, some organizations also have requirements for effective information acquisition in social media. For instance, the National Institute of Health (NIH) can detect bursty flu events from spreading tweets about the symptoms of flu; the Food and Drug Administration (FDA) can also detect the outbreaks of food-borne diseases happening in certain regions from related posts in social networks. However, when faced with millions of posts generated every second, it is challenging to extract useful information from the timelines effectively. In addition, users can miss important information easily because the timelines are updating at an extremely fast pace.

As shown in Fig. 1, rather than directly extract useful and bursty events from large-scale text streams, we take a user-centric view by choosing a subset of user as "social sensors", who can act as intermediates to filter out irrelevant information and aggregate high-quality information effectively. Then, decision makers can obtain timely information about bursty events by just reading the tweets posted or retweeted by these social sensors.

To illustrate the effectiveness of user-centric tracking for information acquisition, we choose the incident of *expired vaccine* in Sina Weibo as a showcase. This incident was initiated by parents who speculated that a health center of Nantong City, China used expired vaccine to vaccinate children on October 11th, 2018. We then crawled all the related posts with the keywords *"expired vaccine"* from October 11th, 2018 to October 13th, 2018, and plotted the volumes of posts over time, as shown in Fig. 2. From the figure, we can see that the event was firstly released by an influential user, *Nantong Hotline*, who has more than 100,000 followers on Sina Weibo. However, in the first several hours, the volume of posts related to the event remained at a low level, until the time of 21:00, more than five hours later, at which the volume started to increase tremendously. While if we set the user *Nantong Hotline* as sensor, we would acquire the information immediately once the sensor releases the news.

Comparatively speaking, the keyword-based methods detect the event only when the frequencies of certain words surge to a

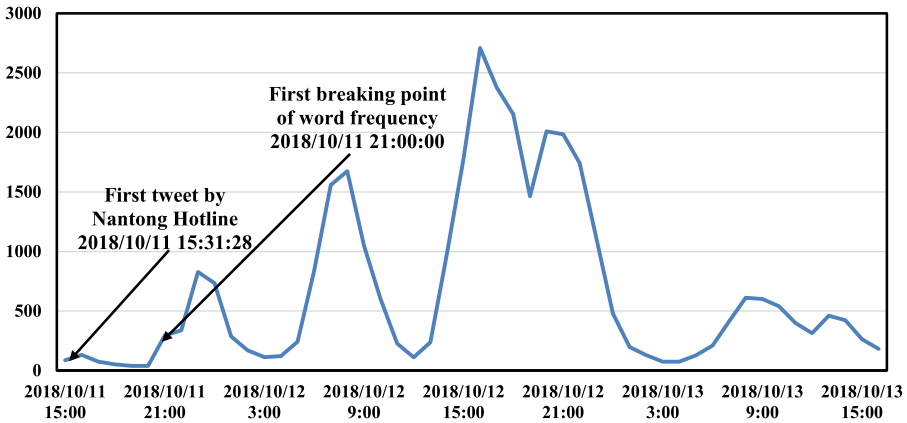

Fig. 2. The volumes of posts related to the incident of " expired vaccine" over time.

certain level, and the event can be represented by a set of keywords. Moreover, the selected sensors that might actively participate in their interested topics, and therefore the aggregated information acquired from them can be regarded as a set of filtered information with fewer noises. Therefore, good placement of social sensors in the network helps to get earlier and more comprehensive information than the keywords tracking methods.

Therefore, the core of the framework is how to select social sensors that can provide maximum coverage of the diffusing information about particular topics. Fig. 1 also displays that each individual user can act (*e.g.*, post or retweet) in response to the tweets on timelines, which provides clues for identifying the users who are more sensitive to the diffusing events. We formally define the social sensing network and then formulate the social sensing maximization problem for the effective information acquisition framework.

3.2. Social sensing network

As shown in Fig. 1, the users in social media posting/retweeting posts compose a bipartite network, which we call a *social sensing network*. The introduced network explicitly exhibits different events diffusing in the network and the users participating in the events. More formally, let $G = (V, E, C)$ be a social sensing network, where V and C represent two types of nodes. Each node $u \in V$ denotes a user, and $e \in E$ represent an edge that connects user node u with event node c, indicating that the user u participates in the diffusion process of the event $c \in C$. Meanwhile, the corresponding participation time stamp t_{uc} is also recorded along with the connecting edge. For example, in the social sensing network of Fig. 1, node u_1 participates in events C_1 and C_2 with edges connecting the corresponding event nodes, and the corresponding participation times are t_{11} and t_{12}, as shown on the edges. The starting time of an event c can be defined as the earliest time when the event is detected, *i.e.*, $t_c = \min_{u \in V} t_{uc}$. Then, the delay time of user u in event c is $t_{uc} - t_c$, reflecting its agility in detecting the event c.

Obviously, the defined social sensing network G is different from a traditional social network, in which there exist two types of heterogeneous nodes, with the relationship indicating users' joint participation in events, rather than real social relationships. The user-event relations can be further extended to user-user relationship, in which we can connect users who participate in the same event and regard them as "friends". The bipartite network is more beneficial to clearly display the participation structure. Particularly, when addressing the delay time for social sensing, we need to incorporate the participation time, which is better captured as link property in our bipartite network.

4. Problem statement

The general goal of the SSM problem is to select a subset of social sensors based on the event participation history to sense the diffusing cascades and trendy topics, in order to achieve effective information acquisition. Then, we can transform the problem into an optimized search in the social sensing network, in pursuit of maximum detection rates of diffusing events within a minimal delay time. Specifically, let the set of social sensors be A and the corresponding event set detected by A be C_A, and then a reward score $R(A)$ can be gained to evaluate the ability of the sensor set in acquiring the information concerning the detection likelihood and delay time. When a user u is selected as a social sensor and added to the set, it can yield extra gains since it detect more events. Additionally, it incurs an additional cost $f(u)$ due to the efforts required to watch the new posts or reposts of the sensor. The total cost can be the summation of the cost of each sensor, which can be defined as $f(A) = \sum_{u \in A} f(u)$. Generally, the efforts that can be spent on placing social sensors are limited and assumed to be constrained by a budget b, such that the total cost $f(A)$ should not exceed the budget. Note that the total cost may also be in other forms when needed.

Then, the problem can be defined as searching for a subset of nodes from social sensing network to achieve the maximum total reward. We call this the *Social Sensing Maximization* (SSM) problem, and formally define it as follows:

Definition 1. (**Social sensing maximization problem**). Given a social sensing network $G = (V, E, C)$ and a budget b for the social sensors, the social sensing maximization problem is to select a set of social sensors $A \subseteq V$ to maximize the reward $R(A)$ subject to the budget constraint $f(A) \leq b$.

Therefore, first of all, we need to address the form of reward function (social sensing ability) for each sensor in order to evaluate the selection criteria toward social sensing maximization. Intuitively, the key goals for information acquisition in social network should include both information comprehensiveness and timeliness. In other words, whether the information acquired from social sensors can cover most of the trendy topics and diffusing events in social networks, and whether such information is gathered with less delay time, can together determine the quality of social sensing. In prior work, the two aspects are denoted as Detection Likelihood (*DL*) and Detection Time (*DT*) respectively. However, the proposed SSM should not emphasizes on a single perspective, and thus a multi-criteria objective needs to be set up for the SSM problem. One natural solution is to use a weighted value to balance the amount and timeliness of detected information, that is weighted Detection Likelihood and Time (*wDLT*). However, since no priors can be used to set the weight values for each objective, and hence they can only be adjusted in an ad hoc manner through repeated experiments.

Therefore, to refrain from trying weights for each objective separately, we adopt a priority view rather than a weighted view to address the challenges of multi-criteria objectives. We assume that a quality sensor should first detect as many events as possible to guarantee information comprehensiveness. Then, under the condition that most trendy topics have been covered by the social sensors, we can proceed to consider the delay time of detecting these events. Therefore, to make sure that the reward evaluation function prioritizes sensors with high detection likelihood, a large penalty would be imposed on user u if a diffusing event c is not

detected by u. In contrast, if an event c is detected by the sensor u, the sensor can be endowed with a reward $R_c(u) = t_c - t_{uc}$, which measures the actual delay time of the event c to user u. Regarding the goal of social sensing maximization, we adopt the least delay time as the reward for detecting the corresponding event, whereas other late detections provide no extra rewards due to redundant detections. Thus, we propose a novel objective function as *preference-based Detection Ability* (*pDA*),

$$R_c(u) = \begin{cases} t_c - t_{uc}, & \text{if } c \text{ is detected} \\ -\infty, & \text{if } c \text{ is not detected} \end{cases}. \tag{1}$$

Therefore, in the following section, we adopt *pDA* as the reward function for each user in the SSM problem. Based on the above definitions, the SSM problem can be mathematically formulated as follows:

$$\begin{aligned} &\max_{A} \quad \sum_{c \in C} \max_{u \in A} R_c(u) \\ &s.\,t. \quad \sum_{u \in A} f(u) \le b \end{aligned} \tag{2}$$

5. Algorithm LeCELF

5.1. Greedy method and re-evaluations

Given the mathematically formulated SSM problem, we find it difficult to obtain a global optimal solution, since the problem can be analogized to *budgeted set cover, influence maximization*, and *sensor placement problem*, which can be proven to be NP-hard. One possible solution to solve NP-hard problems is to exploit a greedy method to derive local optima, and a nondecreasing and submodular objective function can guarantee a good approximation, which is at least $1\text{-}1/e \approx 63\%$ of the optimal solution (Nemhauser, Wolsey, & Fisher, 1978). Specifically, a function with submodularity satisfies the fact that the marginal reward gained by adding an element s to a smaller set S_1 is no less than that when the element is added to a larger set $S_1 \subseteq S_2$, which can guarantee a near optimal approximation when exploiting a greedy method. As a matter of fact, our proposed reward function *pDA* for SSM problem is also submodular and nondecreasing.

Theorem 1. *The objective function pDA for the SSM problem is submodular.*

Proof 1. Denote *pDA* as a nondecreasing set function F, and assume $\forall S \subseteq T \subseteq V$, $\phi = T - S$. For user $\forall s \in V \setminus T$, if the event set s detects have no intersection with that of ϕ, then $F(S \cup s) - F(S) = F(T \cup s) - F(T)$ according to definition of F. Once s and ϕ have joined in at least one or more of the same events, the marginal gain of s will be re-evaluated by removing the overlapping parts with ϕ, leading to $F(S \cup s) - F(S) > F(T \cup s) - F(T)$. In summary, *pDA* satisfies the requirement of *submodularity*.

Therefore, we exploit the simple greedy method to select social sensors, and in each iteration, the user with the maximal marginal gain is chosen as a social sensor and added to the social sensor set A. However, although submodularity can guarantee a good approximation, the naive greedy algorithm is too time-consuming when the social sensing network contains a large number of users. Specifically, in each iteration, all the candidates have to be probed by repeatedly evaluating the marginal gain after a new node added to the social sensor set, since the events detected by both any candidate and social sensors can not account for the marginal gain of that candidate, which we call *re-evaluation*.

Though the greedy method can guarantee a near-optimal solution, the high computational complexity caused by numerous repeated re-evaluations would limit the application of the method. Several methods have been proposed in order to reduce re-evaluations for each candidate, and among them CELF (Leskovec et al., 2007) achieves dramatic speedup with a good approximation guarantee through a "lazy-evaluation" mechanism. Reflected in the social sensing maximization, the mechanism maintains a ranking list to keep track of the marginal gain of each sensor in the iterative process. In the beginning of each iteration, all the sensors are first marked as "*not re-evaluated*". Then, the sensors are re-evaluated in a decreasing order of the ranking list, and reinserted into the ranking list again. If the re-evaluated marginal reward of the top sensor is found to be greater than that of the candidate sensors below it, the top sensor can directly be chosen and the re-evaluations for the others are not needed. If it is not the case, the sensor would be inserted into the ranking list according to the re-evaluated values.

It is worth noting that an underlying condition in CELF is that the marginal reward of a candidate between two iterations does not change dramatically, and thus, the possibility of the top candidate staying at or near the top will be high. Under such condition, the number of re-evaluations is small enough, and hence, the efficiency of CELF can be guaranteed. To ensure this condition, a prerequisite is that the candidate sensors have few intersections with selected social sensors in common detected events, which we call the *loose coupling effect*. According to the objective function of the SSM problem, for any candidate sensor, the common event detections with selected sensors would contribute no rewards to sensing the diffusing information, and therefore, if the candidate sensors are tightly coupled, the marginal rewards will change greatly when a candidate is selected as the social sensor, which further influences the efficiency of CELF.

5.2. List-enhanced CELF

As a matter of fact, in real social media scenarios, candidates become more tightly coupled as the size of social sensors increases.

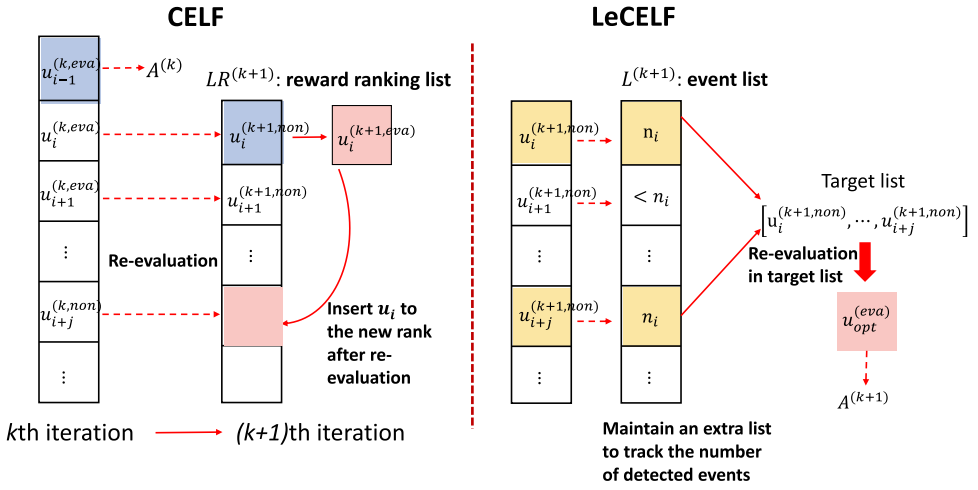

Fig. 3. Re-evaluation comparison between CELF and LeCELF.

For example, some users are more likely to be involved in different diffusing events, and then they show a tight coupling structure with other candidates. Once these users are selected as social sensors, the marginal rewards of other candidates will decreases greatly compared with the previous iteration, spending much time in re-evaluating these candidates. Particularly, when the set of social sensors becomes larger, the tight coupling effects would become more significant and the "lazy-greedy" strategy would be inefficient. Therefore, we proceed to propose a novel method to enhance the CELF algorithm by introducing an extra list. For the sake of clear illustration, we show the ranking list and the re-evaluation procedures in Fig. 3. We denote the ranking list of the k-th iteration as $LR^{(k)} = [R_{i-1}^{(k,eva)}, R_i^{(k,eva)}, ..., R_{i+j}^{(k,non)}, ...]$, where $R_i^{(k,eva)}$ represents the re-evaluated reward of u_i in the k-th iteration with the marker *eva*, whereas the marker *non* means that the user has not yet been re-evaluated. According to CELF, u_{i-1} would be added to the sensor set in the k-th iteration. Then, in the $(k+1)$-th iteration, we can first re-evaluate the marginal reward of the top user u_i in the list. With respect to the formula of the objective function, the social sensing reward of u_i can be re-evaluated by removing the redundant detections from the detected events C_i compared to the events detected by the current social set C_A^k.

$$R_i^{(k+1,eva)} = R_i^{(0)} - \sum_{c \in C_i \cap C_A^k} \delta_i^c, \tag{3}$$

where R_i^0 represents the initial reward u_i would receive when added to an empty sensor set.

According to CELF, it requires that $R_i^{(k+1,eva)} \geq R_{i+1}^{(k+1,non)}$ to make u_i be added to the sensor set. While in reality, the top user may drop in several ranks Δ in the ranking list after reevaluation, such that $R_i^{(k+1,eva)} \leq R_{i+\Delta}^{(k,non)}$, and more re-evaluations would be needed in this case as shown in Fig. 3. Obviously, the number of dropped ranks Δ is closely related to the set intersection $\mathcal{X}_i^k = C_i \cap C_A^{(k)}$. When $\mathcal{X}_i^k \to \varnothing$, *i.e.*, the sensor is loosely coupled with previously added sensors, the marginal reward would be higher and the sensor is more likely to remain at or near the top of list, in which CELF would perform well at avoiding re-evaluations. However, if $\mathcal{X}_i^k$ is large, indicating a tight coupling effect, the ranks would drop greatly, and hence, more re-evaluations are required.

Therefore, rather than only emphasize the marginal rewards, the number of detected events is deemed to be crucial in selecting eligible social sensors in an effective manner. Specifically, as shown in Fig. 3, we introduce an extra event list, which records the number of events that each user can detect, except for the reward ranking list. The general principle is that the users with more detected events are more likely to be subjected to the tight coupling effects, and hence can enjoy higher marginal rewards. Thus, the target list including the potential optimal sensor in the current iteration is constructed by comparing the number of detected events of the nodes in the event list. Then, the sensors in this shrunken target list will be re-evaluated, which can greatly reduces the number of re-evaluation times. As can be seen from Fig. 3, the event list plays an indispensable role in the newly proposed method, and we refer to it as **L**ist-**e**nhanced CELF (LeCELF). The algorithmic details are presented in Algorithm 1.

As can be observed from in Algorithm 1, the improved algorithm LeCELF introduces the event list L to record the number of detected events for each user $v \in V \setminus A$ in the network. We then sort the candidate sensors according to the number of detected events in decreasing order, and the sensors at the top of the list are chosen to form the target list, in which several sensors may have the same values (lines 8). Then, the re-evaluation strategy in CELF is applied only to the target list to seek the optimal node and insert it into the social sensor set A (lines 12–20). In addition, the event list L has to be updated when adding a sensor to the set A, *i.e.*, all the events detected by the sensor should be removed to avoid redundant detection, and subsequently, the number of detected events of each sensor in L should be modified (lines 21–25).

5.3. Discussions

Compared to CELF, our proposed algorithm LeCELF maintains a list recording users' event detection likelihood as a supplement for the marginal gain of each user. LeCELF succeeds in obtaining a focal group of sensors which potentially contains the local optimal

Input: $G = (V, E, C)$, Reward R, Event list L, Budget b, Target list TAR
Output: Social sensor set A

$A \leftarrow \emptyset$
for all $s \in V$ **do**
    $\delta_s \leftarrow \infty$
end for
while $\exists s \in V \setminus A : f(A \cup s) \leq b$ **do**
    $TAR \leftarrow \emptyset$
    $max = 0$
    $TAR \leftarrow argmax_v L$ ▷ construct the target list
    for all $s \in TAR$ **do**
        $\delta_s \leftarrow \infty; cur_s \leftarrow False$
    end for
    for all $s \in TAR$ **do**
        $s^* \leftarrow argmax \delta_s$
        if cur_{s^*} **then**
            $A \leftarrow A \cup s^*; break$
        else
            $\delta_{s^*} \leftarrow R(A \cup s^*) - R(A)$
            $cur_{s^*} \leftarrow true$
        end if
    end for
    for all $c \in C_{s^*}$ **do** ▷ user s^* detect events C_{s^*}
        for all $u \in U_c$ **do** ▷ users that detect event c
            $L_u \leftarrow L_u - 1$
        end for
    end for
end while
return A

Algorithm 1. LeCELF.

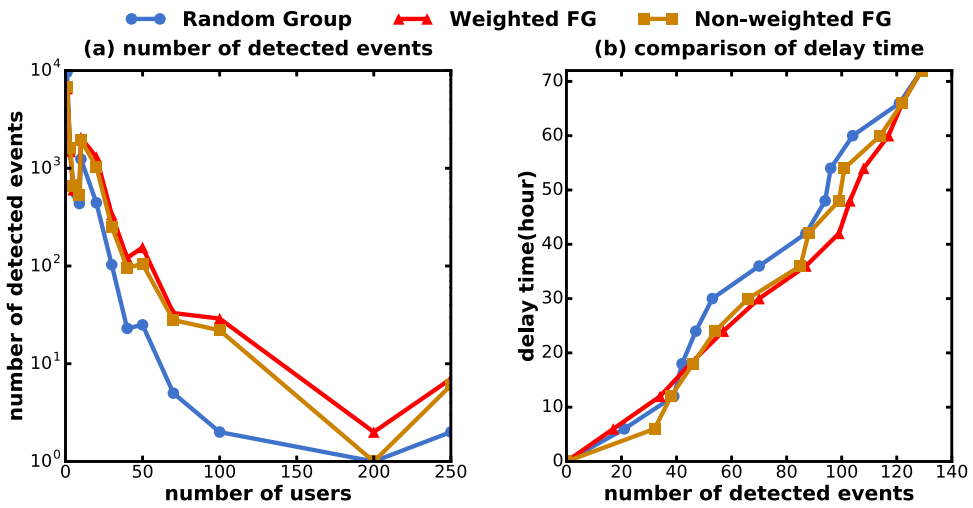

Fig. 4. Comparisons between random group and friend group in detecting events.

node, *i.e.*, the sensors that can detect the most events in the current iteration. As a matter of fact, the computational time of CELF is shown to increase nonlinearly with the size of sensors k. However, regarding LeCELF, the construction of target list in each round requires $\Omega_1(\|V\|)$, and choosing the local optimal sensor from the target list costs $\Omega_2(\|top\_list\|)$. Moreover, maintaining the event list L costs another $\sum_{c\in c^{s^*}} size(c)$. Therefore, the overall complexity of LeCELF is $k(\Omega_1(\|V\|) + \Omega_2(TL_i) + \sum_{c\in c^{s^*}} size(c))$, which almost increases linearly with k (*i.e.*, the size of the social sensor set) and can be easily extended to large-scale social sensing networks and large social sensor sets.

6. Algorithm FRIENDOM

Although the algorithm LeCELF returns the social sensor in an efficient manner, it can still possibly be trapped at local optima. Therefore in this paper, we seek to weaken the local trap effect by introducing a randomness mechanism, which has already been exploited in other algorithms, such as genetic algorithms. Instead of greedily searching for the local optimal sensor in each iteration, we introduce a more flexible strategy by considering the users' coparticipation behaviors in social sensing network, which is merely addressed in CELF-type algorithms. The "friendships" formed by the coparticipation behaviors are different from those in traditional social network, which indeed provide a new perspective to choose quality sensors.

We first illustrate the properties of "friendships" in social sensing network through observations from the real-world microblog user-event participation data. To start, a set of 20,000 individual users in the social network are randomly selected as a "random group" (*RG*), while "friends" of the users in the random group are randomly chosen to form a "friend group" (*FG*) of equal size. The "friend group" is formed by two different strategies: one is to select users according to the occurrence of their participation, *i.e.*, the "friends" with larger detection likelihood have higher probability to be chosen, which we call the "Weighted friend group" (*Weighted FG*); the other is to select users uniformly to yield the "Non-weighted friend group" (*Non-weighted FG*). We then compare the number of detected events between these different groups. As demonstrated in Fig. 4(a), compared to the random group, users in the FG can detect more events, which further implies that the pure random friends cannot guarantee the quality of sensors, while "friendships" in social sensing network can guide better choices of social sensors. Moreover, we also validate the delay time with respect to these groups by randomly selecting 20 users to form the "random group", and randomly choosing 20 "friends" of these users to form the Weighted FG and Non-weighted FG. We extract 140 widespread events that are detected at least once in each group, and show the delay time of detection by these groups respectively. Fig. 4(b) shows that users in the FG would take less delay time than the RG in detecting the same number of events.

It is worth noting that the analysis can be replicated with the same observations that the coparticipators of a random user may be involved in more events with less delay time. We call such phenomenon in social sensing network "*participation paradox*". Similar phenomenon has also been witnessed in prior studies, known as "friendship paradox" (Evans, Kairam, & Pirolli, 2010; Feld, 1991). It has been proven that a randomly selected "friend" of a node has higher degree on average than itself (Cohen et al., 2003). Thus, such phenomenon sheds light on finding social sensors in a more heuristic manner. Inspired by the phenomenon, we propose a new random selection algorithm, named FRIENDOM. Specifically, in each iteration, instead of greedily choosing the local optimal element from the event list L, as in the LeCELF algorithm, FRIENDOM starts with a randomly selected node u from the top N_1 nodes in the event list (line 4–5). Then, considering the "participation paradox" in the social sensing network, for each event c in the corresponding detected event set C_u, other nodes that participate in the same event c together with u, *i.e.*, the "friends" of u are inserted into the set S (line 7–11). Subsequently, we re-evaluate all the marginal rewards of the sensors in the set S and choose the optimal one in the current selection (line 12–16). Algorithm 2 provides the details for FRIENDOM.

In a nutshell, the proposed algorithm FRIENDOM can be viewed as a combination of randomness mechanism with the computed

Input: Graph $G = (V, E, C)$, Reward R, Event list L, Budget b
Output: Social sensor set A

$A \leftarrow \emptyset$; $E \leftarrow \emptyset$; $N_1 = 1000$; $N_2 = 2000$
while $\exists s \in V \setminus A : f(A \cup s) \leq b$ **do**
    $S \leftarrow \emptyset$
    $L \leftarrow$ sort L with decreasing order
    seed $u \leftarrow$ random u from $L[: N_1]$
    $C_u \leftarrow C_u \setminus E$
    for all $c \in C_u$ **do**
        for all $v \in V(c)$ **do**
            add v to S
        end for
    end for
    $S \leftarrow S[: N_2]$ ▷ according to occurrence of v
    for all $v \in S$ **do**
        $R_v \leftarrow R(A \cup v) - R(A)$
    end for
    $s^* \leftarrow argmax_{v \in S} R_v$
    add s^* to A; add C_{s^*} to E
    for all $c \in C_s^*$ **do**
        for all $v \in c_v$ **do**
            $L_v \leftarrow L_v - 1$
        end for
    end for
end while
return A

Algorithm 2. FRIENDOM.

marginal rewards. The random user acting as the seed is selected from a pool of users with relatively high reward, which guarantees the overall performance. Instead of choosing the local optimal node in each iteration, the randomness driven by the phenomenon of the "participation paradox" in social sensing network can help broaden the range of candidates, since some current suboptimal sensors would turn out to perform better than the sensor set formed by local optima.

We also analyze the time complexity of FRIENDOM. The major procedures of the algorithm consist of two parts, *i.e.*, selecting the sensors and maintaining the event list L. As stated in Algorithm 2, it takes $\Omega_1(\|aver_v{*}aver_c + N_2\|)$ to select sensors in each iteration, where $aver_v$ denotes the average number of nodes in each event, and $aver_c$ denotes the number of events in which a node may participate. This procedure is indeed a constant, such that the overall complexity of FRIENDOM grows linearly with the size of the sensor set. However, since the algorithm introduces randomness schema, the results may fluctuate to some degree. Therefore, in practice, we can run the algorithm repeatedly and return the set that achieves the optimal detection performance.

7. Experiments

7.1. Experimental data

We conduct experiments on two datasets: one is a synthetic disease spreading based on a real-world contact network, the other is an event participation dataset from microblog.

7.1.1. Disease spreading network

We simulate the spread of infectious disease in a real-world high-resolution human contact network (Salathé et al., 2010). The contact network records both the interaction time and interaction partners among 788 individuals at a U.S. high school during a typical school day, revealing a high-density network with typical small-world properties. We apply a general Susceptible-Exposed-Infected-Recovered (SEIR) model to simulate the disease spreading process. We consider simulations of 3 days length, with 12-hour simulation time steps. The simulation initially starts from a randomly selected individual, and transmits from node i to its neighbor j with the probability $p_{ij} = 0.003$. Thus the probability of transmission per time step (12 h) from an infectious individual to a susceptible individual is $1 - (1 - 0.003)^{w_{ij}}$, where w_{ij} is the weight of the contact edge. Once a susceptible individual becomes exposed, it will become infectious after 6 hours and starts to spread the disease to others. The infection time distribution is assumed to follow a Poisson distribution with λ equals to the weight of the contact edge. Each infectious node can recover with a probability of $1 - 0.95^t$ per time step, and t represents the number of time steps after being infectious. We run the simulations for 10,000 times and record the spreading details in the contact network to construct the social sensing network, with each individual infected by a disease as a user detecting a diffusing event.

7.1.2. Weibo network

Weibo is one of the largest microblog platforms in China. We treat the posts and their reposts with the same hashtag (*i.e.*, embedded in the symbol "#") as the same event due to the mechanism in microblog that a hashtag generally denotes a particular topic under heated discussion. We construct an event-diffusing dataset based on the marketing information in Weibo, which often contains sales and campaigning keywords such as "present", "gift" and "free". Specifically, we extract 8489 events from the dataset via keyword matching in the prefix of hashtags (##), for the period from September 1, to September 7, in 2013, with each event diffusing for 3 days. For repeated participating users, only their earliest posts are included in the dataset. Each event participation record contains detailed information about the user and the time. We split the dataset into training set and testing set according to the time of events. The events and the participating users in the first 4 days are regarded as training set, which has 3750 events with 0.45 million users; and the remaining data is used as testing set with 4739 events and 1 million users; we refer to the testing set as *Weibo-Test*.

In addition, to further validate the event detection performances, we further construct another independent testing set in a different timespan from September 22 to September 28 in 2013, which has no intersections with the training set. This testing set has 10,012 events with 1.67 million users, and we refer to it as *Weibo-INDTest*.

7.2. Objective function comparison

We first examine the effectiveness of different objective functions in detecting diffusing events on all the datasets. As mentioned in Section 4, we compare the following objective functions: *DL, DT, PA, PA/DT, wDLT* and *pDA*. Note that *PA* denotes the number of users who get involved in an event before it is detected, and *PA/DT* is a ratio between *PA* and *DT*, aiming at achieving both goals simultaneously. For fair comparison, we uniformly adopt the LeCELF algorithm to obtain optimal solutions under the same budget constraint with different settings of objective functions. Meanwhile, we also propose two heuristic objective functions, *CasCount* and *CasTime*. *CasCount* selects social sensors based on the number of events in which the user participates, whereas *CasTime* prefers the nodes that have minimal cumulative delay time among all the users.

As shown in Fig. 5, the curves of both *wDLT* and *pDA* lie in the upper region for the three datasets, which means that given the same number of detected events, *wDLT* and *pDA* achieve the earliest event detections in comparison with other objective functions. Considering that both objective functions *wDLT* and *pDA* incorporate *DL* and *DT* and they also have similar sensing performances, we further compare the results between the two objective functions, and the results are depicted in Fig. 6. It can be seen that the weight

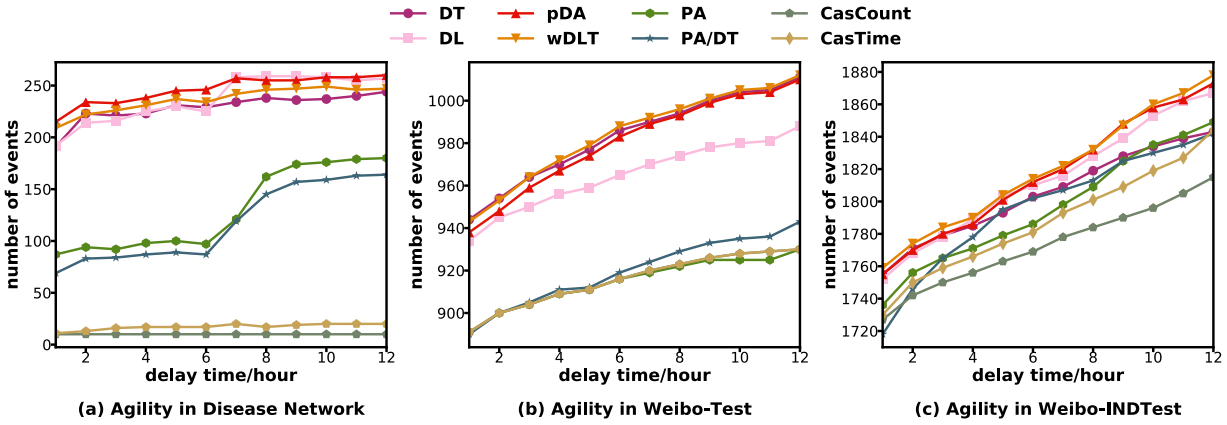

Fig. 5. Performance comparison with different objective functions.

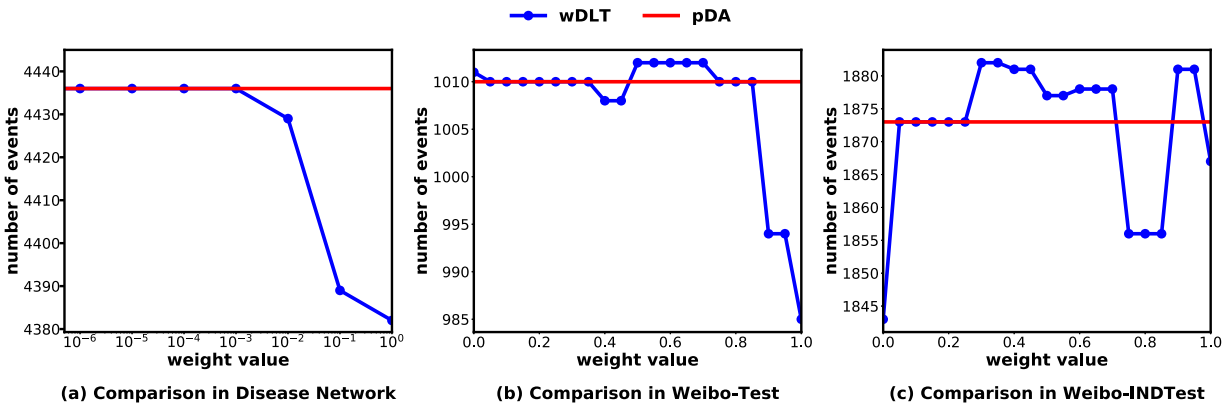

Fig. 6. Performance comparison under different weights for *wDLT* and *pDA*.

value can greatly influence the detection performances with the function *wDLT*, leading to unstable social sensing performances. Particularly, the number of events detected by *wDLT* in the *disease spreading network* exhibits a declining trend and finally equals that of *DT* when the weight value is greater than 1. This detailed comparison clearly demonstrates the effectiveness and rationality of the novel objective function *pDA*.

7.3. Performance comparison

With respect to the objective function *pDA*, we compare the proposed algorithm LeCELF and FRIENDOM with other baseline methods in terms of detection performances and computational efficiency. The baseline methods include *CasCount, CasTime*, and CELF, which have been introduced in previous sections. We also compare the proposed algorithms with the following baseline methods. Note that for better illustration of the differences between algorithms in the disease spreading network dataset, we set the result of *CasCount* as baseline 10, and draw the differences of that between other functions and the baseline.

- CELF++ (Goyal et al., 2011b): An improved greedy algorithm based on CELF.
- WgRandom: Randomly sample sensors from the top 1000 users in ranking list according to *pDA* by using the number of detected events as weight.
- *K_sum* (Pei, Muchnik, Andrade Jr, Zheng, & Makse, 2014): Select sensors according to the sum of degree of the nearest neighbors, where the degree denotes the number of detected events. Note that this method requires the whole network structure and thus can only be applied in the disease spreading network.

We regard the number of sensors as the budget and vary the number from 10 to 60, and it is worth noting that the results of the methods with randomness (*e.g.*, WGRandom, *K_sum*) are reported based on the average performance with 10 replications. Fig. 7 shows the comparative results of the number of detected events under the setting of specific number of sensors. CELF-type algorithms indeed produce the same detection performances and differs only in the computational time. Thus, we only show the results of the proposed LeCELF. In general, the proposed LeCELF and FRIENDOM outperform all the baseline methods in the three datasets, and FRIENDOM performs the best in both of the Weibo dataset, while it is interesting to observe that in the synthetic disease spreading

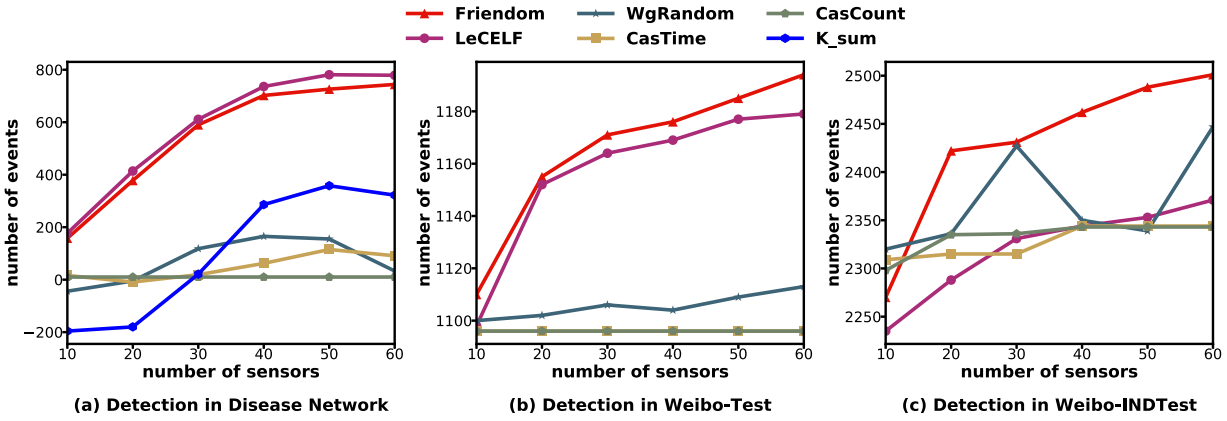

Fig. 7. Detection comparison of different algorithms.

network, LeCELF performs the best. The reason of the inconsistent performances is that the distribution of the number of events detected differs in the two networks. In the real-world Weibo network, the number of detected events follows a long tail distribution. However, in the simulated disease network, all the nodes are initially randomly selected and share the same probability of being infected, resulting in homogeneous infection probability. In this case, FRIENDOM cannot play to its strengths to spread out the sensors in the network and reduce the effect of local optima traps. Hence, the sensors selected by FRIENDOM perform worse than those selected by LeCELF in this case. *K_sum* shows an inclining trend as it gradually picks more diverse sensors but is still inferior to the proposed methods.

We then compare the agility of sensors, *i.e.*, the delay time in detecting the same number of events. As shown in Fig. 8, LeCELF and FRIENDOM still outperform other algorithms, and FRIENDOM performs a bit better than LeCELF in the three datasets. Overall, FRIENDOM enjoys advantages over LeCELF when considering both goals for social sensing maximization.

In addition, we conduct significance test to demonstrate the robustness of the proposed algorithms. Specifically, we repeatedly simulate the disease spreading process for 30 times, and obtain the social sensors from each simulation dataset for each method respectively. We then implement a two-sample one-tailed *t*-test between the proposed methods and the baseline methods. As can be seen in Table 1, no matter on the number of detected events or the delay time, the proposed LeCELF and FRIENDOM can achieve significantly better results than the baseline methods.

7.4. Analysis of redundant detection

We further analyze the difference between the event sets detected by different sensors. We generally assume that each quality social sensor can detect particular subset of events with few redundancy, or in other words, an effective social sensor should avoid redundancy in event detection because the redundancy would not bring in extra rewards. In order to investigate the selected sensors in terms of the detection quality, we further show the ratios of redundant events under different algorithms. Specifically, if a event is detected by more than two sensors, we regard such detection as a redundant detection.

Fig. 9shows the ratios of redundant events detected by different methods with respect to the numbers of sensors. As can be observed from Fig. 9, the heuristic methods such as *CasCount, CasTime*, and *WgRandom* have obviously higher ratios of redundancy, while those of LeCELF and FRIENDOM remain at a relatively low level. Note that FRIENDOM has a little bit more redundant event

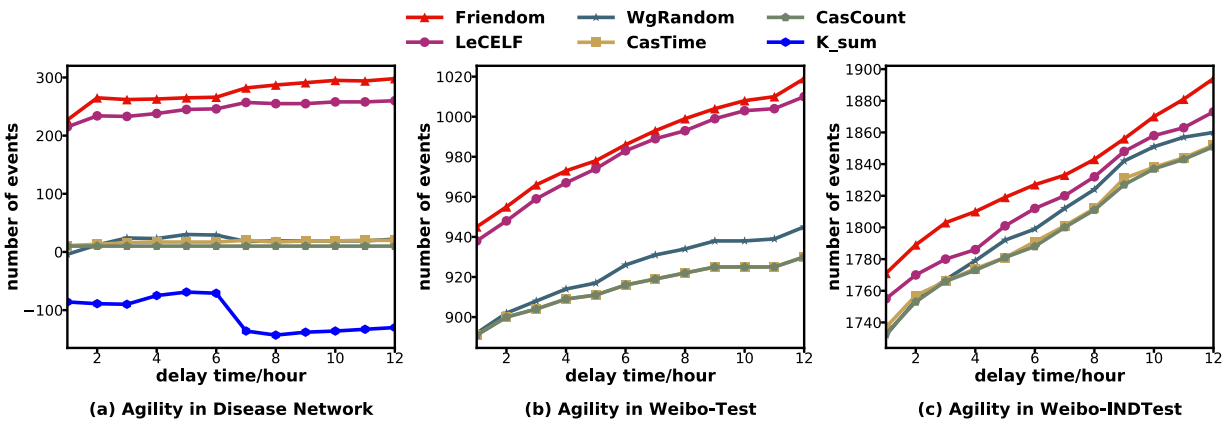

Fig. 8. Agility comparison of different algorithms.

Table 1
Significance test on disease spreading network.

Methods	Baselines	Detection events	delay time
FRIENDOM	> CELF	0.0014***	0.012**
	> *CasCount*	0.001***	0.0036***
	> *CasTime*	8.5E-04***	0.0029***
	> *WgRandom*	0.0011***	0.004***
	> *K_sum*	7.89E-05***	7.02E-04***
LeCELF	> =CELF	–	–
	> *CasCount*	0.0011***	0.0046***
	> *CasTime*	7.41E-04***	0.0039***
	> *WgRandom*	4.85E-04***	0.005***
	> *K_sum*	2.42E-05***	8.88E-04***

[1] Note: The value in the table represents the calculated p-value of *t*-test. The symbols > means that the method is advantageous over the compared method; ***, ** and * indicate the significance levels at 0.001, 0.01 and 0.05, respectively. The symbol "-" indicates all the CELF-type methods indeed have the same results and would show no differences in terms of the detection performances.

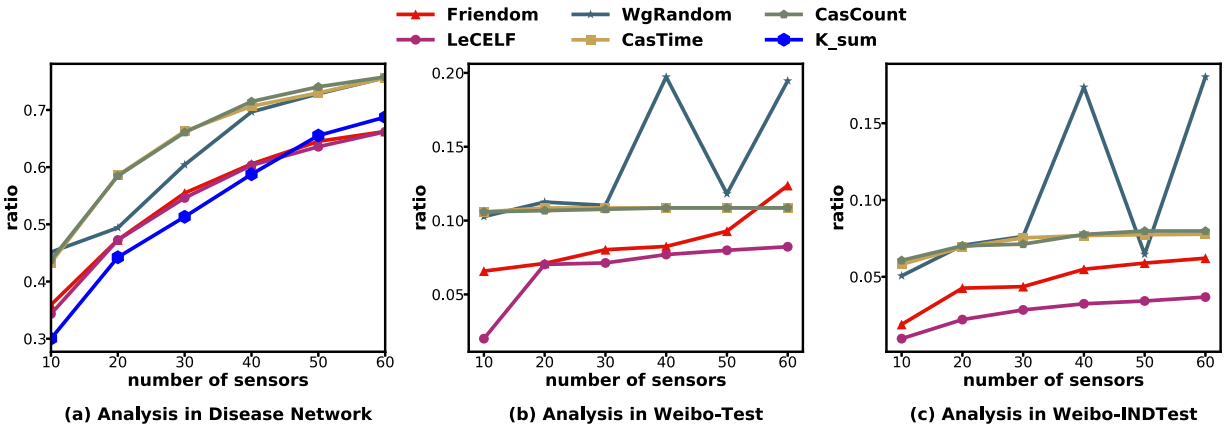

Fig. 9. Comparison of redundant events detected by different algorithms.

detections compared with that of LeCELF. Since the FRIENDOM allows for a randomness schema in forming the sensor set, it is more likely that the sensors with duplicate detections would be chosen than the local greedy search in LeCELF. Such redundancy in FRIENDOM would not depreciate the detection performances compared with the local greedy search as shown in Fig 7, which may be possibly due to the fact that alert sensors can still get involved in a few most popular events but they still maintain their sensing ability on particular topics. Thus, though the randomness tolerates a moderate level of redundancy in event detection, it may guarantee a more robust detection.

7.5. Runtime comparisons

Moreover, we also compare the running time of various algorithms, and the results are shown in Fig. 10. Note that we neglect the initialization process and count the runtime starting from the second iteration. It is shown that in the both datasets, the runtime of LeCELF is significantly less than that of CELF. For example, when the sensor set size is 60, the runtime of the two algorithms are 14.60s and 29.52s, respectively, with 50.54% reduction in the Weibo-Test, whereas for in the simulation disease dataset, they are 17.83s and 99.63s, respectively. CELF++ performs the worst among the three CELF-type methods, reaching 95.04s in real-world dataset and 314.46s in the simulation dataset. This phenomenon also demonstrates the analysis presented in 5.2 that as more sensors are selected, the more distance from the re-evaluated node to the top node. Correspondingly in LeCELF, the size of the latent candidate list lessens, and the growing rate of the running time decreases as more sensors are selected. Comparatively speaking, the improved CELF++ cannot accelerate the efficiency in SMM problem because it recomputes the marginal gain twice for each node. The runtime of FRIENDOM is between those of LeCELF and CELF, and it also remains stable against the number of sensors. We can see that as the number of sensors increases, the running time of FRIENDOM gradually approaches and is even less than that of LeCELF. For example, the runtime is approximately 7.9% faster than LeCELF, while it is 54.44% faster than CELF when the number of sensors is 60. Moreover, the runtime of the *K_sum* is similar to that of LeCELF because the data renewal mechanism of each node is the same as that in the disease network.

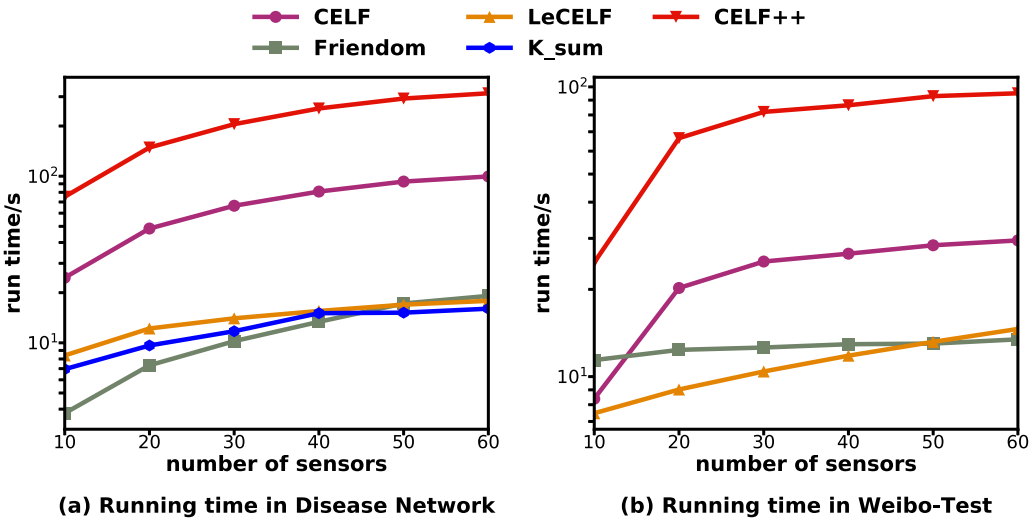

Fig. 10. Experimental results for the runtime.

8. Conclusion

In this paper, we proposed an effective information acquisition framework to obtain comprehensive and timely information in social media. The major contribution is the novel perspective of acquiring information by setting up social sensors in the network, i.e., acquiring information in user-centric view. To achieve the goal of social sensing maximization, the sensing ability of each individual user is measured a newly proposed measure, *pDA*, to incorporate both information comprehensiveness and timeliness. We proposed two algorithms toward the optimization goal: the first one is LeCELF, which greatly reduces the repeated re-evaluations aroused from a tight coupling structure and improves the "lazy-evaluation" strategy of CELF. Additionally, a heuristic algorithm with randomness, FRIENDOM, is also proposed to avoid local traps and achieve better detection performances. From a practical perspective, the algorithms can be applied to various scenarios in which we need to timely acquire high-quality information from large-scale and fast-spreading information streams for particular purposes, including outbreak detection for news agencies, earthquake detection, food-borne disease detection for official organizations, etc.

Acknowledgments

Dr. Guannan Liu was supported by National Natural Science Foundation of China (NSFC) (71701007). Dr. Junjie Wu was supported by National Natural Science Foundation of China (NSFC) (71725002, 71531001, U1636210, 71471009), and Fundamental Research Funds for the Central Universities.

Supplementary material

Supplementary material associated with this article can be found, in the online version, at 10.1016/j.ipm.2019.01.009 .

References

Allan, J., Papka, R., & Lavrenko, V. (1998). *On-line new event detection and tracking. Proceedings of the 21st annual international ACM SIGIR conference on research and development in information retrieval*. ACM37–45.

Brants, T., Chen, F., & Farahat, A. (2003). *A system for new event detection. Proceedings of the 26th annual international ACM SIGIR conference on research and development in information retrievalSIGIR '03*New York, NY, USA: ACM330–337.

Cha, M., Mislove, A., & Gummadi, K. P. (2009). *A measurement-driven analysis of information propagation in the flickr social network. Proceedings of the 18th international conference on world wide web*. ACM721–730.

Chen, W., Wang, Y., & Yang, S. (2009). *Efficient influence maximization in social networks. Proceedings of the 15th ACM SIGKDDinternational conference on knowledge discovery and data mining*. ACM199–208.

Christakis, N. A., & Fowler, J. H. (2010). Social network sensors for early detection of contagious outbreaks. *PloS One, 5*(9), e12948.

Cohen, R., Havlin, S., & ben Avraham, D. (2003). Efficient immunization strategies for computer networks and populations. *Physical Review Letters, 91*, 247901.

Diao, Q., Jiang, J., Zhu, F., & Lim, E.-P. (2012). *Finding bursty topics from microblogs. Proceedings of the 50th annual meeting of the association for computational linguistics: Long papers - volume 1ACL '12*536–544.

Domingos, P., & Richardson, M. (2001). *Mining the network value of customers. Proceedings of the 7th ACM SIGKDDinternational conference on knowledge discovery and data mining*. ACM57–66.

Evans, B. M., Kairam, S., & Pirolli, P. (2010). Do your friends make you smarter?: An analysis of social strategies in online information seeking. *Information Processing & Management, 46*(6), 679–692.

Feld, S. L. (1991). Why your friends have more friends than you do. *American Journal of Sociology, 96*(6), 1464–1477.

Garcia-Herranz, M., Moro, E., Cebrian, M., Christakis, N. A., & Fowler, J. H. (2014). Using friends as sensors to detect global-scale contagious outbreaks. *PloS One, 9*(4), e92413.

Goyal, A., Bonchi, F., & Lakshmanan, L. V. (2010). *Learning influence probabilities in social networks. Proceedings of the 3rd ACM international conference on web search and data mining*. ACM241–250.
Goyal, A., Bonchi, F., & Lakshmanan, L. V. (Bonchi, Lakshmanan, 2011a). A data-based approach to social influence maximization. *Proceedings of the VLDB Endowment, 5*(1), 73–84.
Goyal, A., Lu, W., & Lakshmanan, L. V. (Lu, Lakshmanan, 2011b). *Celf + +: Optimizing the greedy algorithm for influence maximization in social networks. Proceedings of the 20th international conference companion on world wide web*. ACM47–48.
Hart, W. E., & Murray, R. (2010). Review of sensor placement strategies for contamination warning systems in drinking water distribution systems. *Journal of Water Resources Planning and Management, 136*(6), 611–619.
Hill, S., Provost, F., Volinsky, C., et al. (2006). Network-based marketing: Identifying likely adopters via consumer networks. *Statistical Science, 21*(2), 256–276.
Huang, J., Peng, M., Wang, H., Cao, J., Gao, W., & Zhang, X. (2017). A probabilistic method for emerging topic tracking in microblog stream. *World Wide Web, 20*(2), 325–350.
Kellerer, H., Pferschy, U., & Pisinger, D. (2004). *Introduction to np-completeness of knapsack problems. Knapsack problems*. Springer483–493.
Kempe, D., Kleinberg, J., & Tardos, É. (2003). *Maximizing the spread of influence through a social network. Proceedings of the 9th ACM SIGKDDinternational conference on knowledge discovery and data mining*. ACM137–146.
Khuller, S., Moss, A., & Naor, J. S. (1999). The budgeted maximum coverage problem. *Information Processing Letters, 70*(1), 39–45.
Kwak, H., Lee, C., Park, H., & Moon, S. (2010). *What is twitter, a social network or a news media? Proceedings of the 19th international conference on world wide web*. ACM591–600.
Laylavi, F., Rajabifard, A., & Kalantari, M. (2017). Event relatedness assessment of twitter messages for emergency response. *Information Processing & Management, 53*(1), 266–280.
Lei, S., Maniu, S., Mo, L., Cheng, R., & Senellart, P. (2015). *Online influence maximization. Proceedings of the 21th ACM SIGKDDinternational conference on knowledge discovery and data mining*. ACM645–654.
Leskovec, J., Krause, A., Guestrin, C., Faloutsos, C., VanBriesen, J., & Glance, N. (2007). *Cost-effective outbreak detection in networks. Proceedings of the 13th ACM SIGKDDinternational conference on knowledge discovery and data mining*. ACM420–429.
Li, C., Sun, A., & Datta, A. (2012). *Twevent: Segment-based event detection from tweets. Proceedings of the 21st ACM international conference on information and knowledge managementCIKM '12*155–164.
Liu, G., Zhang, J., & Chen, G. (2014). An approach to finding the cost-effective immunization targets for information assurance. *Decision Support Systems, 67*, 40–52.
Liu, Y., Liu, X., Gao, S., Gong, L., Kang, C., Zhi, Y., et al. (2015). Social sensing: A new approach to understanding our socioeconomic environments. *Annals of the Association of American Geographers, 105*(3), 512–530.
Madan, A., Cebrian, M., Lazer, D., & Pentland, A. (2010). *Social sensing for epidemiological behavior change. Proceedings of the 12th ACM international conference on ubiquitous computing*. ACM291–300.
Mahmoody, A., Riondato, M., & Upfal, E. (2016). *Wiggins: Detecting valuable information in dynamic networks using limited resources. Proceedings of the 9th acm international conference on web search and data mining*. ACM677–686.
Mirzasoleiman, B., Badanidiyuru, A., Karbasi, A., Vondrák, J., & Krause, A. (2015). *Lazier than lazy greedy. Aaai*1812–1818.
Nemhauser, G. L., Wolsey, L. A., & Fisher, M. L. (1978). An analysis of approximations for maximizing submodular set functionsi. *Mathematical Programming, 14*(1), 265–294.
Ostfeld, A., Uber, J. G., Salomons, E., Berry, J. W., Hart, W. E., Phillips, C. A., et al. (2008). The battle of the water sensor networks (BWSN): A design challenge for engineers and algorithms. *Journal of Water Resources Planning and Management, 134*(6), 556–568.
Pei, S., Muchnik, L., Andrade Jr, J. S., Zheng, Z., & Makse, H. A. (2014). Searching for superspreaders of information in real-world social media. *Scientific Reports, 4*, 5547.
Petrović, S., Osborne, M., & Lavrenko, V. (2010). *Streaming first story detection with application to twitter. Human language technologies: The 2010 annual conference of the north American chapter of the association for computational linguisticsHLT '10*Stroudsburg, PA, USA: Association for Computational Linguistics181–189.
Sakaki, T., Okazaki, M., & Matsuo, Y. (2010). *Earthquake shakes twitter users: Real-time event detection by social sensors. Proceedings of the 19th international conference on world wide web*. ACM851–860.
Salathé, M., Kazandjieva, M., Lee, J. W., Levis, P., Feldman, M. W., & Jones, J. H. (2010). A high-resolution human contact network for infectious disease transmission. *Proceedings of the National Academy of Sciences*.
Schubert, E., Weiler, M., & Kriegel, H.-P. (2014). *Signitrend: Scalable detection of emerging topics in textual streams by hashed significance thresholds. Proceedings of the 20th ACM SIGKDDinternational conference on knowledge discovery and data mining*. ACM871–880.
Takahashi, T., Tomioka, R., & Yamanishi, K. (2011). *Discovering emerging topics in social streams via link anomaly detection. Data mining (ICDM), 2011 IEEE 11th international conference on*. IEEE1230–1235.
Tang, Y., Xiao, X., & Shi, Y. (2014). *Influence maximization: Near-optimal time complexity meets practical efficiency. Proceedings of the 2014 ACMSIGMODinternational conference on management of dataSIGMOD '14*New York, NY, USA: ACM75–86.
Tu, Y.-N., & Seng, J.-L. (2012). Indices of novelty for emerging topic detection. *Information processing & management, 48*(2), 303–325.
Weng, J., & Lee, B.-S. (2011). Event detection in twitter. *ICWSM, 11*, 401–408.
Xie, W., Zhu, F., Jiang, J., Lim, E., & Wang, K. (2016). Topicsketch: Real-time bursty topic detection from twitter. *IEEE Transactions on Knowledge and Data Engineering, 28*(8), 2216–2229.
Yan, X., Guo, J., Lan, Y., Xu, J., & Cheng, X. (2015). *A probabilistic model for bursty topic discovery in microblogs. Proceedings of the twenty-ninth AAAIconference on artificial intelligenceAAAI'15*353–359.
Zhang, J., Wang, Y., & Vassileva, J. (2013). Socconnect: A personalized social network aggregator and recommender. *Information Processing & Management, 49*(3), 721–737.
Zhao, J., Lui, J. C., Towsley, D., & Guan, X. (2014). Whom to follow: Efficient followee selection for cascading outbreak detection on online social networks. *Computer Networks, 75*, 544–559.