The Social Amplifier – Reaction of Human Communities to Emergencies

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

I. 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 [2], 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 [4], [5]. 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 [1]. While simulations on realizations from these models can help explore the properties of networks [3], 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 certain coverage level of the network, our method outperforms the use of either random sampling, as well as single signal analysis.

Y. Altshuler, E. Shmueli and A. Pentland are with MIT Media Lab. E-mail: {yanival,shmueli,sandy}@media.mit.edu.

M. Fire and Y. Elovici are with Deutsche Telekom Lab & Department of Information Systems Engineering, Ben-Gurion University. E-mail: {mickyfi,elovici}@bgu.ac.il

A.M. Bruckstein is with Computer Science Department, Technion. E-mail: freddy@cs.technion.ac.il

D. Lazer is with College of Computer and Information Science & Department of Political Science, Northeastern University. E-mail: d.lazer@neu.edu

II. The Social Amplifier

The proposed method is comprised of three stages as follows.

In the initial stage, 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.

Given available resources ϵ, we select network nodes, $v_1, \ldots, v_n$ such that those nodes have the highest degrees in the network and the set $S_M = \bigcup_{1 \leq i \leq n} E^{1.5}(v_i)$ does not contain more than ϵ portion of the edges, where $E^{1.5}(v)$ 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.

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 [2], 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).

In the second stage, for each day during the test period, and each phone social network, we extract a set of 21 topological features, such as the In Degree, Out Degree, Number of Strong Connected Components, Subgraph Density, etc.

In the third stage, we detect anomalies in the dynamics of the social network around the network hubs, using the Local-Outlier-Factor (LOF) algorithm. Applying the LOF algorithm on each hub, detects days which anomaly features occurred. Then, by using ensemble of all the hubs, we detect which dates have the highest probability for anomaly.

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.

III. Validation

A. 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).
- **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.

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

$$E = E_{INITIAL} + E_{AMPLIFIER} + E_{DETECT} \quad (1)$$

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 $\frac{1}{<k>}$ (denoting by $< k >$ the average degree of the network) :

$$\frac{\lambda}{k_{MAX}} \leq \frac{E_{AMPLIFIER}}{E_{DETECT}} \leq \frac{\lambda}{< k >}$$

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 $\frac{\lambda}{<k>}$. We can therefore write :

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

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

We shall therefore rewrite Equation 1 as follows :

$$E_{DETECT} = \frac{E \cdot (1 - \alpha)}{1 + \frac{\lambda}{<k>(1-\alpha\epsilon)+\alpha\epsilon k_{MAX}}} \quad (2)$$

Figure 1 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.

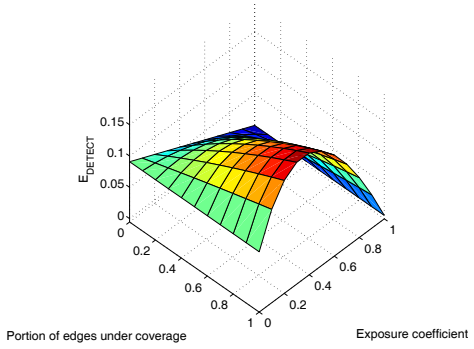

Fig. 1. The dependency of E_{DETECT} on the exposure coefficient α and on the number of edges being monitored. The illustration assumed $k_{MAX} = 10 \cdot < k >$.

B. Simulation

The goal of our simulation was to check how the two methods for selecting the subset of monitored edges (i.e. Social Amplifier vs. Random) influence the time required 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 to detect those events when using the two different methods for selecting the subset of monitored edges.

Our simulation included a tremendously large number of executions ($\approx 10^6$) and used different parameters:

- cp - the coverage percentage of the mobile operator.
- w - the number of initial witnesses to the event.
- c - the confidence level, i.e., the minimum number of "spreading edges" that need to be sensed in order to be confident that an event has occurred.

Following the analysis in Section III-A, we defined the exposure coefficient, denoted by α, as $\alpha = log_2(c)/w$.

Fig. 2 shows the influence of α and cp on $\Delta(\alpha, cp)$ where $\Delta(\alpha, cp)$ is the mean difference in detection time (between the two methods) over all executions with the given cp and $log_2(c)/w = \alpha$. (Note that the original results were smoothened with $R = 0.804$ and $R^2 = 0.646$.) As shown in the figure, for medium α values, the Social Amplifier method outperforms the Random method. In addition, we observe that in this range of medium α values, the advantage of the Social Amplifier method increases with larger cp values.

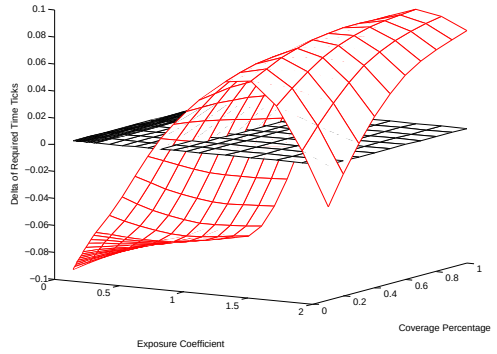

Fig. 2. The influence of α and cp on $\Delta(\alpha, cp)$ as evaluated using simulative environment. $\Delta(\alpha, cp)$ is represented by the red area in the figure. The dark grid represents the fixed $z = 0$ plane. Positive values of $\Delta(\alpha, cp)$ mean an advantage for the Hubs method.

Note that the efficiency of our method as illustrated in Figure 2 closely resembles that of our analytic model, as discussed in Section III-A (Equation 2 and Figure 1).

C. Real World Data

We also validated our method 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.

For evaluating the Social Amplifier 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.

We have divided the anomalies into the following three groups : (1) "Concerts and Festivals" Events that are anomalous, but whose existence is known in advance to a large enough group of people; (2) "Small exposure events" Anomalous events whose existence is unforeseen, and that were limited in their effect; and (3) "Large exposure events" Anomalous events whose existence is unforeseen, that affected a large population.

For each of the events we used the method described in Section II 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 of 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.

Figure 3 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 Figure 1, as concerts and events have the highest exposure value a, and the small exposure events have the lowest value.

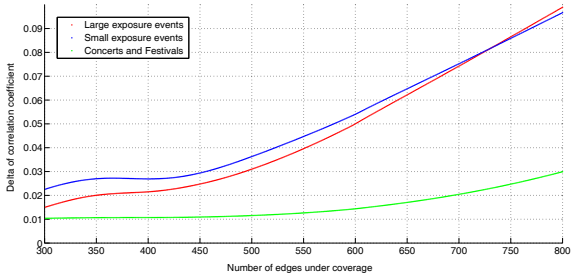

Fig. 3. The changes in the value of $\delta_{|E|}$ for growing numbers of edges being analysed, segregated by the type of event detected. Notice how concerts and festivals that have high exposure value a generate relatively lower values of $\delta_{|E|}$ (but still monotonously increase with $|E|$), 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 is small.

REFERENCES

[1] 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)

[2] Bagrow, J., Wang, D., Barabási, A.: Collective response of human populations to large-scale emergencies. PloS one **6**(3), e17,680 (2011)

[3] Herrero, C.: Ising model in scale-free networks: A monte carlo simulation. Physical Review E **69**(6), 067,109 (2004)

[4] Huberman, B., Romero, D., Wu, F.: Social networks that matter: Twitter under the microscope. First Monday **14**(1), 8 (2009)

[5] 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. Citeseer (2009)