Detecting Anomalous Behaviors Using Structural Properties of Social Networks

Yaniv Altshuler[1], Michael Fire[2], Erez Shmueli[1], Yuval Elovici[2], Alfred Bruckstein[3], Alex (Sandy) Pentland[1], and David Lazer[4]

[1] MIT Media Lab
{yanival,shmueli,sandy}@media.mit.edu

[2] Deutsche Telekom Lab, Department of Information Systems Engineering
Ben-Gurion University
{mickyfi,elovici}@bgu.ac.il

[3] Computer Science Department
Technion – Israeli Institute of Technology
freddy@cs.technion.ac.il

[4] College of Computer and Information Science & Department of Political Science
Northeastern University
d.lazer@neu.edu

Abstract. In this paper we discuss the analysis of mobile networks communication patterns in the presence of some anomalous "real world event". We argue that given limited analysis resources (namely, limited number of network edges we can analyze), it is best to select edges that are located around 'hubs' in the network, resulting in an improved ability to detect such events. We demonstrate this method using a dataset containing the call log data of 3 years from a major mobile carrier in a developed European nation.

Keywords: Mobile Networks, Anomalies Detection, Emergencies, Behavior Modeling

1 Introduction

Analyzing the spreading of information in human networks has long been the focus in many studies of social networks [17, 20]. A main challenge in practical analysis of the way information flows between network's participants is the trade-off between the available analytic resources and the accuracy of the prediction they yield [3, 6].

Imagine a scenario where several people observe some extraordinary event, which triggers a cascading sequence of reports between "social neighbors". In this scenario, it is possible for an external observer to track the volume of network traffic, but not its 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 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 [10], 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? We argue that an efficient monitoring strategy is focusing on network edges that are located in vicinity to network hubs.

We demonstrate our approach 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. The dataset also identifies the active cell towers of each call, thereby offering real time location (with tower resolution) data for each user. All personal details have been anonymized, and we have obtained IRB approval to perform research on it.

The rest of this paper contains related work that is discussed in Section 2, problem's definitions that is presented in Section 3, a demonstration of the proposed method using real world cellular data in Section 4 and concluding remarks appear in Section 5.

2 Related Work

Recent research around the use of mobile network data for detection of extraordinary events, had examined the question pertaining to the area where the event has occurred, and its nature: 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 [10].

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 [26, 27], the role of the link weights [16], as well as the response of the network to node and link removal [2].

Other works had examined the evolution of social groups, developing algorithms capable of identifying "new groups" – a certain kind of anomalous network pattern [23], or ways to cluster networks based on social and behavioral features [11]. In [19] the behavior and social patterns 2.5 million mobile phone users, making 810 million phone calls, were analyzed and resulted in clustering of the network to components showing striking resemblance to the geographical districts the users live in.

In the broader scope, this line of work aims for creating techniques for analyzing mobile phone data as ubiquitous and pervasive sensors networks. These techniques can be used to detect social relations [1, 13], evolving behavioral trends [7, 24], mobility patterns [15], environmental hazards [18, 25], socio-economical properties [14], and various security related features [4, 5].

Another question of interest with this regards is the specific way we utilize our sensors network. Namely, given a large number of sensors, what would be the best way to deploy them, in order to obtain maximal pervasiveness ? Is the optimal deployment must always be symmetric (as some works actually argue against this approach [8, 9, 21]) ?

3 Problem Definitions

We denote the “global social network” as a graph $G =< V, E >$ where V is the set of all nodes and E is the set of directed edges over those nodes (an edge (u, v) exists if and only if there has been a reciprocal call between users u and v).

We assume that occasionally various anomalous events take place in the “real world”, that are being directly observed by some of the network’s users, that subsequently may (or may not) react by calling one of more of their friends (i.e. their neighbors in the 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. :

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

We assume that the operator M is interested to detect anomalous events such as emergencies, and to do so with as high accuracy rate as possible, and using as little resources as possible. We measure the amount of resources required by M 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 the highest detection performance obtainable by monitoring a portion of the edges smaller than ϵ.

Throughout this work we refer to the “1 ego-network” of a node v as the graph $G_1(V_1, E_1)$ such that V_1 contains all of the nodes u such that there exists an edge (v, u) in E, and that E_1 contains all the edges from v to the nodes of V_1. Furthermore, we denote by the “1.5 ego-network” the graph $G_{1.5}(V_{1.5}, E_{1.5})$ such that $V_{1.5} = V_1$, and that $E_{1.5} = E_1 \cup \varDelta_{1.5}$ where $\varDelta_{1.5}$ contains all the edges in E between nodes u_i and u_j such that $u_i, u_j \in V - v$. The definitions of 1 ego-network and the 1.5 ego-network are illustrated in Figure 1.

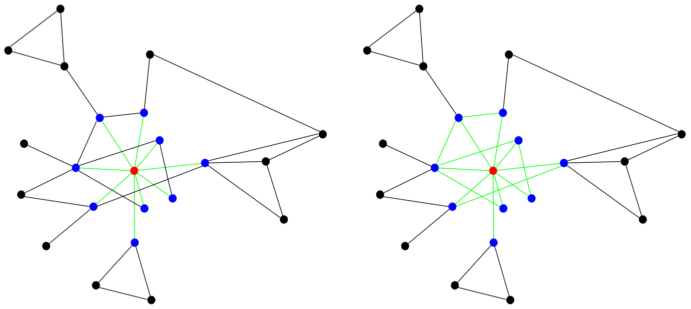


Fig. 1. An illustration of the 1 ego-network around (left chart) and the 1.5 ego-network (right chart) around the same node v (marked in red). The nodes $V_1 = V_{1.5}$ are marked in blue. The E_1 edges (left chart) and the $E_{1.5}$ edges (right chart) are marked in green.

3.1 Network Sampling

In this work we compare two alternatives for sampling a pre-defined number of network's edges. The basic method is simply to randomly select edges, with uniform distribution (this will serve as the baseline for comparison). The second method we propose in this work is to use the edges of the 1.5 ego-networks around network 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.

Specifically, 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 $S_M = \bigcup_{1 \leq i \leq n} E^{1.5}_M(v_i)$ does not contain more than ϵ portion of the edges, where $E^{1.5}_M(v)$ 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 :

$$E^{1.5}_M(v_k) = E_M(v_k) \bigcup \{(u_1, u_2) | (u_1, u_2) \in E_M \wedge u_1 \in E(v_k) \wedge u_2 \in E(v_k)\}$$

3.2 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. In other words, using the LOF algorithm for each network node we detected times where anomaly features occur. Using majority voting between all the hubs under monitoring we detected the times with the highest probability for anomalies.

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.

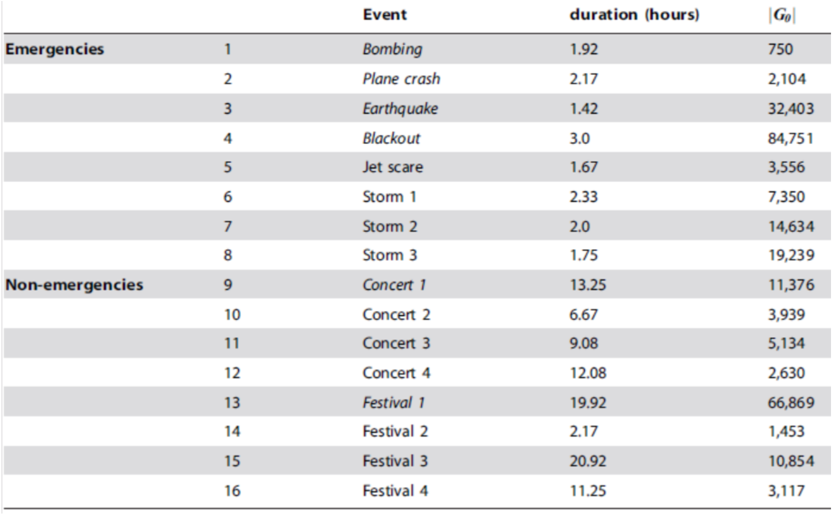
4 Emergency Detection using Real World Data

For evaluating our proposed monitoring method 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 2 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 Figure 2.

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



		Event	duration (hours)	$\|G_0\|$
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. 2. 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 [10].

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

We rank each day between 0 and 1, according to its "anomalousness", based on the method explained above. 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: $(\mathcal{V}_{BASE}, \mathcal{V}_{HUBS})_{|E|}$, corresponding to the two sampling methods used (e.g. the random network sampling for $\mathcal{V}_{BASE}$ and the hubs-sampling for $\mathcal{V}_{HUBS}$), 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 $\mathcal{V}_{HUBS}$ and $\hat{\mathcal{V}}$, and the correlation coefficient of $\mathcal{V}_{BASE}$ and $\hat{\mathcal{V}}$, namely :

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

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

Figure 3 presents the values of $\delta_{|E|}$ for number of monitored edges ranging between 300 and 800. It can be seen how the hubs-sampling 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).

Figure 4 presents the values of $\delta_{|E|}$ for number of monitored edges between 300 and 800, for the three types of events.

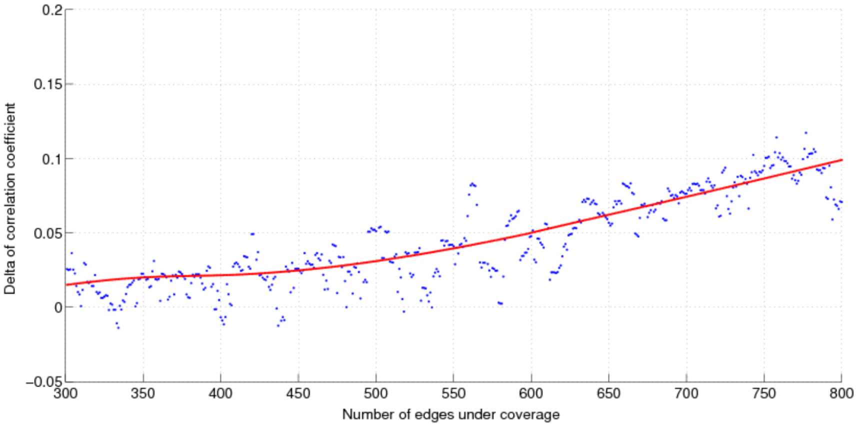


Fig. 3. The changes in the value of $\delta_{|E|}$ for growing numbers of edges being analyzed, evaluated using real anomalies and mobile calls data collected from a developed European country for a period of 3 years. Positive values indicate a higher detection efficiency of hubs-sampling compared to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

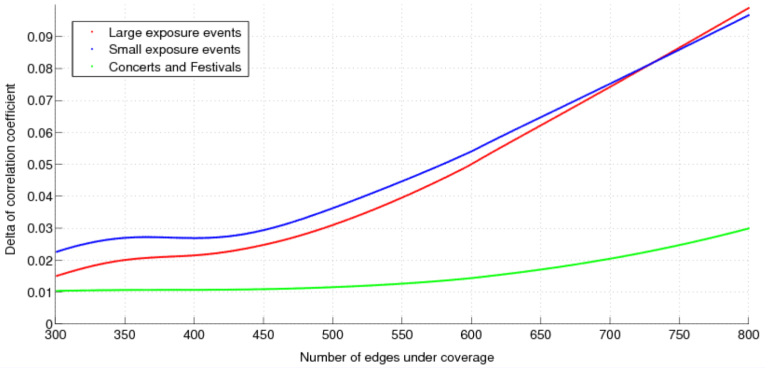


Fig. 4. The changes in the value of $\delta_{|E|}$ for growing numbers of edges being analysed, segregated by the type of event detected. Notice how concerts and festivals that have high exposure value a generate relatively lower values of $\delta_{|E|}$ (but still monotonously increase with $|E|$), 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 between the two methods is small.

5 Conclusions

In this paper we examined the problem of monitoring resources allocation for analyzing mobile networks, and have shown that by using social features of the network (namely, focusing on network hubs) prediction accuracy of anomalous events can be significantly increased. Specifically, we have shown that focusing on the neighborhood around a hub (the connections among the alters) will detect 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 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 more edges than required for an arbitrary node).

We anticipate, however, that this methodology could be further extended and refined. For example, as 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 [12,22]). It is also possible that the incorporation of other kinds of information about the properties of nodes would greatly improve the model.

References

1. Nadav Aharony, Wei Pan, Cory Ip, Inas Khayal, and Alex Pentland. Social fmri: Investigating and shaping social mechanisms in the real world. *Pervasive and Mobile Computing*, 2011.
2. R. Albert, H. Jeong, and A.L. Barabási. Error and attack tolerance of complex networks. *Nature*, 406(6794):378–382, 2000.
3. Y. Altshuler, N. Aharony, M. Fire, Y. Elovici, and A Pentland. Incremental learning with accuracy prediction of social and individual properties from mobile-phone data. *CoRR*, 2011.
4. Y. Altshuler, N. Aharony, A. Pentland, Y. Elovici, and M. Cebrian. Stealing reality: When criminals become data scientists (or vice versa). *Intelligent Systems, IEEE*, 26(6):22–30, nov.-dec. 2011.
5. Y. Altshuler, Y. Elovici, A.B. Cremers, N. Aharony, and A. Pentland. Security and privacy in social networks. *Recherche*, 67:02, 2012.
6. Y. Altshuler, M. Fire, N. Aharony, Y. Elovici, and A Pentland. How many makes a crowd? on the correlation between groups' size and the accuracy of modeling. In *International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction*, pages 43–52. Springer, 2012.
7. Y. Altshuler, W. Pan, and A Pentland. Trends prediction using social diffusion models. In *International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction*, pages 97–104. Springer, 2012.
8. Y. Altshuler, I.A. Wagner, and A.M. Bruckstein. On swarm optimality in dynamic and symmetric environments. volume 7, page 11, 2008.

9. Y. Altshuler, V. Yanovsky, I. Wagner, and A. Bruckstein. Swarm intelligencesearchers, cleaners and hunters. *Swarm Intelligent Systems*, pages 93–132, 2006.
10. J.P. Bagrow, D. Wang, and A.L. Barabási. Collective response of human populations to large-scale emergencies. *PloS one*, 6(3):e17680, 2011.
11. Carter T. Butts. The complexity of social networks: theoretical and empirical findings. *Social Networks*, 23(1):31 – 72, 2001.
12. H. Choi, S.H. Kim, and J. Lee. Role of network structure and network effects in diffusion of innovations. *Industrial Marketing Management*, 39(1):170–177, 2010.
13. N. Eagle, A. Pentland, and D. Lazer. Inferring social network structure using mobile phone data. *Proceedings of the National Academy of Sciences (PNAS)*, 106:15274–15278, 2009.
14. Nathan Eagle, Michael Macy, and Rob Claxton. Network diversity and economic development. *Science*, 328(5981):1029–1031, 2010.
15. Marta C. Gonzalez, Cesar A. Hidalgo, and Albert-Laszlo Barabasi. Understanding individual human mobility patterns. *Nature*, 453(7196):779–782, 06 2008.
16. M.S. Granovetter. The strength of weak ties. *American journal of sociology*, pages 1360–1380, 1973.
17. B.A. Huberman, D.M. Romero, and F. Wu. Social networks that matter: Twitter under the microscope. *First Monday*, 14(1):8, 2009.
18. Pavlos S. Kanaroglou, Michael Jerrett, Jason Morrison, Bernardo Beckerman, M. Altaf Arain, Nicolas L. Gilbert, and Jeffrey R. Brook. Establishing an air pollution monitoring network for intra-urban population exposure assessment: A location-allocation approach. *Atmospheric Environment*, 39(13):2399 – 2409, 2005. ¡ce:title¿12th International Symposium, Transport and Air Pollution¡/ce:title¿ ¡xocs:full-name¿12th International Symposium, Transport and Air Pollution¡/xocs:full-name¿.
19. Renaud Lambiotte, Vincent D. Blondel, Cristobald de Kerchove, Etienne Huens, Christophe Prieur, Zbigniew Smoreda, and Paul Van Dooren. Geographical dispersal of mobile communication networks. *Physica A: Statistical Mechanics and its Applications*, 387(21):5317 – 5325, 2008.
20. J. Leskovec, L. Backstrom, and J. Kleinberg. Meme-tracking and the dynamics of the news cycle. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 497–506. Citeseer, 2009.
21. Alan T. Murray, Kamyoung Kim, James W. Davis, Raghu Machiraju, and Richard Parent. Coverage optimization to support security monitoring. *Computers, Environment and Urban Systems*, 31(2):133 – 147, 2007.
22. V. Nicosia, F. Bagnoli, and V. Latora. Impact of network structure on a model of diffusion and competitive interaction. *EPL (Europhysics Letters)*, 94:68009, 2011.
23. G. Palla, A.L. Barabasi, and T. Vicsek. Quantifying social group evolution. *Nature*, 446(7136):664–667, 2007.
24. Wei Pan, Nadav Aharony, and Alex Pentland. Composite social network for predicting mobile apps installation. In *Proceedings of the 25th Conference on Artificial Intelligence (AAAI)*, pages 821 – 827, 2011.
25. R. Puzis, Y. Altshuler, Y. Elovici, S. Bekhor, Y. Shiftan, and A.S. Pentland. Augmented betweenness centrality for environmentally-aware traffic monitoring in transportation networks.
26. B. Waclaw. Statistical mechanics of complex networks. *Arxiv preprint arXiv:0704.3702*, 2007.
27. S. Wassermann and K. Faust. Social network analysis: Methods and applications. *New York*, 1994.