2014 IEEE 15th International Conference on Mobile Data Management

Using Location-based Social Networks for Time-Constrained Information Dissemination

Iouliana Litou, Ioannis Boutsis, Vana Kalogeraki
Department of Informatics
Athens University of Economics and Business
Athens, Greece
{litou, mpoutsis, vana}@aueb.gr

***Abstract*—Location-based social networks have evolved into powerful tools in recent years. The ability to embed location information in Social Networks such as Facebook, Foursquare and Twitter creates exciting opportunities for users to disseminate and exchange geolocation information in a variety of domains. The problem of exploiting the social ties between the users for maximizing information reach has become a topic of great interest, and many challenges have to be met. In this work we study the problem of efficient information dissemination in location-based social networks under time constraints. The objective is to identify a subset of individuals to propagate the information and make intelligent route selection that can result in maximizing the reach within a time window. Our detailed experimental results illustrate the feasibility and performance of our approach.**

I. Introduction

Recently, we have observed the explosive growth of social networks such as Facebook, Twitter and Google+, that enumerate large amount of subscribers. For instance, Facebook has reached over 900 million active users, while Twitter follows with over 550 million users and Google+ with more than 170 million users[1]. These networks have been utilized as major tools for the spread of ideas, information and notifications among their members. Studies reveal that social networks can be exploited not only for "viral marketing" [1] (i.e. promote products to targeted sets of users that further propagate them through the word-of-mouth effect to reach a larger audience), but also for discovering emergent topics [2] and for emergency events alerting, management and public safety [3]. For example, people located in the vicinity of earthquakes share via Twitter, a well known social service for exchanging short text messages, anecdotal information related to the dissemination of seismically activity, that earthquake alerts lag behind firsthand notification [4], [5]. Studies reveal that depending on the size and location of the earthquake, scientific alerts can take between 2 to 20 minutes to publish, while using Twitter's notification capabilities people were notified about the occurrence of the earthquakes shaking within seconds of their occurrences.

Thus, social networks (*i.e.,* Twitter, Facebook, LinkedIn) can play a major role in effective emergency notification due to their ability to (1) reach millions of users, especially family and friends, (2) become alternative communication mediums when the wireless and telecommunication networks are congested during emergencies, and (3) provide cost-effective solutions since they are able to reach great amounts of social users without additional infrastructure costs. Furthermore, the study of social relationships and interactions in social networks may provide important insights for gathering information and planning evacuations during rescue efforts. However, adopting location-based social networks as an effective communication medium for emergency alerting raises considerable challenges. Challenges lie in the level of availability and responsiveness expected from these infrastructures in delivering notifications under *time constraints* to reach all recipients interested in receiving the information (these can be people located in the area of the event *i.e.,* students in a campus, as well as their relatives and friends).

In this paper we study the problem of using location-based social networks for efficient dissemination of information under time constraints. Specifically, we examine how efficiently a location-based social network, such as Twitter, can be deployed for emergency notification. Twitter has the ability to broadcast and forward messages to users and is primarily used via mobile devices[2] so that users can be informed at anytime, anywhere as long as they can access the network. Our objective is stated as follows: Given a location-based social network comprising a number of mobile users, the social relationships among users, the set of recipients and the timeliness requirements, our goal is to select an appropriate subset of users to propagate the information such that (1) the expected spread of information is maximized, (2) time constraints in the dissemination of the information are satisfied and (3) costs are considered. Cost is defined as the amount of messages that need to be exchanged among users. Thus, it could be either monetary (for an SMS) or resource allocation cost. Our primary focus is on information that needs to be propagated under time constraints, such as emergency information, where the notification about the event needs to be propagated under strict time constraints.

We approach the problem in two phases. We first use a crawling phase where user profiles are built, social relationships are inferred and effective dissemination paths among the users of the social network are computed. In the second phase, namely reaction phase, we aim at reducing the search space by considering only users in the social network that are in the proximity of the interested to the event users (i.e. users related to the event) since the event may not be of interest to all users of the network. Then, we select a small number of seed users that will allow us to efficiently disseminate the emergency

[1]http://www.go-gulf.com/social-networking-users.jpg

[2]https://blog.twitter.com/2013/new-compete-study-primary-mobile-users-on-twitter

978-1-4799-5705-7/14 $31.00 © 2014 IEEE
DOI 10.1109/MDM.2014.26

information to all interested recipients during the emergency event. Reduction of costs is accomplished by avoiding push-based broadcasts, which is important in emergency events as communication is typically over-utilized in such scenarios [5].

Existing information dissemination techniques are not adequate to solve these problems. The problem of maximizing the spread of influence in social networks has been addressed in [6], [7], [8], [9], but none of these works consider time constraints. Only recent efforts recognize that time plays an important role in the influence spread [1], [10]. However, contrary to our approach, these efforts assume that the influence flow is known and aim at maximizing the influence in the entire network rather than identifying and informing an appropriate subset of nodes that would be most interested in the information. Furthermore, both works study cases of viral marketing campaigns or voting systems, rather than emergency response situations that have to operate under tight time deadlines and resource savings. Emergency response outside social networks has also been studied. The use of geographical notification systems has been considered in [11]. The purpose of the presented system is to construct overlays that support location-based regional multicasting where they also consider issues of providing reliable storage of social information under extreme regional conditions. Traditional approaches such as multicast [12], [13] and publish/subscribe systems [14] are not appropriate for our setting since they will inform only subscribed users, while we need to alert all users associated with the emergency event, that is, all users that would be interested in the event and not just subscribed users. Furthermore, the set of users to be informed by our system is determined based not only on locational criteria, but also on relationship criteria, so it is not considered to be a strictly location-based approach. On the other hand, approaches like flooding and gossiping [15], [16], [17] will inform most of the users interested in the event, but they will also produce a lot of spamming to the other users and load to the network. Contrary to our work, none of these approaches takes real-time constraints into account and no social networks are exploited in the dissemination process.

Our paper makes the following contributions:

- We present LATITuDE (using Location based sociAl networks for Time constrained InformaTion DissEmination), a system that solves the problem of efficient dissemination of emergency information in location based social networks under time constraints. The problem is NP-hard [18] and we provide a greedy algorithm with an approximation rate of 1-1/e.
- We extend our approach so that we consider the authoritativeness of a user in the spread of information.
- We perform extensive experiments to validate our approach. Our experimental results illustrate that our approach is practical and effectively addresses the problem of informing a large amount of users with the least messages within a deadline when an emergency event occurs, and outperforms its competitors.
- Furthermore, we verify our solution using a real case scenario of emergency event by deploying it on tweets concerning Sandy Hurricane, a major emergency event that occurred in 2012 [19].

II. Model and Problem Definition

In this section, we present our system model and formulate the NP-complete problem addressed in this paper.

A. Model

A social network is typically presented as a weighted social graph $G = (V, E, W)$, in which each vertex $u \in V$ corresponds to a user in the social network, each edge $e_{uv} \in E$ represents a social relation formed among users (u, v), and the weight $w_{uv} \in [0, 1]$ reflects the "strength" of the relationship among the users (u, v), based on their social interactions.

Each user $u \in V$ that connects to the network is associated with a unique id and is used each time the user logs into the system. This for example might be the user's id in the social network. For our Twitter dataset we consider the screen name of the user as id. In Twitter, the screen name follows immediately after the '@' character in the body of the text (i.e. tweet) whenever a user addresses a specific user in the network. For example *"@NYTimes: Scientists have warned of #NYC storm peril; rising sea levels, extreme weather http://ht.ly/eUwqS #Sandy #Climate"* is a tweet addressed to The New York Times, where NYTimes is the screen name of "The New York Times" account in Twitter.

We assume that user relationships take place and evolve over the lifespan of the network (in order to instantiate an edge among a pair of users the corresponding users should be socially connected and have interacted). For each user, we keep track of his interactions. Whenever a user u posts a message referring to any other user v in the network, the list of interactions I_u of user u is updated. The list of interactions consists of tuples of the form $< v, m_v, timestamp_v, time_v >$, where v is the receiver of the message m_v, $timestamp_v$ denotes the unix timestamp when the message was sent and $time_v$ denotes the time used by u to process and deliver the message to user v, that involves the amount of time to process and disseminate a message to other users, if further propagated. Here u and v correspond to user ids, but to keep it simple we use this notation for presenting a user. The data of the interactions list are retrieved and processed for determining the weights of the social relationships among users in the graph.

Several mechanisms exist for assigning weights to edges as stated in the related work section [20], [21], [22]. Our model is generic enough and can be extended so that several metrics can be considered. Yet, finding the most appropriate metric is not subject of this study so we consider a simple but popular metric based on the frequency of communication.

Frequency of communication (FC). It is usual that within a social network users that are closely related interact more frequently. Thus, the strength of the tie between users can be measured using the frequency of communication. We define frequency of communication between a pair of users (u, v) as follows:

$$f_{uv} = \sum |m_v| / \sum_{i \in I_u} \sum |m_i| \tag{1}$$

where $\sum |m_v|$ denotes the total amount of messages in the interaction list I_u of user u referring to user v and $\sum |m_i|$ denotes the total amount of messages in I_u referring to any other user i.

Thus, f_{uv} expresses the amount of messages sent by user u to v out of the total messages sent by user u to any other user i in the network. Note that f_{uv} differs from f_{vu}, since Twitter presents large asymmetry in the relations due to broadcasters and miscreants [23]. That is, there are users with a high number of followers but few followees (broadcasters) and the opposite (miscreants). We normalize the weight for all users, by computing the weight $w_{uv} \in W$ of the edge $e_{uv} \in E$ as:

$$w_{uv} = f_{uv}/max\{f_{uv'} : v' \in neighbor(u)\} \quad (2)$$

The above metric is used for automatically computing the "strength" of the social tie among users, without user participation. The weight of the edges takes values in the range of $(0, 1]$, where a value of 1 denotes a strong relationship between the users. As stated earlier, other metrics can also be defined. For example, in emergency situations, users may be given the chance to define a list of "emergency contacts" or the system may denote some nodes as "authorities" that should always be given higher priority in the dissemination process. For example, for users v that constitute authorities, the weight of the edge $e_{uv} \in E$ equals 1, for all nodes u that have v in their interaction list I_u, regardless of the metric (further discussed in Section IV-C).

In our system we differentiate among the following roles for the users in the location-based social network, based on the way they are related to the emergency event (as shown in figure 1): (a) *Interested nodes* denote users that are interested in the occurrence of the event; these are subset of nodes in the social graph that are either known to be geographically in the proximity of the event or have strong social ties with users related to the event. They constitute the set of nodes our system aims reaching through the dissemination process. (b) *Reachable nodes* are the set of nodes that are accessible immediately after the occurrence of the event, *i.e.* these are connected to the network and thus information can be directly delivered to them to further disseminate it. Reachable nodes are not necessarily a subset of the interested, but we assume that these are in the proximity of interested nodes on the social graph or some type of authority (discussed in Section IV-C). For instance, during the Sandy Hurricane that took place in October 2012 and severely affected New York City [19], all citizens needed to know about the event. Thus, the *interested set* consists of all New York citizens and their relatives that may not be presented to New York and also users that were about to visit New York. When the event took place, not all users had access to the Internet and social networks. The users with active mobile devices that were connected to the network constitute the *reachable nodes*. Among these users, the system needs to identify the most efficient users to propagate the information needed to the interested users and these constitute the *seed nodes*.

B. Problem Definition

Given a weighed social graph G, a subset of vertices $S \subset V$ that constitute the *interested nodes*, a subset of vertices $R \subset V$ that constitutes the *reachable nodes* a time bound $Deadline$ and an integer k, so that $k < |V|$, our goal is to extract a set of nodes $M \subset R$, denoted as seed set, to maximize the expected number of interested nodes in S that will be informed by nodes in M before the $Deadline$, under the condition $|M| \leq k$.

Thus, our problem is to define the maximum amount of nodes from S that will be informed within the $Deadline$, provided that the amount of k seeds cannot be exceeded. We also note, that not all nodes are online in our system. Thus, our goal is to maximize the spread of the information, from the k nodes selected to disseminate the information, under the restriction that nodes selected are reachable.

Traditional influence maximization problems [9], differ from our formulated problem in two ways. First, traditional models ignore any time delays in propagating messages among users in a social network. The propagation delay corresponds to the time to propagate a message along multiple edges between users, (*i.e.,* through the wireless or wired communication medium) and includes the time to receive, process and selectively propagate the messages. In traditional models any propagation delay is thought to be constant or simply ignored. Contrary to traditional models, we associate each edge with a *latency*, which denotes that if a message was sent at time t from node u along the edge e_{uv}, and δ is the estimated delay to forward the message across the edge, then we expect that the message will reach node v at time instance $t + \delta$.

A second important difference, is that, we aim at propagating the information to a subset of the nodes in the location-based social network, S, which are affected by the event, referred as *interested nodes*. Hence, our goal is to maximize the amount of *interested nodes* that will be informed, instead of maximizing the amount of informed nodes on the whole graph. A fundamental challenge in our setting is that the nodes' reachability derives from their physical connectivity. That introduces additional constraints on the availability of the nodes to act as seeds, since only a subset R of nodes can be accessed and seeds must be selected among those.

Similarly to the Independent Cascade (IC) Model [9], there exists only one chance that a node will forward the message to its neighbors, after it is informed.

The traditional influence maximization problem, without considering time constraints, has been shown to be NP-complete [9], [18]. Considering time deadlines imposes additional complexity to the problem. Since the problem is NP-complete we develop an approximation algorithm to solve the problem efficiently, within the defined time constraints.

III. THE LATITUDE SYSTEM

In this section we provide an overview of our LATITuDE system. The system is implemented with the following five main components that work in concert: i) a Profiling Component, ii) a Social Graph Component iii) a Latency Estimation Component, iv) a Path Generator Component and v) a Dynamic Notification Component.

The Profiling component is responsible for maintaining data about user interaction in the location-based social network. It uses raw data collected by the network to build user profiles and extract statistics. It is responsible for updating user's interaction lists (discussed in II-A). These data are later used by the Social Graph Component for computing the weight of the edges on the Social Network. We consider the graph given and only aim at computing the strength of the relationships formed between users. Data selected and

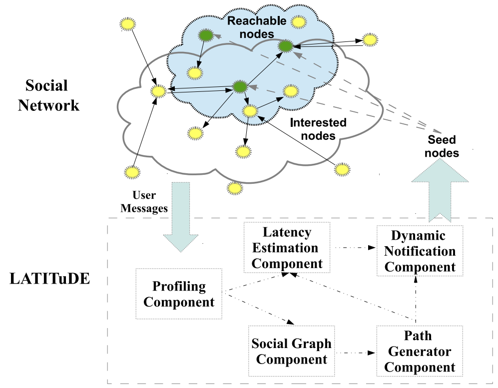

Fig. 1. The LATITuDE Architecture.

computed by the Profiling component are also used for the latency estimation in the delivery of the messages.

The Path Generator Component computes for each node $u \in V$ in the social graph, the effective paths beginning at node u. The effective paths for a node u represent the likelihood that messages generated or messages that reach node u, will be propagated along those paths. In the simple case, all paths can be examined to decide whether a message will be effectively disseminated along them. However, to reduce the search space, we consider only those paths whose probability exceeds a tunable threshold ϵ, and whose length is at most L hops. The Latency Estimation Component implements a probabilistic model to estimate the latency of the messages disseminated along the path. This is used to identify the probability of a message being delivered among two nodes in the social Graph before a predefined $Deadline$. The $Deadline$ is a relative value, and denotes a time interval starting from the occurrence of an event, within which all interested users must be notified. All the above components are deployed in the crawling phase.

Finally, the Dynamic Notification Component is triggered in the reaction phase when an emergency event occurs. This component utilizes the information stored and maintained by the previously discussed components to identify the initial receivers of the message (*i.e.* seeds) so as to maximize the spread of information among interested users within the $Deadline$, while using the least amount of messages.

The overall architecture and the interaction between different components is depicted in Figure 1 and the corresponding functionalities are described in the following sections.

IV. Efficient Dissemination of Emergency Information

A. Crawling Phase

During the crawling phase, the following pieces of information are precomputed: the effective paths of the nodes, and the latencies along the paths are estimated. Below we give a detailed description of their functionality.

1) Computing Effective Paths: To reduce the complexity of the Social Graph component during the reaction phase, we precompute and store the most effective paths for each node $u \in V$. The set of effective paths contains all paths whose probability exceeds a threshold ϵ, and their length is at most L edges (in our experiments we choose a length of at most 7 hops based on the findings of the small-world experiment [24] where they show that a message traverses an average of up to 5.2 links in the network, note though that choosing shortest paths has been proved equally effective). When a path's probability is below ϵ, this path is not further considered. This process is performed by the Path Generator Component.

The paths are explored using a Depth First Search (DFS) approach. However, for each node $u \in V$, we do not keep track of the whole path to any other node. We only keep the destination node $v \in V$ and the path probability to that node. Since there may be multiple paths between a pair of nodes (u, v), only the most optimistic path probability is maintained. So, for each node $u \in V$ we compute a set of nodes $v \in V$ towards which u has effective paths, along with the efficiency of the path. The path probability is calculated as:

$$p(u, v) = \Pi_{u',v' \in path(u,v)} w(u', v'), \in [0, 1] \tag{3}$$

where $w(u', v')$ denotes the weight of edge u', v' in the original graph $G(V, E, W)$. Essentially, this metric denotes the probability that the message generated at node u is delivered along the entire path to reach node v, computed by multiplying the probabilities of each individual edge in the path being traversed. We choose this path probability metric in our experimental evaluation, similarly to [8].

2) Estimating the latency of a path: In this step, we use the information stored in the Profiling Component and the effective paths computed in the Path Generator Component to estimate the delivery delay along these paths. For all effective paths among two nodes in the location-based social network we use our probabilistic model to estimate the probability of time constraints being met.

In order to select the most appropriate paths the next step is to estimate whether the message will be successfully delivered along these paths given the real-time requirements. Thus, for each path we determine the probability that the message will be received on time. It has been shown in [25] that in location-based social networks like Twitter the delivery time of a message follows the Power Law Distribution [26]. That means that the delivery times $time_v$ among two users should cluster around a typical value. Hence, we compute the probability $p_{u \to v}(Delivery_{u \to v} < Deadline)$, using the CDF of the Power Law Distribution. We compute the $Delivery_{u \to v}$ based on the past user behavior, i.e. previous message exchanges among users, provided by the Profiling Component. Essentially, the delivery metric is determined from the motifs of communication among users.

A quantity x obeys Power law when it is drawn from a probability distribution $p(q) \propto q^{-\alpha}$. The α parameter, known as the exponent or scaling parameter, is a constant parameter. To define the probability mentioned above we need to use the CDF of the Power Law distributed variable, denoted as $P(q)$ (with q being a variable), which is defined as $P(q) = Pr(Q \geq q)$, where Q is the observed value. Thus, $P(q) = \int_q^\infty p(q')dq' = (\frac{q}{q_{min}})^{-\alpha+1}$, with $q_{min} > 0$ being a lower bound on the Power Law behavior. Finally, α is computed as: $\alpha = 1 + n[\sum_{i=1}^{n} ln \frac{q_i}{q_{min} - 1/2}]^{-1}$, where n denotes the amount of q_i values.

In our approach, the q_i values represent the message delivery times among two users u and v. Thus, we use the information stored in the user's u profile to obtain the observed delivery times $time_v$ to user v denoted as $Delivery_{u \to v}$, in order to compute the probability $p_{u \to v}(Deadline)$ and thus the probability that the delivery time will exceed the Deadline value. The lower bound q_{min} is set as the minimum measured delivery time for a message among u and v. Since the path might follow more than one hops the observed values should represent the summation of all combinations of observed times in the specific path. For instance, for the path $a \to b \to c$ the set of q_i values will be generated as follows: $\{(time_{a \to b} + time_{b \to c}) | \forall a \to b, \forall b \to c\}$, retrieved from the Profiling Component. The probability is computed as:

$$p_{u \to v}(Delivery_{u \to v} < Deadline) = 1 - (p_{u \to v}(Deadline)) \tag{4}$$

Hence, we estimate the probability of a specific path to effectively deliver a message before the deadline, from its starting node u to the destination node v.

Though Power Law Distribution is chosen, since it is the most common distribution characterizing the messages delivery times over social networks, any other distribution may be used to extract probabilities of timely delivery. Note that during emergency events, users tend to use the network more frequently and dissemination is expanded rapidly [5]. Our predictions underestimate this factor, which could make our model more efficient, by assuming that dissemination distribution during the event will not change drastically.

B. Reaction phase

The basic functions of the reaction phase is to locate the interested and the reachable users on the social graph and select among the reachable an initial set of nodes (*seeds*) that have the most effective paths with the interested users, so as to increase chances of the message being effectively delivered during the propagation. This process is executed by the Dynamic Notification Component of the system. Note that our System cannot affect the dissemination process, except from choosing the appropriate seeds. Thereafter its up to the users to further propagate the information.

We consider the process of locating interested users quite trivial, since they can be detected for instance based on their latest geographical location or by given lists (e.g. students of a University Campus). Detecting the reachable users is also trivial (e.g. users that are online), so we will not further discuss these processes and consider them given.

1) Seed Selection: Given a number of candidate effective paths and the users that our system aims at reaching (*i.e.* all users interested in the event), our goal is to select the seeds so as to maximize the information dissemination to the interested users when the event occurs .

Let R be the set of reachable nodes (discussed in Section II-A). We compute the efficiency of the candidate effective paths, that begin with a reachable node $u \in R$ and end with a targeted interested node $v \in S$ as follows:

$$ef(u,v) = w_1 p(u,v) + w_2 p_{u \to v}(Delivery_{u \to v} < Deadline) \tag{5}$$

In this paper we consider the same weight for the path probability and the delivery probability as $w_1 = w_2 = \frac{1}{2}$ (*i.e.,* we consider equally important that the message is timely received and successfully reaches the receiver), however, these parameters are tunable and can be adjusted at run-time based on the goals of LATITuDE, to either maximize the information spread or to deliver more messages within the deadline. After computing the efficiencies of the candidate effective paths, the greedy algorithm starts to choose the set of Seed nodes M. The greedy process for the seed selection is described below.

Greedy Node Selection: Consider a set of seeds M, a set of nodes A that consist of the nodes that we expect to be informed by M with high probability, thus $M \subseteq A$, the set of interested nodes S, the set of reachable nodes R, and a candidate seed $u \in R$ to be added in M. The greedy step for the selection of u is:

$$M \cup u \ s.t. \ \sigma(u) = max\{\sigma(u) : \forall u \in R \setminus A\} \tag{6}$$

$$where \quad \sigma(u) = \frac{(\sum_{v \in S \setminus A} ef(u,v))^2}{|Q \setminus A|}$$

In the above equation $Q \subseteq S$ and $\forall v \in Q$ there exist an effective path $path(u,v)$ starting by u.

Intuitively, $\sigma(u)$ computes the number of interested nodes $v \in S$ informed when u is selected as the root of the dissemination process, while taking into account the average probability that those nodes are effectively informed. Nodes that are possibly informed by previous seeds selected, *i.e.,* nodes in A, are not added to the efficiency of u. Thus, the algorithm may produce a seed set that has multiple seeds reaching the same node in S, *i.e.,* consider nodes in $u_1, u_2 \in M$ that both have paths to node $v \in S$. This case is not undesirable since the chance of v being informed is increased.

Algorithm Description: Initially the seed set is empty, $M = \emptyset$. In the first iteration, the efficiencies of all reachable nodes $u \in R$ are computed. Only paths ending to a node $v \in S$ are considered in the computation of the efficiency of u. When all efficiencies are computed the best node is added to the seed set and all possibly informed nodes $v \in A$ by this node are identified. The set A is computed using Monte-Carlo simulations while considering as the only seed the node u, selected to be added at the seed set M. We only add nodes that are informed with high probability θ as derived by the simulations (in our experiments we set θ to 0.9 and 10000 iterations are set to be executed for the Monte-Carlo Simulations). In the next iteration, efficiencies for the remaining reachable nodes are computed. Interested nodes in A are not added to the efficiency of a node, and nodes in A are not candidate seeds. The process is repeated until all k seeds are selected, or no more new seeds can be added in the set.

Monotonicity, Submodularity and non-negativity. The function suggested for the seed selection process is monotone, submodular and by-definition non negative. So our greedy algorithm, according to [27] achieves an approximation rate of 1-1/e. Let $F(M)$ denote the set of interested users that will be informed by the seed M. We have $F(M) \leq F(M \cup u)$ since a node u as selected by $\sigma(u)$ adds interested nodes to the set of informed that are not already informed by M, that is u will add nodes to $F(M)$. So our function is monotone. Our function is also submodular. Let M_1 and

M_2 denote different set of seeds with $M_1 \subseteq M_2$. We have $F(M_1 \cup u) - F(M_1) \geq F(M_2 \cup u) - F(M_2)$, since in $\sigma(u)$ function when a node is added its efficiency over previously informed nodes is ignored but exists. So if $M_1 \subseteq M_2$ u added to M_1 adds more value to $F(M)$, that is greater amount of previously non-informed interested users is added to $F(M)$. Thus, our greedy algorithm has an approximation rate of 1-1/e.

Worst Case Complexity of Seed Selection. Since our goal is to identify k seeds, we have at most k iterations to select them, with one seed being selected in each iteration. In every iteration we compute the most efficient node. Thus, for each node $u \in R$, we need to compute $ef(u, v)$ (that costs $O(1)$) with every node $v \in S$, that is $O(|R||S|)$, if every node in R can reach every other node in S. Determining the node with the maximum value costs only $O(1)$, since we store the best node as we traverse through nodes in R. For the Monte-Carlo simulation to determine the nodes that will be informed by the selected seed node, we have to iterate through the S nodes and define if the nodes are informed, that costs $O(1)$, so the complexity of this step is $O(|S|)$. Thus, the worst-case complexity of our algorithm is $O(k|R||S|)$.

2) Dissemination of Information.: Assuming a selected set of seed nodes M, our model propagates the emergency information as described below. We denote A_t as the set of nodes informed at step t. At the beginning, only the seed nodes are informed and thus $A_0 = M$. At step $t + 1$, every node $u \in A_t$ can inform each of its currently uninformed neighbors v and the probability of node u to inform v is given by the weight of the edge (u, v), w_{uv}. Each node has a single chance of informing its currently uninformed neighbors about the event in step $t+\delta$, where δ represents the delay of a user to forward a message as estimated from the Profiling Component.

C. Authorities as seeds

During major emergency events, there are often principal emergency authorities, such as city officials, emergency personnel etc., that are given higher priority in the alerting and notification process. Authorities often use social media for issuing emergency alerts or spreading emergency information[3]. Thus, we also present an alternative approach, where certain nodes are given higher priority in the seed selection process according to their importance or reliability.

Authorities are considered more reliable and thus user may be less reluctant in further propagating their messages. Thus they act as better seeds. To incorporate them in our algorithm we consider that among the k seeds selected, a part of it will constitute authorities and they are added to the set regardless of their score estimation at the greedy step. Thus we initially select all authorities and then seed selection as described in IV-B follows for the remaining seeds. Let T be the set authorities. The seed set M is now formed as: $\forall u \in T, M \cup u$, and then the greedy step takes place until all k seeds are added to M or no more seeds can be added.

Since we have no designated users with distinct roles in our dataset, we define as authorities nodes that have a high amount of followers compared to the ones they follow. Note that there might be other ways to define authorities, but this is not subject of this study. So for our experiments we choose a fixed number of authorities (that is set to 5% of all seeds) and we add nodes as described below:

$$T \cup u \ s.t. \ d(u) = max\{d(u) : \forall u \in V\} \tag{7}$$
$$where \quad d(u) = inDegree(u) - outDegree(u)$$

Nodes are added to T until the desired number of authorities is satisfied. Then, we add to the seed set the nodes that were selected as authorities and out of the k nodes that we wish to use as seeds we select the $k - |T|$ nodes to be added in the seed set using our algorithm. Unlike previously, the algorithm has an initial set of seeds M that consists of the nodes in T and is non-empty and the set A consists of all users possibly informed by T. The algorithm then selects a seed set of $k-|T|$ seeds or less if no more nodes can be added, and thus the final seed set consists of at most k nodes in total (authorities and users selected as seeds). The dissemination process information varies in how the authorities inform their users. In the non-authoritative seeds, the communication is personal, and thus follows the out-degrees of the seeds, *i.e.,* user $u \in M$ directly communicates with its out neighbor v. Authorities do not follow personal communication flow, but a rather broadcast form of communication. So for the authorities, all followers are informed, and so, the flow of communication follows the in-degrees. The authorities are also promoted for efficiently delivering the information timely, that is they have higher chances of timely informing their followers.

V. Experimental Evaluation

We have implemented our LATITuDE system and tested it with two real-world Twitter datasets. The first dataset consists of tweets related to the area of Dublin without any specific topic of interest. The second dataset consists of tweets referring to the Sandy Hurricane.

The experimental evaluation focuses on: (i) Execution time compared to state-of-the-art techniques, (ii) Number of informed users among users that are interested in receiving alerts, (iii) Number of informed users before the Deadline expires, (iv) Scalability based on the amount of seed nodes selected, and (v) Performance under different deadlines.

Note that, proving the efficacy of our metric for computing the edges weight, is out of the scope of our study. Both state-of-the-art algorithms and our LATITuDE system, are given the same social graph as input. Thus, we argue that given any other weighted social graph and past interactions between users, LATITuDE would perform equally well compare to any other influence maximization algorithm.

A. Evaluation using Twitter data

First we consider a dataset that is composed of 513.449 tweets posted by 175.974 unique users in the city of Dublin, crawled from Dec 2012 to Mar 2013. The tweets were collected by the Twitter Streaming API[4] and were filtered so that only tweets related to the area of Dublin are contained in the dataset. The filter is applied: (i) to the "Location" field

[3]http://www.w3.org/2007/06/eGov-dc/papers/ElectronicgovernmentLibaryofCongress

[4]https://dev.twitter.com/docs/streaming-apis

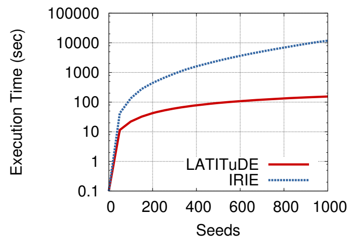

Fig. 2. Dublin Execution Times with 10%Reachable

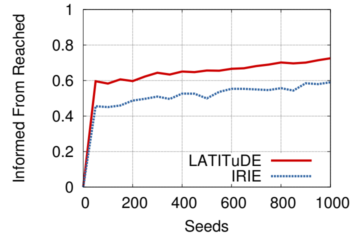

Fig. 3. Dublin Informed From Accessible with 10% Reachable

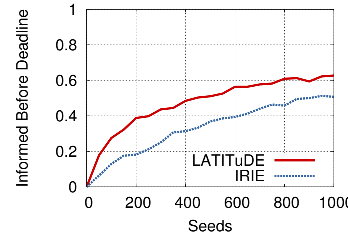

Fig. 4. Dublin Informed Before Deadline with 10% Reachable

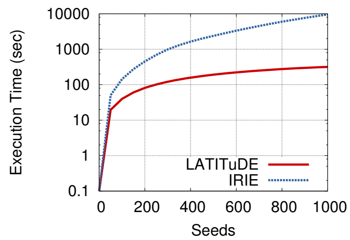

Fig. 5. Dublin Execution Times with 20% Reachable

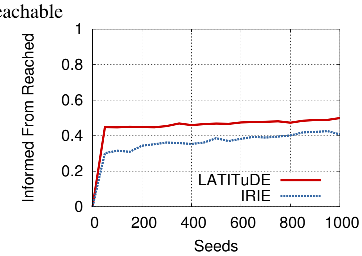

Fig. 6. Dublin Informed From Accessible with 20% Reachable

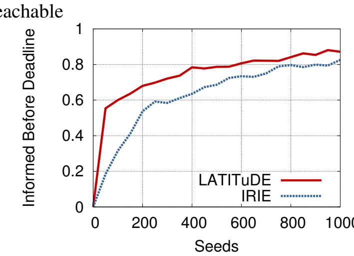

Fig. 7. Dublin Informed Before Deadline with 20% Reachable

exported from the user account, (ii) to the GPS coordinates included in the tweet or (iii) at parts of the text of the tweet[28]. After the above tweets are extracted, all tweets from user accounts presented in the dataset are gathered using the REST API[5] and users' ID. Tweets in our dataset have the following structure: <*Tweet ID, User ID, UTC/GMT timestamp, Latitude, Longitude, ID of tweet replying, ID of user replying, Number of retweets, Source (Andoid, iPad), Text*>. User ID is used to obtain the screen name of the user, the UTC/GMT timestamps are used to estimate the latency for the tweet to be propagate along edges and the Text is used to extract the users mentions with the '@' symbol. No further information related to the users is used. The latency is estimated based on users regularity of communication (e.g. user u contacts user v twice a day). The higher the regularity is, the lower the latency is estimated. We do not consider any anonymization issues [29] since we assume that the system is used by a trusted agency.

1) Parameter Setting: Our threshold parameter ϵ for the efficient paths was set to $\frac{1}{320}$. We choose this in concert with the datasets used in [8], [1]. Since our dataset did not contain delivery times, for the experiments we have assigned the $Delivery_{u->v}$ times based on the regularity of communication between users. Thus, each user is assigned a random lowest delivery time t_u that ranges in $[5, 10]$ and the $Delivery_{u->v}$ is computed as t_u/r_{uv} where $r_{uv} = 1/\log(d_{uv}+1)$ with $(d_{uv}+1)$ denoting the average time lapse (e.g. days) between u and v communication. Hence, we expect that a user that communicates very frequently with another user is going to inform him/her quickly when an incident occurs, compared to a non-frequent contact. Nevertheless, the selection of the delivery times is orthogonal to our proposed methods.

2) LATITuDE Evaluation over IRIE: In this section we present the performance of our approach and we compare it with the state-of-the-art algorithm IRIE [30], which is the fastest algorithm in the literature that we know of and is able to perform influence maximization equally effectively to its competitors. IRIE is a system based on Influence Ranking and Influence Estimation. It has a greedy step in which the node with the largest marginal spread is added to the seed set. The spread is computed based on Monte-Carlo simulations. After a seed is added its additional influence is estimated.

[5]https://dev.twitter.com/docs/api

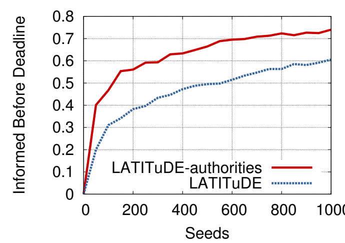

Fig. 8. Dublin Informed Before Deadline with 20% Reachable

In the first set of experiments we present our approach for the Dublin Dataset when we set 10% of the users as reachable (2392) out of the interested set (this corresponds to a set of 23913 nodes). We have set a deadline of 15 seconds to send the messages to all the interested nodes. Figure 2 shows the execution times of each approach as a function of various amounts of seed nodes. As can be observed, LATITuDE needs a lot less time to execute compared to IRIE that reaches up to 11816 seconds for 1000 seed nodes, when our approach needs only 156 seconds. This is mainly due to our crawling phase in which effective paths are precomputed and it shows that IRIE cannot be used for emergency response systems. In figure 3 we present the percentage of interested nodes that each approach can inform, relative to the users informed if all reachable nodes were selected as seeds. LATITuDE manages to inform more users than IRIE at all times, with a percentage that ranges from 7% to 12%, although our approach needs a lot less time to execute. Finally in figure 4 we illustrate the percentage of the interested nodes that were informed before the deadline expires. As can be observed, LATITuDE informs the greatest amount of users within the deadline, due to the Latency Estimation Component, that considers the delivery times for each path. As the number of seed nodes increases, IRIE reduces the gap from LATITuDE, since more seeds can

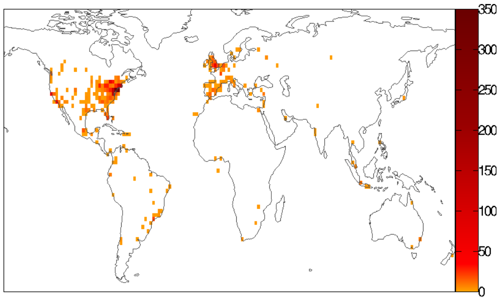

Fig. 9. Sandy Tweets frequent locations

inform more users in fewer hops. However the difference is still more than 10 units per cent for 1000 seeds. It is also important to mention that we start to consider the time interval until the deadline only after the seed selection for both approaches is completed; thus, the gap among LATITuDE and IRIE for the percentage of users being timely informed would be a lot larger if execution times were considered.

In the second set of experiments we consider 20% of the users being reachable (this corresponds to 4784 users) and a deadline of 30 seconds (we have computed that in our dataset 30 seconds is the average time needed to forward a message using our technique, without considering the delay estimation, by setting $ef(u, v) = p(u, v)$). As can be observed in figure 5 the execution times slightly decreases over the reachability of 10% for both techniques but overall they have the same behavior. Figure 6 presents the percentage of interested nodes that were informed and we see that the percentage has been reduced, since we have the same amount of seeds as previously but relatively to 20% reachable nodes instead of 10%, that is if all 20% act as seeds we expect more users to be informed compare to 10%. However, the percentage of users that LATITuDE manages to inform is more than 9 units percent over IRIE at all times. In figure 7 we illustrate the percentage of the interested nodes that were informed within the time constraints. As can be observed more users are informed within deadline, as expected, since we set a laxer deadline.

We note that in figures 4, 7 as the number of seeds increases, the gap between LATITuDE and IRIE is reduced. This is because the intersection of the seed sets selected by LATITuDE and IRIE contains more nodes as the seed set increases. We expect that they would converge as the amount of seed nodes increases. The sharp angle at 50 seeds occurs because we set seeds from 0 to 1000 seeds with step 50.

The additional nodes added to the seed set do not provide any tremendous difference to the nodes being informed and a stabilized performance is observed. This may be due to the sparsity of the graph. Also, when the number of reachable nodes is increased the performance for both algorithms drops, since more users are informed by a broadcast approach, yet LATITuDE manages to inform up to 50% compare to broadcasting with a lot less messages on the network.

We argue that 10% to 20% of reachable users is a realistic proportion since when emergency events occurs not many users have access to the Internet and are connected to a social network service. Yet, there is no study revealing the amount of users that are actually accessible via social networks immediately after an emergency.

Finally, we compare our approach with the alternative of considering the authorities for the seed selection. We have set a deadline of 15 seconds with 20% reachable nodes. When using the authorities in the Dublin Dataset there is no particular difference in the number of interested users being informed and thus we do not provide that figure. However, there is a noticeable difference in the number of interested users being timely informed about the event as shown in figure 8. Thus, although the same amount of users is informed when the authorities are employed at the LATITuDE system, we manage to inform up to 21% more users within the deadline.

B. Evaluation of LATITuDE on Real Case Emergency

Our second dataset consists of tweets referring to the occurrence of Sandy Hurricane and are collected as described in [31]. We used 160704 tweets that correspond to 156490 users and 39100 edges for a time period from October 22 to December 31 of 2012. The dataset does not contain all fields that are needed, so we further crawled Twitter to extract the timestamps when they were published and the GPS locations of the tweets that were tagged with geolocation. Figure 9 shows the locations where tweets concerning the Sandy event were published. As can be seen, the event was of global interest, but most tweets were published in the areas that were directly affected by the occurrence of the Hurricane (i.e. N-E America).

User tweeting behavior for the Sandy Hurricane is presented in figure 10, where it can be observed that few users post a large amount of tweets. Also in our datasets a few users have many followers. These observations are in accordance with [23]. The results are similar for both datasets, and so we can conclude clusters of broadcasters, acquaintances and miscreants exist in our datasets as well and that our datasets are representative of users actual behavior in the network.

1) LATITuDE performance on Sandy Hurricane: We conducted the experiment of comparing the authorities over the simple seed selection process on the Sandy Hurricane dataset. We set the interested users at 19334 nodes which constitutes 1/8 of the users in our dataset. We consider this to be a representative amount since as can be seen in figure 9 the event draws global attention with a large amount of users not being geographically interested in the occurrence of the hurricane. We consider 20% of interested users as reachable nodes (3867) and the deadline was set to 15 seconds.

In figure 11 we present the percentage of interested nodes informed. We set the authorities to be 5% of the seed set. For small seed sets, the simple seed selection manages to inform greater amount of interested users. This is because the selected authorities might be few hopes away of the interested users, thus the simple seed selection may choose better seeds. As the amount of seeds increases, we see that the gap decreases until both algorithms perform equally well. However, as can be seen in figure 12, authorities achieve a higher percentage of timely informed users.

Discussion: From the experimental evaluation we conclude that when we have tighter deadlines, it is better to add authorities to the seed set, while if we aim at maximum reachability with a relaxed deadline, LATITuDE seed set selection is

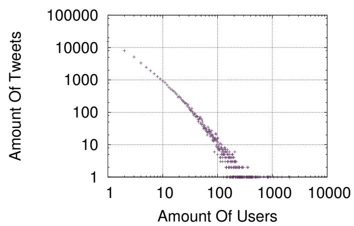

Fig. 10. User tweeting behavior during the Sandy Hurricane

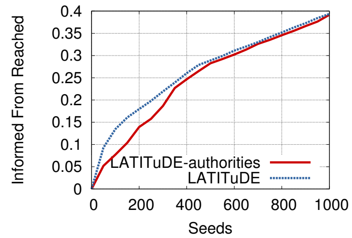

Fig. 11. Sandy Informed From Accessible with 20% Reachable

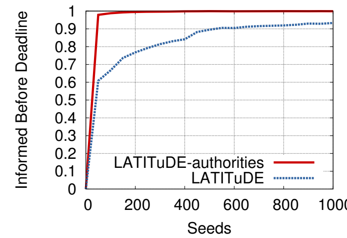

Fig. 12. Sandy Informed Before Deadline with 20% Reachable

preferable. We also state that in the case of emergencies the frequency of communication increases, thus both models perform much better in informing almost every user timely. That proves that our predictions underestimate the factor of increased communication rate (section IV-A), and that leads our model to actually perform better than estimated.

VI. Related Work

Social Networks have attracted interest in exploiting their capabilities in recent years. Recent studies reveal that up to 70% of users have access to social networks via mobile devices[6], providing challenges to explore the abilities that social network offer.

Social network of users is typically depicted as a weighted graph. Nodes of the graph represent users of the network and the edges express relations between pair of users. The weight of the edges captures the strength of the tie, expressing the affect that a user has over his/her fellow or social relationship strength. Inferring the edges' weight has been part of several studies. Edgerank is a metric recently announced by Facebook for determining which items to populate on users news feed. Affinity of users, interactions between users, sort of interaction and time elapsed between last interaction are considered in the metric to determine the strength of the bond and order news feeds [21]. Goyal *et al* [6] consider a number of metrics, including the Weighted Cascade (WC) model, the Trivalency model and learning model. The WC model defines the weight of the edges e_{uv} as $1/in-degree(u)$ for all $v \in neighbor(u)$. The Trivalency model randomly and uniformly peaks values from the set $\{0.1, 0.01, 0.001\}$. In the learning model they use the Expectation Maximization based method suggested by Saito *et al* [32] on a training set to extract edge probabilities. The asymmetry presented in social networks is being captured by the metric defined by Hangal *et al* in [20] where they argue that the strongest paths in global social search may be better that the shortest, so they aim in improving global social search by exploiting edges weights in the network. Xiang et al. [22] suggests a latent variable model for predicting the strength of a social tie, based on interaction and user affinity. However, data required by their model may not be publicly available due to user's privacy settings and thus their metric may not be applied in all applications.

Social networks as dissemination mechanisms have been part of many studies, in a variety of fields of viral marketing and voting systems. The goal is the influence maximization among all users of the network, with a small seed set to spread the information. Kempe et al. [9] were the first to formulate the problem of influence maximization in social networks given the weights of the social graph and suggested cascade models followed by users in the social networks, namely Independent Cascade (IC) Model and Linear Threshold Model (LT). In the Independent Cascade model whenever each user informed has a single chance of spreading the information to his/her neighbors and in the Linear Threshold model, users are associated with a threshold and when the sum of the edges of his/her already informed neighbors reaches the threshold, the user is then informed. Hangal et al. [20] explore the ability of social networks to disseminate a specific piece of information to particular users, in a way similar to the real life, where users try to get in touch with an expert using their already existing social ties. Influence probability and propagation is also a case of study in [33], where past propagation traces are considered in order to accurately predict the information spread among users. None of the above studies considers time in the propagation process.

Closer to our study are the works of Chen et al. [10] and Liu et al. [1]. Chen et al. associate with its edge aside from the probability of propagating the information, the meeting probability. Based on that, they suggest variations of Independent Cascade and Linear Threshold model to estimate the timely propagation of information. Liu et al. [1] suggest a simulation based approximation algorithm for solving the problem of influence maximization under time constraints in social networks and associate each edge with probabilities expressing the chance of distributing the information within a specific time lapse. Efficient influence maximization in time diffusion networks is studied by Gomez-Rodriquez and Schölkopf [7] in which information or influence can be propagated in different rates across different edges. Wang et al. [8] suggest variations of the Independent Cascade model similar to [10] in order to minimize the computational cost. Similarly to ours, paths whose probabilities are below a given tunable threshold θ are excluded. These approaches, contrary to our work, consider the influence graph given and try to maximize the influence spread in the entire network rather than a particular subset of users in the network that are most likely interested in receiving the information. They also focus on viral marketing campaigns or voting systems that do not have to operate under tight time constraints and resource saving, unlike emergency notification. Finally, none of the above works consider location information in the dissemination process.

Emergency notification with users having different roles according to their correlation to the event is studied by Kyungabaek *et al* [11]. They suggest two types of users,

[6]http://blogs.adobe.com/digitalmarketing/digital-marketing/mobile/adobe-2013-mobile-consumer-survey-71-of-people-use-mobile-to-access-social-media/

the ones located to the area of the event and those socially connected to them. They propose a system for alerting users that is aware of the geographies and social ties of the users. Social diffusion process is customized so that good propagation initiators are selected. In our previous work [18] we have proposed ESCAPE, which is an information dissemination approach. ESCAPE differs from LATITuDE in a number of ways. ESCAPE does not consider neither real-time constraints in the dissemination process nor the authoritativeness of the users in the spread of information, both of which are important criteria in emergency response situations that have to operate under tight time deadlines and resource savings. Contrary to ESCAPE, LATITuDE does not aim at just maximizing the spread of information to a subset of users, but in doing so timely. Furthermore, in this work we validate the efficiency of our solution not only with Twitter data, but also with data obtained from a major emergency event, the Sandy Hurricane.

VII. Conclusions

In this paper, we have presented our LATITuDE middleware that investigates the relationships and interactions among the members of a social group, and develops a dissemination mechanism to maximize the information reach within a time constraint after the occurrence of an emergency event. As we illustrate in our experimental evaluation, LATITuDE is able to execute orders of magnitude faster than the state-of-the-art technique IRIE as the number of seeds increases, while we also manage to inform more interested users. Furthermore, LATITuDE is able to inform considerably more users within a predefined deadline due to its intelligent seed selection process.

Acknowledgments

This research has been co-financed by the European Union (European Social Fund ESF) and Greek national funds through the Operational Program "Education and Lifelong Learning" of the National Strategic Reference Framework (NSRF) - Research Funding Program:Thalis-DISFER, Aristeia-MMD, Aristeia-INCEPTION Investing in knowledge society through the European Social Fund, the FP7 INSIGHT project and the ERC IDEAS NGHCS project.

References

[1] B. Liu, G. Cong, D. Xu, and Y. Zeng, "Time constrained influence maximization in social networks," in *ICDM*, Brussels, Belgium, December 2012.

[2] F. Alvanaki, S. Michel, K. Ramamritham, and G. Weikum, "See what's enblogue: real-time emergent topic identification in social media," in *EDBT*, Berlin, Germany, March 2012, pp. 336–347.

[3] A. M. Lesperance, M. A. Godinez, and J. R. Olson, *Social networking for emergency management and public safety*. Pacific Northwest National Laboratory Richland, WA, 2010.

[4] J. P. Bagrow, D. Wang, and A.-L. Barabasi, "Collective response of human populations to large-scale emergencies," *CoRR*, vol. abs/1106.0560, 2011.

[5] M. Guy, P. Earle, C. Ostrum, K. Gruchalla, and S. Horvath, "Integration and dissemination of citizen reported and seismically derived earthquake information via social network technologies," in *IDA*, Tucson, AZ, 2010.

[6] A. Goyal, F. Bonchi, and L. V. Lakshmanan, "A data-based approach to social influence maximization," vol. 5, no. 1, Seattle, WA, August 2011, pp. 73–84.

[7] M. G. Rodriguez and B. Schölkopf, "Influence maximization in continuous time diffusion networks," in *ICML*, Edinburgh, Scotland, June 2012.

[8] C. Wang, W. Chen, and Y. Wang, "Scalable influence maximization for independent cascade model in large-scale social networks," *Data Mining and Knowledge Discovery*, vol. 25, no. 3, pp. 545–576, 2012.

[9] D. Kempe, J. Kleinberg, and É. Tardos, "Maximizing the spread of influence through a social network," in *SIGKDD*, Washington, DC, USA, August 2003, pp. 137–146.

[10] W. Chen, W. Lu, and N. Zhang, "Time-critical influence maximization in social networks with time-delayed diffusion process," in *AAAI*, Toronto, Ontario, Canada, July 2012, pp. 592–598.

[11] K. Kim, Y. Zhao, and N. Venkatasubramanian, "Gsford: Towards a reliable geo-social notification system," in *SRDS*, Irvine, CA, October 2012, pp. 267–272.

[12] J. Singh and J. Bacon, "Event-based data control in healthcare," in *Middleware*, Leuven, Belgium, December 2008, pp. 84–86.

[13] M. Iqbal, X. Wang, D. Wertheim, and X. Zhou, "Swanmesh: A multicast enabled dual-radio wireless mesh network for emergency and disaster recovery services." *JCM*, vol. 4, no. 5, pp. 298–306, 2009.

[14] J. J. Ordille, P. Tendick, and Q. Yang, "Publish-subscribe services for urgent and emergency response," in *COMSWARE*, Dublin, Ireland, June 2009, pp. 8:1–8:10.

[15] A. Boutet, D. Frey, R. Guerraoui, A. Jégou, and A.-M. Kermarrec, "Whatsup decentralized instant news recommender," in *IPDPS*, Boston, MA, May 2013.

[16] P. Eugster, P. Felber, and F. Le Fessant, "The "art" of programming gossip-based systems," *SIGOPS Oper. Syst. Rev.*, vol. 41, no. 5, pp. 37–42, Oct. 2007.

[17] R. Friedman, D. Gavidia, L. Rodrigues, A. C. Viana, and S. Voulgaris, "Gossiping on manets: the beauty and the beast," *SIGOPS Oper. Syst. Rev.*, vol. 41, no. 5, Oct 2007.

[18] I. Litou, I. Boutsis, and V. Kalogeraki, "Efficient dissemination of emergency information using a social network," in *MUD*, Athens, Greece, March 2014.

[19] "Sandy hurricane," http://en.wikipedia.org/wiki/Hurricane_Sandy.

[20] S. Hangal, D. MacLean, M. S. Lam, and J. Heer, "All friends are not equal: Using weights in social graphs to improve search," in *SNA-KDD*, Washington, DC, July 2010.

[21] D. Taylor, "Everything you need to know about facebooks edgerank," *The Next Web*, vol. 9, May 2011.

[22] R. Xiang, J. Neville, and M. Rogati, "Modeling relationship strength in online social networks," in *WWW*, Raleigh, NC, USA, April 2010.

[23] B. Krishnamurthy, P. Gill, and M. Arlitt, "A few chirps about twitter," in *WOSN*, Seattle, WA, USA, August 2008, pp. 19–24.

[24] J. Travers and S. Milgram, "An Experimental Study of the Small World Problem," *Sociometry*, vol. 32, no. 4, pp. 425–443, Dec 1969.

[25] N. O. Hodas and K. Lerman, "How visibility and divided attention constrain social contagion," in *SocialCom-PASSAT*, Amsterdam, Netherlands, September 2012, pp. 249–257.

[26] A. Clauset, C. R. Shalizi, and M. E. J. Newman, "Power-law distributions in empirical data," *SIAM Rev.*, vol. 51, no. 4, Nov. 2009.

[27] G. Nemhauser, L. Wolsey, and M. Fisher, "An analysis of approximations for maximizing submodular set functions - i," *Mathematical Programming*, vol. 14, no. 1, pp. 265–294, 1978.

[28] G. Valkanas and D. Gunopulos, "Location extraction from social networks with commodity software and online data," in *ICDMW*, Brussels, Belgium, December 2012.

[29] S. Das, Ö. Egecioglu, and A. El Abbadi, "Anónimos: An lp-based approach for anonymizing weighted social network graphs," *IEEE Trans. Knowl. Data Eng.*, vol. 24, no. 4, pp. 590–604, 2012.

[30] K. Jung, W. Heo, and W. Chen, "Irie: Scalable and robust influence maximization in social networks." in *ICDM*, Brussels, Belgium, December 2012, pp. 918–923.

[31] A. Zubiaga and H. Ji, "Tweet, but Verify: Epistemic Study of Information Verification on Twitter," *Social Network Analysis and Mining*, In Press.

[32] K. Saito, R. Nakano, and M. Kimura, "Prediction of information diffusion probabilities for independent cascade model," in *KES*, Zagreb, Croatia, Sep 2008, pp. 67–75.

[33] F. Bonchi, "Influence propagation in social networks: A data mining perspective," *IEEE Intelligent Informatics Bulletin*, vol. 12, no. 1, pp. 8–16, 2011.