J Stat Phys (2013) 152:399–418
DOI 10.1007/s10955-013-0759-z

The Social Amplifier—Reaction of Human Communities to Emergencies

Yaniv Altshuler · Michael Fire · Erez Shmueli · Yuval Elovici · Alfred Bruckstein · Alex (Sandy) Pentland · David Lazer

Received: 13 April 2013 / Accepted: 23 April 2013 / Published online: 2 July 2013

Abstract This paper develops a methodology to aggregate signals in a network regarding some hidden state of the world. We argue that focusing on edges around hubs will under certain circumstances amplify the faint signals disseminating in a network, allowing for

This work was supported in part by the Robert Shillman Fund for Global Security—Technion North-Eastern Partnership and the Defense Threat Reduction Agency.

Y. Altshuler · E. Shmueli · A. Pentland
MIT Media Lab, Cambridge, MA, USA

Y. Altshuler
e-mail: yanival@media.mit.edu

E. Shmueli
e-mail: shmueli@media.mit.edu

A. Pentland
e-mail: sandy@media.mit.edu

M. Fire · Y. Elovici
Deutsche Telekom Lab, Department of Information Systems Eng., Ben-Gurion University, Beer-Sheva, Israel

M. Fire
e-mail: mickyfi@bgu.ac.il

Y. Elovici
e-mail: elovici@bgu.ac.il

A. Bruckstein
Computer Science Department, Technion, Haifa, Israel
e-mail: freddy@cs.technion.ac.il

D. Lazer (✉)
College of Computer and Information Science & Department of Political Science, Northeastern University, Boston, MA, USA
e-mail: d.lazer@neu.edu

D. Lazer
John F. Kennedy School of Government, Harvard University, Cambridge, MA, USA

more efficient detection of that hidden state. We apply this method to detecting emergencies in mobile phone data, demonstrating that under a broad range of cases and a constraint in how many edges can be observed at a time, focusing on the egocentric networks around key hubs will be more effective than sampling random edges. We support this conclusion analytically, through simulations, and with analysis of a dataset containing the call log data from a major mobile carrier in a European nation.

Keywords Network science · Mobile phone networks

1 Introduction

Imagine a scenario where some set of individuals witness an extraordinary event which impels them to communicate regarding that event to other individuals, who in turn will communicate with yet others. In this scenario, it is possible for an external observer to witness the fact of communication, but not the content. How might that observer effectively make the inference that an extraordinary event has occurred?

This is in fact a plausible scenario, with the existence of communication systems (most notably phones) where timing and volume of traffic is observed, but (typically) not content. Mobile phones are particularly notable in this regard, because of how pervasive they are. Here we build on work examining detection of anomalous events in networks [7], but with the focus on how to aggregate those signals in a computationally efficient fashion. That is, if one cannot observe all nodes and edges, how best to sample the network?

Analyzing the spreading of information has long been the central focus in the study of social networks for the last decade [6, 17, 18]. One of the main challenges associated with modeling of behavioral dynamics in social communities with respect to anomalous external events stems from the fact that it often involves stochastic generative processes. A further challenge is the trade off that exists between coverage and prediction accuracy [3, 5, 6]. While simulations on realizations from these models can help explore the properties of networks [16], a theoretical analysis is much more appealing and robust. The results presented in this work are based on a pure theoretical analysis, validated both by extensive simulations as well as by real world data derived from a unique dataset.

Contribution In this work we present an innovative approach for studying the network dimension of the changes that take place in social communities in the presence of emergencies. We do so using a mechanism we call a "*Social Amplifier*"—a method for analyzing local sub-networks spanning certain high-volume network nodes. The innovation in our proposed approach is twofold: (a) using a non-uniform sampling of the network (namely, focusing on activity in the social vicinity of network hubs), and (b) projecting the network activity into a multi-dimensional feature space spanned around a multitude of topological network properties. We show using both simulation and real world data that starting with certain coverage level of the network, our method outperforms the use of either random sampling, as well as single signal analysis.

Validation We first validate our technique using an analytic model that predicts the efficiency of our method for various network scenarios. Then, we conduct extensive experimental analysis, simulating various networks in which we examine the way information regarding an emergency spread. Using these results we demonstrate an assessment of the efficiency of our method compared to the conventional random network sampling. We further

validated our proposed methods using a comprehensive dataset, containing the entire internal calls as well as many of the incoming and outgoing calls within a major mobile carrier in a west European country, for a period of roughly 3 years. During this period that mobile users have made approximately 12 billion phone calls. We used the company's log files, providing all phone calls (initiator, recipient, duration, and timing) and SMS/MMS messages that the users exchange within and outside the company's network. All personal details have been anonymized, and we have obtained IRB approval to perform research on it.

Paper organization The rest of the paper is organized as follows: Sect. 2 discusses related work. Section 3 contains the problem's definitions. Section 4 presents the methodology used for testing our proposed mechanism. Section 5 discusses the Social Amplifier mechanism. An in-depth analysis of the technique using a simulated environment is presented in Sect. 6 whereas its demonstration using with real world cellular data is given in Sect. 7. Section 8 contains discussion and concluding remarks.

2 Related Work

There is an emerging literature on the use of network data to detect extraordinary events, especially around the use of mobile phone data. One important line of research examines the question pertaining to the area where the event has occurred, and its dependence on the nature of the event: a bomb attack is narrowly localized in space, thus likely the anomalous calling activity will be limited to the immediate neighborhood of the event. This was observed in an analysis of mobile data in the vicinity of a bomb attack [7].

Another critical question is: Whom do we call during an emergency? Do we call our closest ties or best friends? Or the acquaintance who we perceive as being physically closest to the event (and thus most likely to be affected)? Or do we call emergency personnel? The call data allow us to distinguish these scenarios—indeed, we can determine from the historical call pattern who are the individuals with whom a user communicates most frequently, as well as behavioral characteristics that might allow us to infer relationship type, as well as their most likely location during the emergency, together with the range of places they normally visit. This allows us to look for trends in the calling patterns during the emergency. An extended analysis of traffic patterns also shows that emergencies tend to induce cascading calling patterns, helping the spread of situation awareness in the population [7].

It has been recently shown that in trying to assess the societal changes and anomalous patterns that emerge in response to emergencies and security related events, it is crucial to understand the underlying social network [24, 25], the role of the link weights [15], as well as the response of the network to node and link removal [2]. Past research [20] had pointed out the existence of powerful patterns in the placement of links, where—as predicted in Granovetter's seminal work [15]—that clusters of strongly tied together individuals tend to be connected by weak ties. It was also shown that this finding provides insight into the robustness of the network to particular patterns of link and node removal, as well as into the spreading processes that take place in the social network [22, 23]. This latter result plays a key role in the intellectual basis of our understanding of the spread of situational awareness in an emergency.

Analysis of mobile networks traffic demonstrates a clear correlation between anomalies in the network traffic and a distinct security event. Indeed, in the event of a large scale explosion it is likely to assume that the authorities would be notified about it very close to the time of explosion. This, however, is not the case for many other scenarios. For example, the

large fire outbreak that had taken place in northern Israel at the beginning of December 2012 [26] was identified and discussed by numerous observers passing through this region hours before it caught the attention of relevant controllers in the fire department. Had an ability to detect anomalies in mobile networks been active during this point, an alert, titled "potential security event", focused on the specific geographic region, would have been produced, resulting in the dispatching of a local firefighting units, who could have easily extinguished the fire at this point.

One of the first works that examined the timing of people's communication behavior found that the lengths of the time gaps between two consecutive phone calls is distributed following a power law principle [8]. An extension of this work [10] showed that when an anomalous event is present, various similar statistical properties of the network dynamics gets distorted. Other works had examined the evolution of social groups, aiming for the development of algorithms capable of identifying "new groups"—a certain kind of anomalous network pattern [21].

The broader objective of this paper is to contribute to the ongoing development of more efficient techniques for utilizing mobile phones as a ubiquitous large scale sensors network. These techniques use mobile data to detect social relations [1, 12], mobility patterns [14], socio-economical properties [13], and security related features [4].

3 Preliminaries

We denote the "global social network" as a graph $G = \langle V, E \rangle$ where V is the set of all nodes and E is the set of undirected edges over those nodes (an edge (u, v) exists iff there have been reciprocal calls between users u and v).

We assume that occasionally various anomalous events take place in the "real world", that are being observed directly by some portion of the network's users, that subsequently may respond by making one of more phone calls to their neighbors in the global social network.

Given a mobile carrier M, we denote its set of covered nodes (derived from its market share) as $V_M \subseteq V$, and its set of covered edges as $E_M \subseteq E$. An edge is covered by M if at least one of its nodes is covered by M, i.e.:

$$E_M = \left\{ (u, v) \;\middle|\; (u, v) \in E \wedge (u \in V_M \vee v \in V_M) \right\}$$

We assume that the operator M is interested to detect anomalous events such as emergencies, and to do so as fast as possible, with as high accuracy rate as possible, and using as little resources as possible. We measure the amount of resources required by M (or otherwise defined as the complexity of the detection algorithm) as the overall number of edges being analyzed, or monitored. We denote the subset of edges processed by M as the "monitored edges", $S_M \subseteq E_M$.

Given an upper bound, ϵ on the size of $\frac{|S_M|}{|E|}$, we are interested in achieving the highest detection performance (defined later on) that can be obtained by monitoring a portion of the edges smaller than ϵ.

Throughout this work we refer to the 1 ego-network and the 1.5 ego-network of a node v.

The 1 ego-network of v is a graph $G(V_1, E_1)$ such that nearest-neighbor nodes of v comprise the vertices of V_1 and the links between v and V_1 comprise the edges E_1. Similarly, the 1.5 ego-network of v is a graph $G(V_{1.5}, E_{1.5})$, such that $V_{1.5}$ is the same as V_1 and $E_{1.5}$ consists of E_1 plus all direct links between the nearest-neighbors of v. The definitions of 1 ego-network and the 1.5 ego-network are illustrated in Fig. 1.

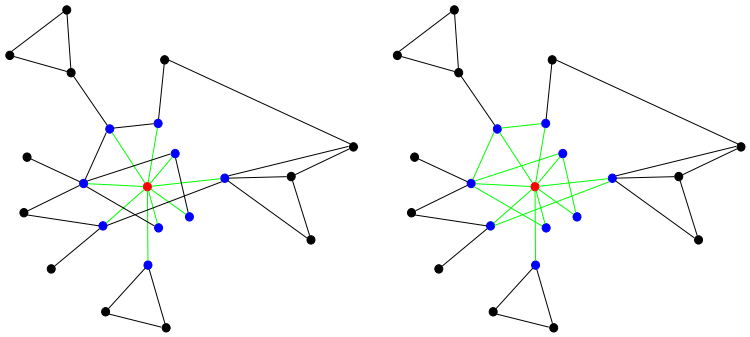

Fig. 1 An illustration of the 1 ego-network around a node v, marked in *red* (*left*) and the 1.5 ego-network around v (*right*). The nodes $V_1 = V_{1.5}$ are marked in *blue*. The E_1 edges (*left chart*) and the $E_{1.5}$ edges (*right chart*) are marked in *green* (Color figure online)

4 Methodology

In the next three sections we analyze the performance of our proposed method by comparing it to a "baseline" algorithm that uses random sampling of the network (instead of hubs-sampling) and tracks the edges of the 1 ego-network (instead of the 1.5 ego-network used in our method).

In Sect. 5 we construct an analytic model depicting the efficiency of our method and the way it changes as a result of changes in the network's conditions. In Sect. 6 we compare the detection time of our method to the baseline detection method using extensive simulations. In Sect. 7 we conduct a posteriori analysis of the detection accuracy of our proposed method, as well as the baseline detection method, and compare them for several different kinds of emergencies.

It should be noted that in order to compare our proposed method to the baseline algorithm it has to be normalized. We do so by presenting all performance results as a function of the number of "overall edges under converge".

Table 1 contains a glossary of the technical annotations used in this paper (defined in detail in the following sections).

5 The Social Amplifier

The proposed method is comprised of three stages: (a) network sampling, by detecting nodes with high incoming and outgoing traffic volume (i.e. hubs), (b) building the social network around the hubs and extracting the topological features of the networks, and (c) analyzing the changes in these features along time, tracking anomalous dynamics that imply the existence of an anomalous event.

5.1 Network Sampling

At the initial step of the analysis process we track the traffic volume in the network's nodes, looking for hubs—nodes with high traffic (either incoming or outgoing). The rationale behind the use of hubs is that hubs are highly likely to be exposed to new information, due to their high degree.

Table 1 A glossary of the main definitions used in this paper

V_M	The nodes of the mobile network.
E_M	The edges of the mobile network.
S_M	The edges actually analyzed by the monitoring system.
E	The overall amount of "energy" available to the monitoring system.
$E_{INITIAL}$	The energy spent on producing a high quality topographical coverage of the network.
$E_{AMPLIFIER}$	The energy spent for maintaining a 1.5 ego-network closure.
E_{DETECT}	The energy spent on the actual detection of the signal.
α	The exposure coefficient of an event (the portion of nodes exposed to the event).
$\langle k \rangle$	The average degree of the network's nodes.
k_{MAX}	The maximal degree of the network's nodes.
λ	The "social amplification constant" of a network.
ϵ	The portion of the edges being monitored (namely, the ratio of $\lvert S_M \rvert$ by $\lvert E \rvert$).
cp	The coverage portion of the mobile carrier from the entire edges (the ratio $\lvert E_M \rvert$ by $\lvert E \rvert$).
w	The initial number of witnesses to an event.
$\mathcal{V}_{BASE}$	A prediction vector generated by the base method.
$\mathcal{V}_{AMPLIFIED}$	A prediction vector generated by the social amplifier.
$\delta_{\lvert E \rvert}$	The difference in the performance of $\mathcal{V}_{BASE}$ and $\mathcal{V}_{AMPLIFIED}$, for $\lvert E \rvert$ monitored edges.
$\Delta(w, c, cp)$	Delta in detection times, for a given parameters w, c, cp.
c	Confidence level (the minimal number of edges generated by nodes exposed to an event, that needs to be monitored in order to deduce the existence of the event, in a sufficient level of confidence).

Given available resources ϵ, we select network nodes, $v_1, \ldots, v_n$, from V_M, such that those nodes have the highest degrees in V_M and the set $S_M = \bigcup_{1 \le i \le n} E_M^{1.5}(v_i)$ does not contain more than ϵ portion of the edges, where $E_M^{1.5}(v)$ denote the 1.5 ego-network around node v, that is—the edges between v and all of v's neighbors, as well as the edges between v's neighbors and themselves:

$$E_M^{1.5}(v_k) = E_M(v_k) \cup \left\{ (u_1, u_2) \;\middle|\; (u_1, u_2) \in E_M \wedge u_1 \in E(v_k) \wedge u_2 \in E(v_k) \right\}$$

The use of the 1.5 ego-network is required in order to analyze not only the overall number of calls in the network (sampled by the hubs), as done in works such as [7], but rather to generate the actual networks around the hubs, in order to enable their in-depth analysis. More specifically, analyzing only the overall number of calls, can only detect massive global events, but not local ones (unless the local events are known in advance, and the local data is analyzed in retrospective). Figure 2 illustrates the importance of using the 1.5 ego-network.

5.2 Features Extractions

For each phone social network we extract the following set of 21 topological features for each day during the test period:

- **In Degree**: The number of incoming edges of the hub.
- **Out Degree**: The number of outgoing edges of the hub.
- **BI Direction Numbers**: The amount of numbers that called the hubs, and that the hub called them.
- **Total Degree**: In degree + out degree of the hub.
- **In Calls**: The number of incoming calls to the hub.

- **Out Calls**: The number of outgoing calls from the hub.
- **Total Calls**: In calls + out calls.
- **Norm In Calls**: the number of in calls divided by the number of total calls.
- **Norm Out Calls**: the number of out calls divided by the number of total calls.
- **Neighborhood Number of Connection**: the number of connections in the 1.5 ego-network of the hub.
- **Number of Strong Connected Components**: the number of strong connected components of the 1.5 ego-network of the hub.
- **Number of Weak Connected Components**: number of weak connected components of the 1.5 ego-network of the hub.
- **Average Number of Strong Connected Components**: the average number of nodes in each strong connected components of the 1.5 ego-network of the hub.
- **Average Number of Weak Connected Components**: the average number of nodes in each weak connected components of the 1.5 ego-network of the hub.
- **Subgraph Density**: the density of the 1.5 ego-network of the hub.
- **Neighborhood Number of Connection**: the number of connections in the network received when removing from the 1.5 ego-network of the hub the hub itself, and all edges directly connected to it.
- **Number of Strong Connected Components**: the number of strong connected components of the 1.5 ego-network of the hub the hub itself, and all edges directly connected to it.
- **Number of Weak Connected Components**: number of weak connected components of the 1.5 ego-network of the hub the hub itself, and all edges directly connected to it.
- **Average Number of Strong Connected Components**: the average number of nodes in each strong connected components of the 1.5 ego-network of the hub the hub itself, and all edges directly connected to it.
- **Average Number of Weak Connected Components**: the average number of nodes in each weak connected components of the 1.5 ego-network of the hub the hub itself, and all edges directly connected to it.
- **Subgraph Density**: the density of the 1.5 ego-network of the hub the hub itself, and all edges directly connected to it.

5.3 Anomalies Detection

In order to detect anomalies in the dynamics of the social network around the network's hubs we use the Local-Outlier-Factor (LOF) anomaly detection algorithm [9]. In other words, using the LOF algorithm for each number we detected days which anomaly features occurred and then by using ensemble of all the hubs that we detected which dates have the highest probability for anomaly (using majority voting method).

We do so by ranking each day according to the number of hubs that reported it as anomalous. Then, for each day we look at the 29 days that preceded it, and calculate the final score of the day by its relative position in terms of anomaly-score within those 30 days. Namely, a day would be reported as anomalous (e.g., likely to contain some emergency) if it is "more anomalous" compared to the past month, in terms of the number of hubs-centered social networks influenced during it. Each day is given a score between 0 and 1, stating its relative "anomaly location" within its preceding 30 days.

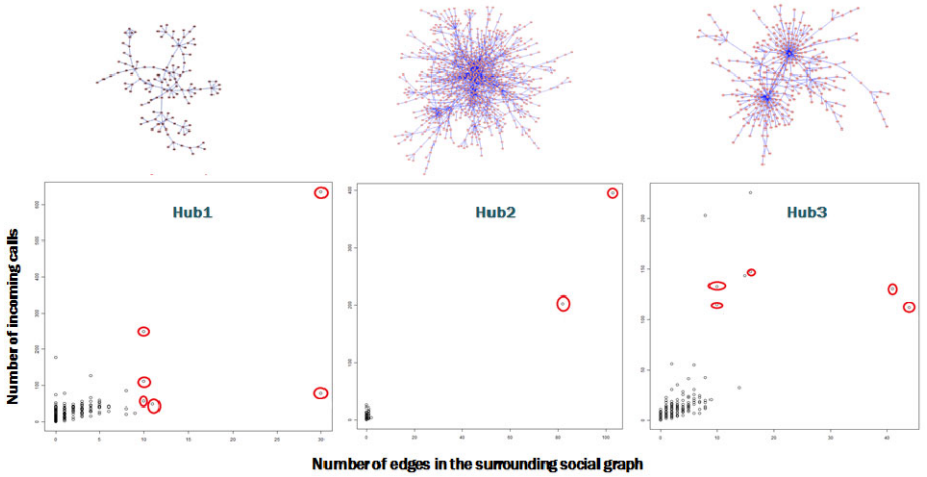

Fig. 2 Three examples of the social networks that are centered around 3 hubs of the network. The top three charts contains the 1.5 ego-network around a hub v (different hub for different chart), from which the hub itself, together with the edged connecting it to the rest of the nodes, were deleted for the sake of visual clarity. The *bottom three charts* represent the classification of days by 2 network properties: the overall number of calls to or from the hub (the *Y axis*), and the overall number of calls between the hub's neighbors (the *X axis*). Each *dot* represents a single day, and is placed according to the values of said properties at that day. Note how even in this 2-dimensional projection, certain emergencies (known in retrospect, and denoted by *red circles*) are detected—which would not have been detected had we used only a single signal (as the chart would collapse downwards had we used only the number of edges in the surrounding social graph, and leftwards had we used only the number of incoming and outgoing edges to the hubs) (Color figure online)

5.4 Analytic Evaluation

Alongside its increased sensing capability, our proposed mechanism has also an additional overhead, in terms of additional edges that should be monitored, compared to the standard approach of "number of calls analysis". This is the result of the following two reasons:

- **Hubs**: Due to their high degree, whenever the edges associated with an additional hub are added to the monitored edges set they increase its size substantially (unlike the addition of a randomly selected node, that is expected to be of a much lower degree). Although this is often compensated by the hubs' high "accessibility" to new information, the hubs-sampling method is expected to achieve poor performance for very low values of k (the number of monitored edges), as this implies very low number of hubs (and hence, low topographical coverage of the network).
- **1.5 Ego-Network**: For some node v, although the number of nodes in its 1 ego-network equals exactly the number of nodes in its 1.5 ego-network, the latter is usually expected to have substantially larger amount of edges. Those edges are used solely for extracting the topological features of the social network, and have little influence on the aggregate number of calls (as the number of calls on the edges directly connected to the hub is significantly larger than the calls on edges connecting two of its neighbors).

We therefore write the utilization of the Social Amplifier mechanism as follows:

$$E = E_{INITIAL} + E_{AMPLIFIER} + E_{DETECT} \tag{1}$$

whereas E is the "energy" supplied to the system for monitoring some k edges, $E_{INITIAL}$ is the overhead spent on monitoring the first few hubs until we achieve good topographical

coverage of the network, $E_{AMPLIFIER}$ is the energy spent on maintaining a 1.5 ego-network closure (that is, the number of edges of the 1.5 ego-network minus the number of edges at the 1 ego-network), and E_{DETECT} denotes the resources spent on the actual detection of the signal.

We note that $E_{INITIAL}$ decreases with the time it takes the detection process to complete. In other words, as the event to be detected is more explicit and broadly observed, it will be detected using a shorter time, which implicitly increases the relative portion of $E_{INITIAL}$. We can therefore write:

$$E_{INITIAL} \approx \alpha \cdot E$$

for $\alpha \in [0, 1]$ the *exposure coefficient* of the event.

Notice that as the exposure coefficient of an event decreases, it means that additional edges (and nodes) are required in order to detect the event. For extreme low values of the exposure coefficient there is no longer much difference between adding "hubs" and adding random nodes (in terms of their degrees) to the monitored set of nodes. This means that the ratio between the number of edges between hubs' neighbors and the edges to and from the hubs increases, resulting in an increase in $E_{AMPLIFIER}$.

Namely, for high exposure coefficient values the ratio between $E_{AMPLIFIER}$ and E_{DETECT} is proportional to the ratio between the average aggregate degrees of hubs' neighbors and the average degree of the hubs themselves. For low exposure coefficient values this ratio converges to $\frac{1}{\langle k \rangle}$ (denoting by $\langle k \rangle$ the average degree of the network):

$$\frac{\lambda}{k_{MAX}} \leq \frac{E_{AMPLIFIER}}{E_{DETECT}} \leq \frac{\lambda}{\langle k \rangle}$$

denoting by k_{MAX} the maximal degree, and for $\lambda \geq 1$ being the *Social Amplification Constant* of the network.

The same effect is obtained when the portion of the edges being monitored ϵ changes, as low values for ϵ cause the ratio $\frac{E_{AMPLIFIER}}{E_{DETECT}}$ to decrease, and very high values of it cause it to converge to $\frac{\lambda}{\langle k \rangle}$. We can therefore write:

$$E_{AMPLIFIER} \approx \frac{\lambda \cdot E_{DETECT}}{\langle k \rangle + \alpha\epsilon(k_{MAX} - \langle k \rangle)} \approx \frac{\lambda \cdot E_{DETECT}}{\langle k \rangle(1 - \alpha\epsilon) + \alpha\epsilon k_{MAX}}$$

We shall therefore rewrite Eq. (1) as follows:

$$E_{DETECT} = \frac{E \cdot (1 - \alpha)}{1 + \frac{\lambda}{\langle k \rangle(1-\alpha\epsilon)+\alpha\epsilon k_{MAX}}} \tag{2}$$

Figure 3 illustrates the behavior of E_{DETECT} as a function of the changes in the exposure coefficient α and in the portion of edges being monitored ϵ. Notice how E_{DETECT} has a non-monotonous dependency on α, obtaining a global maximum for intermediate values.

6 Simulation Evaluation

The goal of our experiments was to check how the two different methods for selecting the subset S_M influence the time required by M to detect an event. In order to achieve this goal we simulated the spreading of events in generated scale-free graphs and measured the time taken for the mobile carrier to detect those events when using different coverage percentages and different methods for selecting the subset S_M.

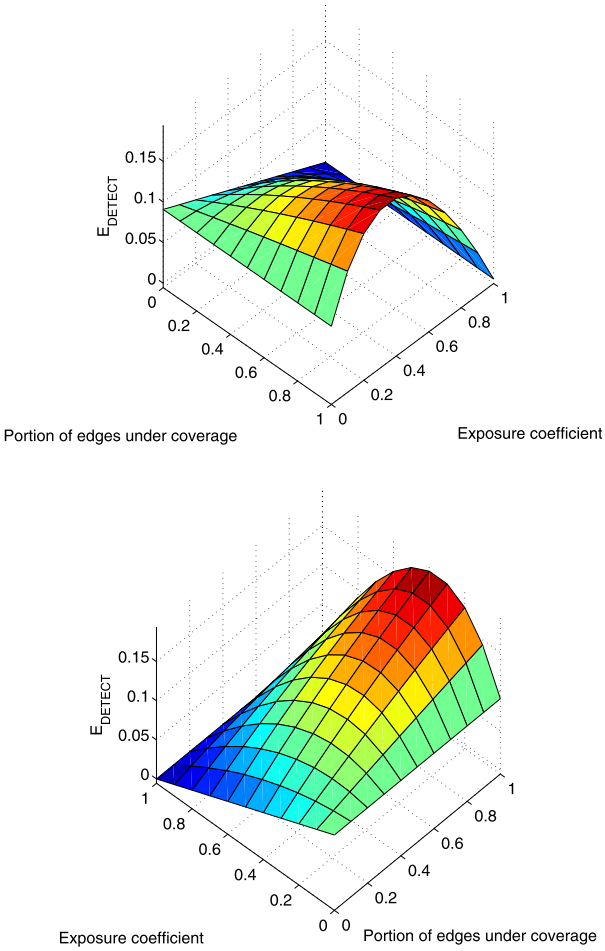

Fig. 3 The dependency of E_{DETECT} on the exposure coefficient α and on the number of edges being monitored. The illustration assumed $k_{MAX} = 10 \cdot \langle k \rangle$

The main steps of our simulation are outlined in Algorithm 1. In Step 1, a scale-free graph, $G\langle V, E\rangle$ with 100,000 nodes is randomly generated. In Step 2, different coverage percentages, cp, for the mobile operator are tested. In Steps 3–5, the nodes and edges of the mobile carrier $M\langle V_M, E_M\rangle$ are randomly selected, based on the coverage percentage cp. In Step 6, different numbers of initial witnesses to the event, w, are tested. The actual set of w witnesses is randomly selected in Step 7. In Step 8, different confidence levels c are tested. The confidence level is the minimum number of "spreading edges" (i.e. phone calls related to the event) that M has to sense in order to be confident that an event has occurred. In Step 9, different methods, f, for selecting S_M is determined. In Steps 10–11, we iterate over the nodes in V in the order determined by f. In Step 12, the current iterated node is added to S_M. In, Step 13, we simulate the spreading of an event on the configuration (G, M, W, S_M, c). This is done by executing Algorithm 2, which returns the number of time steps required to detect the event on the given configuration. Finally, in Steps 14–15, we store the configuration parameters and the resulting number of time steps.

Next, we describe Algorithm 2. At each time step (Steps 3–12), we iterate over the set W of nodes who "know" about the event (Steps 4–10). At first, the set W consists only of the nodes who witnessed the event. Each node v in W calls one of its friend nodes,

Algorithm 1 simulate

1: $G\langle V, E\rangle \leftarrow$ a randomly generated scale-free graph with $|V| = 100{,}000$;
2: **for** $cp \in \{0.1, 0.2, \ldots, 1.0$ **do**
3: $V_M \leftarrow cp \cdot |V|$ random nodes from $|V|$;
4: $E_M \leftarrow \{(u, v) \mid (u, v) \in E \wedge (u \in V_M \vee v \in V_M)\}$;
5: $M \leftarrow \langle V_M, E_M\rangle$;
6: **for** $w \in \{5, 10, 50, 100\}$ **do**
7: $W \leftarrow w$ random nodes from V;
8: **for** $c \in \{4, 8, 16, 32, 64, 128, 256, 512, 1024\}$ **do**
9: **for** $f \in \{Random, Hubs\}$ **do**
10: **for** $i = 1 \rightarrow |V|$ **do**
11: $v \leftarrow$ use f to select the next node from V;
12: $S_M \leftarrow S_M \cup E_M(v)$;
13: $t \leftarrow$ **call** Algorithm 2 with the parameters (G, M, W, S_M, c);
14: $sp \leftarrow |S_M|/|E_M|$;
15: **store result record** $(G, M, W, S_M, c, cp, w, f, sp, t)$;
16: **end for**
17: **end for**
18: **end for**
19: **end for**
20: **end for**

Algorithm 2 spreadEvent(G, M, W, S_M, c)

1: $t \leftarrow 1$;
2: $D \leftarrow \emptyset$;
3: **while** $|D| \leq c$ **do**
4: **for** $v \in W$ **do**
5: $u \leftarrow$ a "new" node from $E(v)$;
6: **if** $(v, u) \in S_M$ **then**
7: $D \leftarrow D \cup \{(v, u)\}$;
8: **end if**
9: $W \leftarrow W \cup \{u\}$;
10: **end for**
11: $t \leftarrow t + 1$;
12: **end while**
13: **return** t;

$u \in E(v)$, to which it didn't call before, and reports the event to it. If the edge (v, u) is being covered by M, then M has managed to detect a "spreading edge", and the edge is added to D (Steps 6–8). Since u knows about the event, it is added to W (Step 9). The main loop ends when the set of detected "spreading edges", D, is large enough, as required by c (Step 3). Finally, the number of time steps passed, t, is returned.

In order to reduce noise, we surrounded Steps 13, 3–19 and 1–20 of Algorithm 1 with three outer loops, each one executing 10 iterations.

The simulation results were analyzed as follows. First, for each coverage percentage cp, number of witnesses w, confidence c, and method f we plotted the required time steps t as

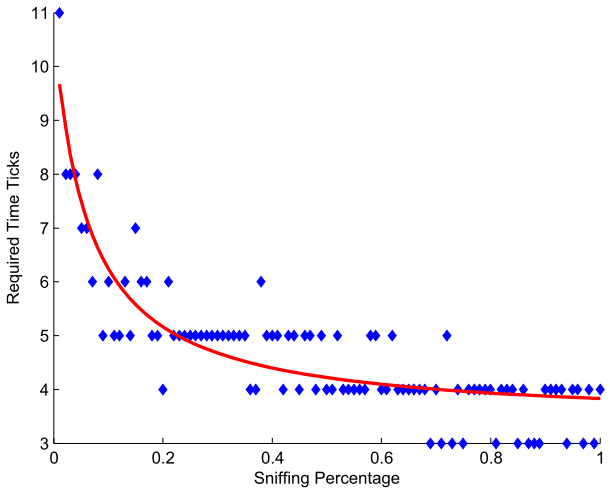

Fig. 4 The required time steps t as a function of the monitoring percentage sp ($cp = 0.1$, $w = 5$, $c = 4$ and $f = Hubs$). The *blue dots* are the original results of the simulation. The *red curve* was obtained after smoothing the points in order to demonstrate their trend (Color figure online)

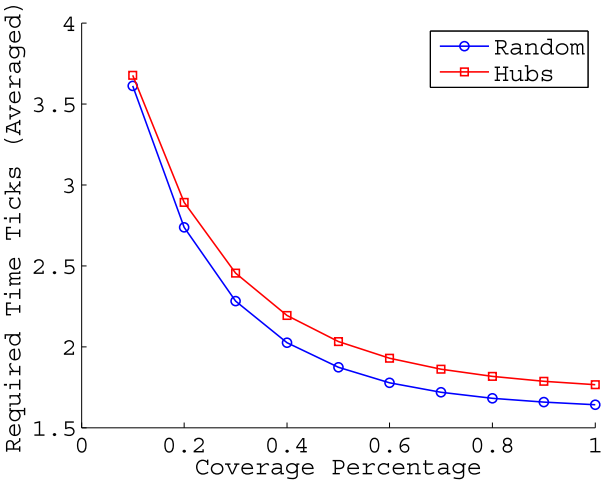

Fig. 5 $\tilde{t}(w, c, cp, f)$ for different values of cp and different methods f (using fixed $w = 5$ and $c = 4$). The *red line* represents $f = Hubs$ and the *blue line* represents $f = Random$ (Color figure online)

a function of the monitoring percentage sp. Figure 4 shows the resulting plot for $cp = 0.1$, $w = 5$, $c = 4$ and $f = Hubs$.

For each such plot, we calculated the value $\hat{t} = \int_0^1 t \, \mathrm{d}sp$. Intuitively, $\hat{t}$ is the averaged required time steps over the different monitoring percentages. (Note that $\hat{t}$ was calculated over the original data points and not over the smoothed ones.) Then, for each w, c, cp and f, we calculated $\tilde{t}(w, c, cp, f)$, which is the average of $\hat{t}$ values for all configurations with the same w, c, cp and f values (recall that Steps 13, 3–19 and 1–20 of Algorithm 1 were repeated 10 times each and therefore there are exactly 10^3 such configurations).

As an illustrating example, Fig. 5 shows the influence of cp on $\tilde{t}(w, c, cp, f)$ for the two different methods f when fixing $w = 5$ and $c = 4$.

Next, for each set of w, c and cp values, we calculated the delta in detection times, denoted as $\Delta(w, c, cp)$ and defined as follows:

$$\Delta(w, c, cp) = \tilde{t}(w, c, cp, Random) - \tilde{t}(w, c, cp, Hubs) \tag{3}$$

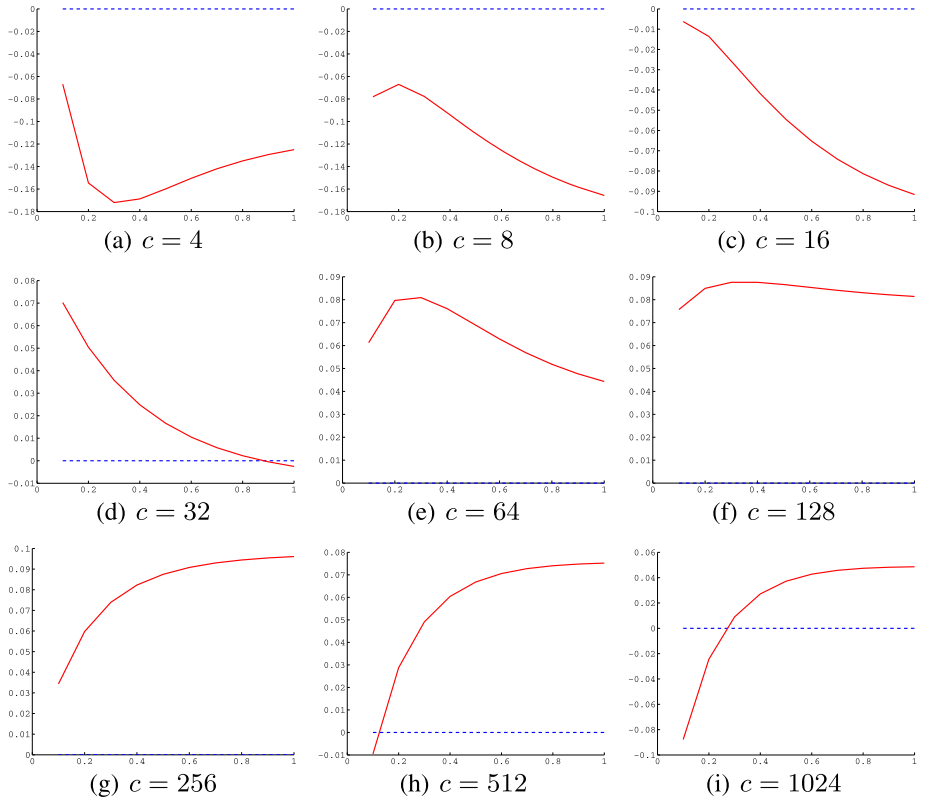

Fig. 6 The influence of cp on $\Delta(w, c, cp)$ for 9 different c values and a fixed $w = 5$. $\Delta(w, c, cp)$ is represented by the *red curve* in the figures. The *blue line* represents a fixed 0. Positive values of $\Delta(w, c, cp)$ mean an advantage for the Hubs method (Color figure online)

Figures 6, 7, 8 and 9 show $\Delta(w, c, cp)$ for $w = 5$, $w = 10$, $w = 50$ and $w = 100$ respectively. Each of these figures includes 9 sub-figures which correspond to the 9 different c values. Each sub-figure illustrates the influence of cp on $\Delta(w, c, cp)$ for fixed w and c values.

Figures 6–9 demonstrate that none of the methods is absolutely superior to the other. More precisely, we observe that for very small and very large confidence values, the Random method outperforms the Hubs method. However, for some range of medium confidence values, the Hubs method becomes superior. We also see that this range depends greatly on the initial number of witnessing nodes w. In other words, for larger values of w, the range of medium confidence values slides to larger confidence values. This makes sense as larger w values require less time steps in order to achieve the same confidence level.

The above observations led us to adopt the definition of the exposure coefficient α (defined in Sect. 5.4), combines w and c into a single variable: $\alpha = \log_2(c)/w$. Figure 10 illustrates the influence of α and cp on $\Delta(\alpha, cp)$ where $\Delta(\alpha, cp)$ is the average of all $\Delta(w, c, cp)$ such that $w \cdot cp = \alpha$. (Note that the original results were smoothened with $R = 0.804$ and $R^2 = 0.646$.) As expected, the Hubs method outperforms the Random method for medium α values. In addition, the advantage of the Hubs method further increases with larger cp values.

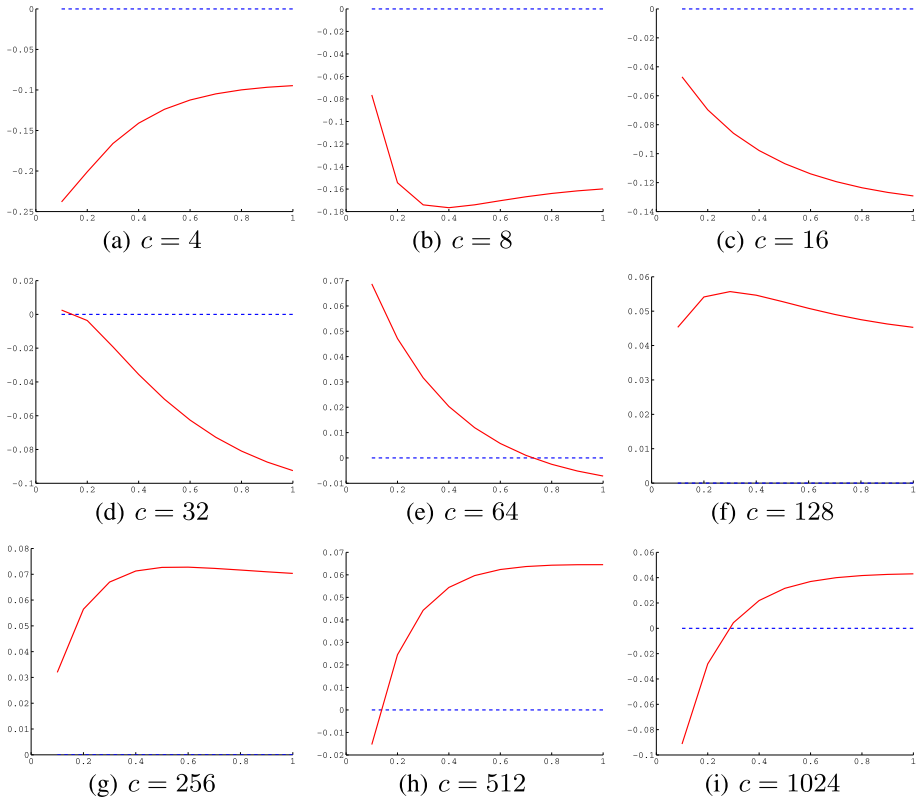

Fig. 7 The influence of cp on $\Delta(w, c, cp)$ for 9 different c values and a fixed $w = 10$. $\Delta(w, c, cp)$ is represented by the *red curve* in the figures. The *blue line* represents a fixed 0. Positive values of $\Delta(w, c, cp)$ mean an advantage for the Hubs method (Color figure online)

Note that the efficiency of our proposed method as illustrated in Fig. 10 closely resembles the prediction of our analytic model, as discussed in Sect. 5.4, and specifically in Eq. (2) and Fig. 3.

7 Emergencies Detection Using Real World Data

For evaluating our proposed Social Amplification technique as an enhanced method for anomalies detection we have used a series of anomalous events that took place in the mobile network country, during the time where the call logs data was recorded. Figure 11 presents the events, including their "magnitude", in terms of the time-span and size of population they influenced.

We have divided the anomalies into the following three groups:

Concerts and Festivals Events that are anomalous, but whose existence is known in advance to a large enough group of people. Those include events number 9–16, as appears in Fig. 11.

"Small exposure events" Anomalous events whose existence is unforeseen, and that were limited in their effect. Those include events numbers 1, 2, 5, 6.

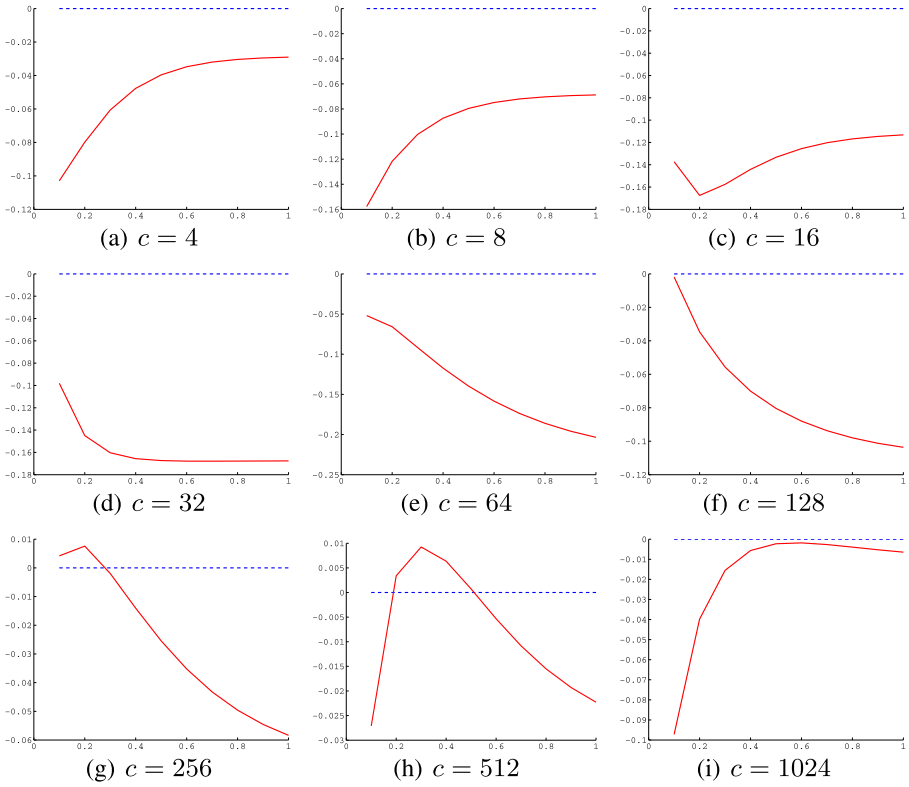

Fig. 8 The influence of cp on $\Delta(w, c, cp)$ for 9 different c values and a fixed $w = 50$. $\Delta(w, c, cp)$ is represented by the *red curve* in the figures. The *blue line* represents a fixed 0. Positive values of $\Delta(w, c, cp)$ mean an advantage for the Hubs method (Color figure online)

"Large exposure events" Anomalous events whose existence is unforeseen, that affected a large population. Those include events numbers 3, 4, 7, 8.

For each of the events we used the method described in Sect. 5 in order to rank each day between 0 and 1, according to its "anomalousness". This was done for increasingly growing number of monitored edges, in order to track the evolution of the detection accuracy. The result of this process was a series a numeric vectors pairs: $(\mathcal{V}_{BASE}, \mathcal{V}_{AMPLIFIED})_{|E|}$, corresponding to the two networks used (e.g. the random network sampling for $\mathcal{V}_{BASE}$ and the social-amplified hubs-sampling for $\mathcal{V}_{AMPLIFIED}$), for $|E|$ edges which were monitored. In addition, we created a binary vector $\hat{\mathcal{V}}$ having '1' for anomalous days and '0' otherwise.

For $|E|$ edges which were monitored we denote by $\delta_{|E|}$ the difference between the correlation coefficient of $\mathcal{V}_{AMPLIFIED}$ and $\hat{\mathcal{V}}$, and the correlation coefficient of $\mathcal{V}_{BASE}$ and $\hat{\mathcal{V}}$, namely:

$$\delta_{|E|} = CORR(\mathcal{V}_{AMPLIFIED}, \hat{\mathcal{V}}) - CORR(\mathcal{V}_{BASE}, \hat{\mathcal{V}})$$

for $(\mathcal{V}_{BASE}, \mathcal{V}_{AMPLIFIED})_{|E|}$, and for $CORR(x, y)$ the correlation coefficient function.

Notice that whereas $\delta_{|E|}$ measures the delta in detection accuracy, it has somewhat similar meaning to $\Delta(\alpha, cp)$, which measures delta in detection speed.

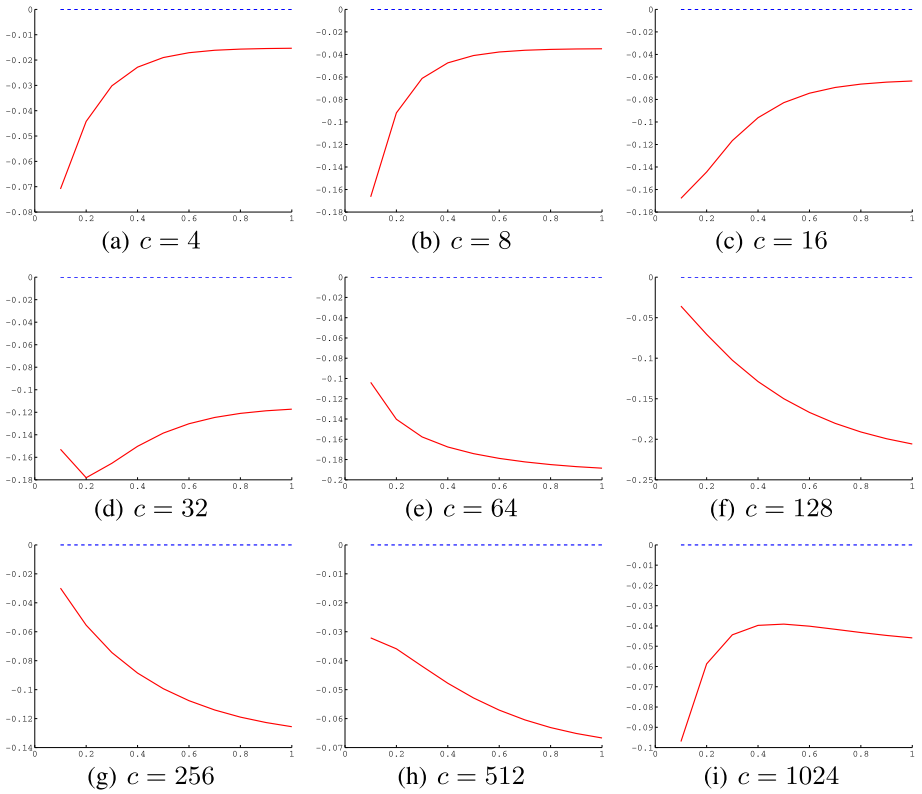

Fig. 9 The influence of cp on $\Delta(w, c, cp)$ for 9 different c values and a fixed $w = 100$. $\Delta(w, c, cp)$ is represented by the *red curve* in the figures. The *blue line* represents a fixed 0. Positive values of $\Delta(w, c, cp)$ mean an advantage for the Hubs method (Color figure online)

Figure 12 presents the values of $\delta_{|E|}$ for number of monitored edges ranging between 300 and 800. It can be seen how the hubs-sampling based social amplifier technique outperforms the basic random-sampling method. Furthermore, it can be seen that the positive delta increases with the increase in the amount of available resources (namely, number of monitored edges). This coincides with the analytic model that appears in Sect. 5 (and specifically, Fig. 3), as well as with the simulative analysis as illustrated in Fig. 10.

Figure 13 presents the values of $\delta_{|E|}$ for number of monitored edges between 300 and 800, for the three types of events. Notice how the results strongly coincide with the analytic model as is illustrated in Fig. 3, as concerts and events have the highest exposure value a, and the small exposure events have the lowest value.

8 Conclusions

This paper has demonstrated that under specified, and fairly broad, conditions, focusing on the neighborhood around a hub (the connections among the alters) enables efficient detection of events external to the network that provoke spreading communication within the network. Hubs act as collectors (and as a result, amplifiers) of social information, through facilitating

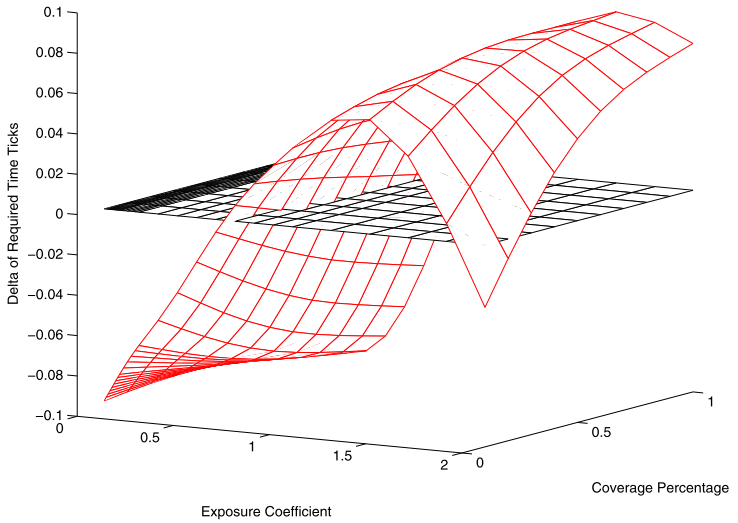

Fig. 10 The influence of α and cp on $\Delta(\alpha, cp)$ as evaluated using simulative environment. $\Delta(\alpha, cp)$ is represented by the *red area* in the figure. The *dark grid* represents the fixed $z = 0$ plane. Positive values of $\Delta(\alpha, cp)$ mean an advantage for the Hubs method (Color figure online)

		Event	duration (hours)	$\lvert G_\emptyset \rvert$
Emergencies	1	*Bombing*	1.92	750
	2	*Plane crash*	2.17	2,104
	3	*Earthquake*	1.42	32,403
	4	*Blackout*	3.0	84,751
	5	Jet scare	1.67	3,556
	6	Storm 1	2.33	7,350
	7	Storm 2	2.0	14,634
	8	Storm 3	1.75	19,239
Non-emergencies	9	*Concert 1*	13.25	11,376
	10	Concert 2	6.67	3,939
	11	Concert 3	9.08	5,134
	12	Concert 4	12.08	2,630
	13	*Festival 1*	19.92	66,869
	14	Festival 2	2.17	1,453
	15	Festival 3	20.92	10,854
	16	Festival 4	11.25	3,117

Fig. 11 A detailed list of the anomalous events that were identified, including their duration (in hours) and the number of population that resided in the relevant region (denoted as G_P). Further details can be found in [7]

the spread of communication in their immediate neighborhood. Traces of small scale information diffusion processes are more likely to be revealed when tracking hubs' activities compared to randomly selected nodes. In this work we show that this effect is so intense that in many cases it outperforms the analysis of significantly larger amount of random nodes (in order to compensate of the fact that the analysis of a single hub requires coverage of much

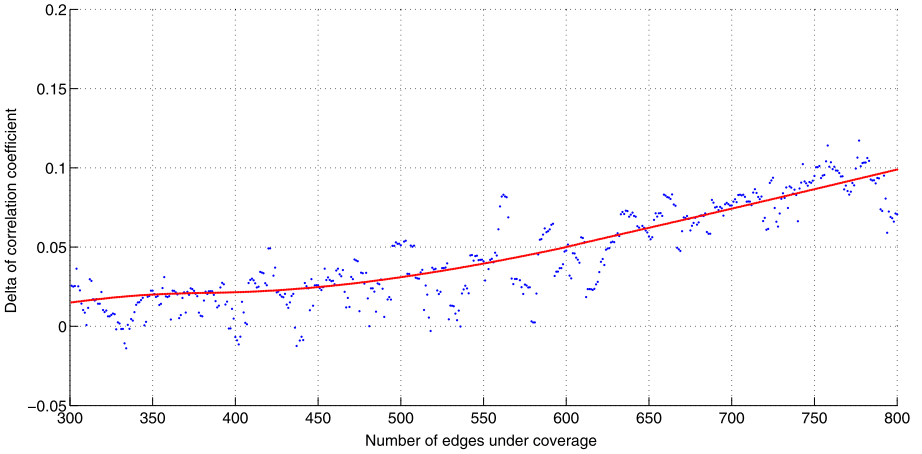

Fig. 12 The changes in the value of $\delta_{|E|}$ for growing numbers of edges being analyzed, evaluated using real anomalies and mobile calls data collected from a developed European country for a period of 3 years. Positive values indicate a higher detection efficiency of the social amplifier method compared to a basic random sampling method, for the same number of monitored edges

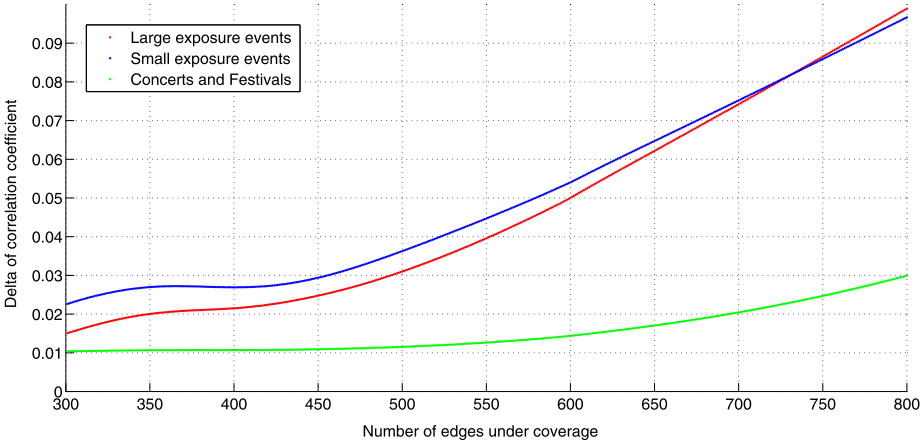

Fig. 13 The changes in the value of $\delta_{|E|}$ for growing numbers of edges being analyzed, segregated by the type of event detected. Notice how concerts and festivals that have high exposure value a generate relatively lower values of $\delta_{|E|}$ (but still monotonously increase with $|E|$), while the small exposure events are characterized by the highest values of $\delta_{|E|}$, specifically for low values of $|E|$. It is important to note that a low value of $\delta_{|E|}$ does not imply that the accuracy of the detection itself is low, but rather that the difference in accuracy between the two methods is small

more edges than required for an arbitrary node). We dub this effect "*Social Amplification*" and demonstrate it both analytically and experimentally.

We anticipate, however, that this methodology could be substantially extended and refined. For example, because hubs can sometimes be major bottlenecks, it is plausible that other neighborhoods within a large scale network would more efficiently act as social amplifiers. For example, it is possible that generally densely connected communities within a network would more efficiently disseminate observable changes in communication behavior,

virtually acting as a kind of "distributed hub" (the dramatic effect of the network topology on the dynamics of information diffusion in communities was demonstrated in works such as [11, 19]). It is also possible that the incorporation of other kinds of information about the properties of nodes would greatly improve the model.

References

1. Aharony, N., Pan, W., Ip, C., Khayal, I., Pentland, A.: Social fMRI: Investigating and shaping social mechanisms in the real world. Pervasive Mob. Comput. **7**(6), 643–659 (2011).
2. Albert, R., Jeong, H., Barabási, A.: Error and attack tolerance of complex networks. Nature **406**(6794), 378–382 (2000)
3. Altshuler, Y., Aharony, N., Fire, M., Elovici, Y., Pentland, A.: Incremental learning with accuracy prediction of social and individual properties from mobile-phone data. CoRR (2011)
4. Altshuler, Y., Aharony, N., Pentland, A., Elovici, Y., Cebrian, M.: Stealing reality: when criminals become data scientists (or vice versa). IEEE Intell. Syst. **26**(6), 22–30 (2011). doi:10.1109/MIS.2011.78
5. Altshuler, Y., Fire, M., Aharony, N., Elovici, Y., Pentland, A.: How many makes a crowd? On the correlation between groups' size and the accuracy of modeling. In: Intl. Conf. on Social Computing, Behavioral-Cultural Modeling and Prediction, pp. 43–52. Springer, Berlin (2012)
6. Altshuler, Y., Fire, M., Aharony, N., Volkovich, Z., Elovici, Y., Pentland, A.S.: Trade-offs in social and behavioral modeling in mobile networks. In: Social Computing, Behavioral-Cultural Modeling and Prediction, pp. 412–423. Springer, Berlin (2013)
7. Bagrow, J., Wang, D., Barabási, A.: Collective response of human populations to large-scale emergencies. PLoS ONE **6**(3), e17680 (2011)
8. Barabasi, A.: The origin of bursts and heavy tails in human dynamics. Nature **435**(7039), 207–211 (2005)
9. Breunig, M.M., Kriegel, H.P., Ng, R.T., Sander, J.: Lof: identifying density-based local outliers. In: ACM Sigmod Record, vol. 29, pp. 93–104. ACM, New York (2000)
10. Candia, J., González, M., Wang, P., Schoenharl, T., Madey, G., Barabási, A.: Uncovering individual and collective human dynamics from mobile phone records. J. Phys. A, Math. Theor. **41**, 224015 (2008)
11. Choi, H., Kim, S., Lee, J.: Role of network structure and network effects in diffusion of innovations. Ind. Mark. Manage. **39**(1), 170–177 (2010)
12. Eagle, N., Pentland, A., Lazer, D.: Inferring social network structure using mobile phone data. Proc. Natl. Acad. Sci. USA **106**, 15274–15278 (2009)
13. Eagle, N., Macy, M., Claxton, R.: Network diversity and economic development. Science **328**(5981), 1029–1031 (2010)
14. Gonzalez, M.C., Hidalgo, C.A., Barabasi, A.L.: Understanding individual human mobility patterns. Nature **453**(7196), 779–782 (2008). http://dx.doi.org/10.1038/nature06958
15. Granovetter, M.: The strength of weak ties. Am. J. Sociol. **78**(6), 1360–1380 (1973)
16. Herrero, C.: Ising model in scale-free networks: a Monte Carlo simulation. Phys. Rev. E **69**(6), 067109 (2004)
17. Huberman, B., Romero, D., Wu, F.: Social networks that matter: twitter under the microscope. First Monday **14**(1), 8 (2009)
18. Leskovec, J., Backstrom, L., Kleinberg, J.: Meme-tracking and the dynamics of the news cycle. In: Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 497–506 (2009). CiteSeer
19. Nicosia, V., Bagnoli, F., Latora, V.: Impact of network structure on a model of diffusion and competitive interaction. Europhys. Lett. **94**, 68009 (2011)
20. Onnela, J., Saramäki, J., Hyvönen, J., Szabó, G., Lazer, D., Kaski, K., Kertész, J., Barabási, A.: Structure and tie strengths in mobile communication networks. Proc. Natl. Acad. Sci. **104**(18), 7332 (2007)
21. Palla, G., Barabasi, A., Vicsek, T.: Quantifying social group evolution. Nature **446**(7136), 664–667 (2007)
22. Pastor-Satorras, R., Vespignani, A.: Epidemic spreading in scale-free networks. Phys. Rev. Lett. **86**(14), 3200–3203 (2001)

23. Pastor-Satorras, R., Vespignani, A.: Evolution and Structure of the Internet: A Statistical Physics Approach. Cambridge University Press, Cambridge (2007)
24. Waclaw, B.: Statistical mechanics of complex networks. arXiv:0704.3702 (2007)
25. Wassermann, S., Faust, K.: Social Network Analysis: Methods and Applications, vol. 8. Cambridge University Press, Cambridge (1994)
26. Ynet: Carmel fire fully extinguished. http://www.ynetnews.com/articles/0,7340,L-3994847,00.html (2010)