Generative Models for Rapid Information Propagation

Kirill Dyagilev
Electrical Engineering
Technion
Haifa 32000, Israel
kirilld@tx.technion.ac.il

Shie Mannor
Electrical Engineering
Technion
Haifa 32000, Israel
shie@ee.technion.ac.il

Elad Yom-Tov
IBM Haifa Research Lab
Haifa, Israel
yomtov@il.ibm.com

ABSTRACT

We consider the dynamics of rapid propagation of information in complex social networks focusing on mobile phone networks. We introduce two models for an information propagation process. The first model describes the temporal behavior of people which leads to the emergence of information propagation events and is based on the existence of two types of subscribers: regular subscribers and subscribers that tend to spread information. The second model describes the topology of paths in which the information propagates from one subscriber to another. We further introduce an efficient algorithm for identification of information propagation events. We then apply our algorithm to a large-scale mobile phone network and demonstrate the correspondence between theoretical expectations and the actual results.

Categories and Subject Descriptors

H.3.m [**Information Storage and Retrieval**]: Miscellaneous

General Terms

Algorithms; Measurement; Theory

Keywords

Information cascades, Telecommunication Networks

1. INTRODUCTION

Until recently, studying social phenomena was limited by the available information, which had to be collected through laborious manual work and personal interviews. The possibility for studying social interactions through electronic records such as web interactions and telephone call records has, in the last few years, opened this field to the study of much larger populations in different parts of the world. Electronic records are also highly advantageous in that they overcome the need for self-reporting of interactions, which is frequently biased, and caused an inherent problem in previous studies. In this paper we study social phenomena using mobile phone operator data, which pertains to a population of over two million people.

The accessibility of large-scale social data lead to an explosion of research in the field of complex networks in general and in the field of social networks in particular [13, 9]. For instance, using specially-designed smart phones, Eagle et.al. [4] showed how social networks could be reconstructed from the locations and interactions between their users. Goldenberg et.al. [5] used web data to show how social structures affect the adoption of new content. In [8] and [7] large–scale telecommunication data were used to design a marketing campaign and to build a fraud detection algorithm correspondingly. Finally, recent work [10, 11] used call data records (CDRs), which record the calling and called numbers of telecom subscribers, to show how social interactions relate to churn in mobile telephone networks.

Identification of people who tend to initiate information propagation events has so far received little attention. The role of hubs in disseminating information and in adoption patterns was investigated in [5], but as a static process, that is, as the overall effect of well-connected people on their peers, without regard to the interaction process itself. In this paper, we investigate the dynamics of this process, i.e., the actual sequences of information-passing events, which lead to a change in peoples' behavior. Our analysis is thus not limited to specific people who have an important role in the network, but instead shows how interactions between people leads to information dissemination.

Beyond the interest in the social phenomena, the identification of information dissemination events has many useful applications, ranging from deciding on which people should be approached for more successful advertising campaigns to identifying the sources of information of news items and web content.

In this work we focus on rapid propagation of information in the sense that once it is received, it is either transferred to somebody else during a relatively short period of time or will not be transferred to anyone. We refer to this mode of information propagation as *gossip*. We emphasize that usage of the term "gossip" does not imply any specific content of the propagating information but only that is spreads rapidly. We present evidence that gossip propagation processes transfer geographical information.

We take a machine learning perspective in this work and try to learn "simple" parametric models that describe well

1st Workshop on Social Media Analytics (SOMA '10), July 25, 2010, Washington, DC, USA.

the information propagation processes. Our data are a collection of millions of call records that we view as a large heterogenous network where random links form and dissipate. The main challenge from a learning perspective is to explain the dynamics from a statistical perspective in a meaningful way without producing very complex models with many parameters.

We provide two generative models of gossip propagation in mobile phone networks that have a relatively small number of parameters. These models are local, namely, they describe the behavior of a small number of subscribers rather than of the entire mobile call graph. The first model generates sequences of calls that yield an emergence of gossip propagation event. Similar to [6] in this model we assume that there exist two types of people, and these are reflected in their behavior as mobile telephone subscribers: regular subscribers and subscribers that tend to spread information. We refer to the latter type as *gossip–leaders* and assume that only these subscribers can initiate gossip propagation. The second model generates information propagation paths, namely, the ways in which the gossip propagates from one subscriber to another.

We further introduce an algorithm for identification the gossip propagation processes, and test this algorithm on large–scale real–world data.

We analyze properties of our models and show that they explain well the properties of gossip propagation processes we identify. Specifically, the first model predicts well the temporal structure of over 50 percent of gossip propagation events. As the regular calls may be modeled by the standard methods (e.g., [9]) we note that our models explain over 70 percent of daily calls. The second model predicts well the topology of information paths in over 85 percent of gossip propagation events. Of course, to explain additional observed phenomena, we will have to create more complex models. Our work is therefore only a first step in learning probabilistic models that describe the dynamics of large scale cellular networks.

The rest of the paper is organized as follows. In Section 2 we present the two generative models for gossip-propagation processes and discuss their properties. Section 3 introduces an algorithm for identification of gossip-propagation processes. In Section 4 we provide experimental results of this algorithm to large–scale call data records. We show that our models accurately predict properties of above 50 percent of gossip-propagation processes found. We conclude with a summary and discussion of future work in Section 5.

2. PROPAGATION MODELS

In this section we propose two generative models that address two different aspects of the gossip propagation event. The Day Generation Model describes the emergence of sequences of calls that capture the information propagation phenomenon. The Information Flow Tree Model describes the emergence of different topologies of an information flow. Namely, it generates a directed graph that shows paths in which the information propagates from one subscriber to another. The question of a finding a unifying model that would address both of the aspects above is beyond this paper.

Both of our generative models build on a well-known distribution - the Discrete Gaussian Exponential (DGX) - that appears in many contexts of social research [2]. In Section 4 we show that this distribution also fits well the real–world data available to us. For the benefit of the reader, we briefly review their definitions and basic properties in [1].

Some notations: we use $X \sim DGX(a,b)$ to denote that X is drawn from the DGX distribution with parameters a and b. We denote by $P_{DGX}\{\cdot|a,b\}$ and $F_{DGX}\{\cdot|a,b\}$ the corresponding probability distribution and cumulative probability function.

2.1 Day Generating Model

As mentioned, our first model introduces a stochastic process for generating sequences of calls that capture the rapid information propagation phenomenon. Specifically we focus on a mode of propagation of information called an *out–star* in which a single subscriber calls to the majority of the subscribers involved in the specific gossip propagation event. Heuristically we choose the threshold to be 80% of the subscribers.

This mode of propagation may seem too simplistic, however, we show in Section 4 that it includes over 50% of the gossip propagation events observed in the real–world data available to us. Moreover, it has the following "physical" model behind it. We assume the existence of two types of subscribers: regular subscribers and gossip-leaders. The first type includes most of the subscribers and exhibits a "passive" behavior with respect to information propagation, namely, the probability that a subscriber of this type will initiate a gossip or spread it further is very small. The second type contains a small fraction of subscribers that tend to initiate gossip propagation events and spread this information to other subscribers. This is inline with [5] who showed the role of hubs in the information diffusion process. We further assume homogeneity within types, namely, the behavior of all subscribers of the same type can be described as different realizations of a single stochastic process.

Under these assumptions, the most probable scenario of gossip propagation is that it is initiated by a gossip–leader who transfers the information to several regular subscribers. The propagation process then stops since regular subscribers do not tend to spread the information further, hence producing an out–star. We require that the information propagation will be rapid, namely, the time interval between two calls consequent calls of the gossip–leader in the gossip propagation event will be less than T. We further focus only on substantially large gossip propagation events, namely, we require that such an event contains at least K calls.

We now proceed with definition of several notations and introduction of the model. Let call c be described by the four–tuple $\langle n_1, n_2, t_{st}, t_{end}\rangle$, where n_1 is the number of the caller, n_2 is the number of the callee and t_{st} and t_{end} are beginning and end times of the call, respectively.

We introduce a generative model of all calls made by a certain gossip–leader during the day. We expect some of these calls to group to form a gossip–propagation event. Namely, there may be a sequence of at least K calls so that each two consecutive calls occur within less than T minutes from each other.

We further make the following two approximations in this model. First, we assumed that beginning time of the first call. Second, we neglect the effect of the social circle, namely, the gossip leader has a limited number of subscribers in his phone book. We note that the first approximation do not affect the structural properties of the information propagation processes we are interested in. However, the effect of

the social circle might be quite significant. These issues are beyond the scope of this paper.

The inputs of the Day Generating Model are the following: (1) The set R of regular subscribers and the set G of gossip–leaders. (2) Parameters a, b, a_d, b_d, α_a, α_b, β_a, β_b, γ_a and γ_b of the probability distributions involved in the model below.

The call sequence for each gossip–leader $g \in G$ is generated according to the following process:

1. Generate the number N of calls according to the distribution $DGX(a, b)$.
2. Draw the identities $\{r_i\}_{i=1}^{N}$ of N regular subscribers according to a uniform distribution without repetitions over the set R.
3. Generate the beginning time t_{st}^1 of the first call according to a uniform distribution over the day.
4. For $i = 1, 2, ..., N$
   - Generate the duration time $t_{dur} \sim DGX(a_d, b_d)$ of call c_i. Let end time of the call c_i be $t_{end}^i = t_{st}^i + t_{dur}$.
   - Add a call $c_i = \langle g, r_i, t_{st}^i, t_{end}^i \rangle$ to the generated sequence.
   - Generate time interval $\Delta t_i \sim DGX(\alpha_a N^2 + \beta_a N + \gamma_a, \alpha_b N^2 + \beta_b N + \gamma_b)$ (in minutes) between calls c_i and c_{i+1}.
   - Let the beginning time of the next call be $t_{st}^{i+1} = t_{end}^i + \Delta t_i$.

The following proposition considers out–stars generated by a single gossip–leader according to the Day Generation Model. It assesses the first two moments of the number of out–stars of a specific size and of the total number of out–stars.

PROPOSITION 1. *Consider calls generated by a single gossip–leader by the Day Generation Model with parameters as above. Let S_k for $k \geq K$ be the number of generated out–stars containing exactly k calls and let S denote the total number of out–stars. Further, let $p(n) \equiv F_{DGX}\{T|\alpha_a N^2 + \beta_a N + \gamma_a, \alpha_b N^2 + \beta_b N + \gamma_b\}$ be the probability that an interval between two consequent calls in less than T minutes. Then:*
(1) The expected value of S_k is given by

$$\mathbb{E}\{S_k\} = \sum_{n=K}^{\infty} P_{DGX}\{n|a,b\} A_{k-1}(n-1; p(n)),$$

where

$$A_i(m,\gamma) = \gamma^i \mathbb{I}_{\{m=i\}} + \gamma^i (1-\gamma)\left[2 + (m-i-1)(1-\gamma)\right] \mathbb{I}_{\{m>i\}}$$

and $\mathbb{I}_{\{\cdot\}}$ is an indicator function.
(2) The second moment of S_k is given by

$$\mathbb{E}\left\{S_k^2\right\} = \sum_{n=K}^{\infty} P_{DGX}\{n|a,b\} B_{k-1}(n-1; p(n)),$$

where

$$\begin{aligned} B_i(m,\gamma) =& \gamma^i \mathbb{I}_{\{m=i\}} + 2(1-\gamma)\gamma^i \mathbb{I}_{\{m>i\}} \\ &+ (m-i-1)(1-\gamma)^2 \gamma^i \mathbb{I}_{\{m\geq i+2\}} + e(i,m,\gamma), \end{aligned}$$

and the term $e(i, m, \gamma)$ is significantly smaller than the other terms. The exact value of $e(i, m, \gamma)$ may be found in [1].
(3) The expected value of S is given by

$$\mathbb{E}\{S\} = \sum_{n=K}^{\infty} P_{DGX}\{n|a,b\} C(n-1; p(n)),$$

where

$$\begin{aligned} C(m;\gamma) =& \left[\gamma^m - G(K-1, n-1; \gamma)\gamma(1-\gamma)\right] \mathbb{I}_{\{m\geq K-1\}} \\ &(2 + (1-\gamma)(m-1)) \left(\gamma^{K-1} - \gamma^{m+1}\right) \mathbb{I}_{\{m\geq K-1\}}, \end{aligned}$$

and

$$G(a,b;\epsilon) \triangleq \sum_{i=a}^{b} i\epsilon^i = \frac{(b+1)\epsilon^b - b\epsilon^{b+1} - (a+1)\epsilon^a + a\epsilon^{a+1}}{(1-\epsilon)^2}.$$

(4) The second moment of S is given by

$$\mathbb{E}\left\{S^2\right\} = \sum_{n=K}^{\infty} P_{DGX}\{n|a,b\} D(n-1; p(n)),$$

where

$$\begin{aligned} D(m;\gamma) =& \gamma^{K-1} \mathbb{I}_{\{m\geq K-1\}} \\ &+ (1-\gamma) \cdot \sum_{i=1}^{m} \gamma^{i-1} D(m-i-1;\gamma) \mathbb{I}_{\{m\geq K-1\}} \\ &+ 2(1-\gamma) \cdot \sum_{i=K}^{m} \gamma^{i-1} \overline{H}(m-i-1;\gamma) \mathbb{I}_{\{m\geq K\}}, \end{aligned}$$

and

$$\begin{aligned} &\overline{H}(m; p(n)) = p^{K-1}(n) \mathbb{I}_{\{m\geq K-1\}} + (1-p(n)) \\ &\cdot \sum_{i=0}^{m} p^i(n) \overline{H}(m-i-1; p(n)) \mathbb{I}_{\{m\geq K-1\}}. \end{aligned} \quad (1)$$

Proof: We begin by considering some fixed number $N = n$ of calls made by a gossip–leader. In what follows, we calculate the values of $\mathbb{E}\left\{S_k \middle| N = n\right\}$, $\mathbb{E}\left\{S_k^2 \middle| N = n\right\}$, $\mathbb{E}\left\{S \middle| N = n\right\}$ and $\mathbb{E}\left\{S^2 \middle| N = n\right\}$. Then Statements 1 to 4 will follow by the towering property of conditional expectations. We further note that all conditional expectations above equal 0 for $n < K$, thus we focus on $n \geq K$ in the rest of the proof.

This proof proceeds through the following steps. In Step 1 we introduce several auxiliary binary random variables and investigate their properties. In Steps 2 to 4 we show that Statements 1 to 3 follow almost trivially from the properties of the above variables. We conclude with the proof of Statement 4 in Step 5.

Step 1: Let $\{c_i\}_{i=1}^{n}$ denote the calls made by the gossip leader indexed in the chronological order. We denote by l_i for $i = 1, ..., n-1$ the binary variables so that $l_i = 1$ if the length of the time interval between calls c_i and c_{i+1} is smaller than T minutes and $l_i = 0$ otherwise. We note that the variables l_i are i.i.d. and $\mathbb{P}\left\{l_1 = 1 \middle| N = n\right\} = p(n)$. For notational convenience we extend our definition so that $l_i = 0$ almost surely for $i = 0$ and $i > n$.

Consider $n \geq k$. Let F_i^k for all $i = 0, 1, ..., n-k$ be an auxiliary binary variable equal to 1 if calls $\{c_j\}_{i+1}^{i+k}$ form an out–star of size k. It can be seen that $F_i^k = 1$ holds if $l_{i+1} = l_{i+2} = ... = l_{i+k} = 1$ and $l_i = l_{i+k+1} = 0$. Hence, $\mathbb{E}\left\{F_i^k \middle| N = n\right\} = \mathbb{P}\left\{F_i^k = 1 \middle| N = n\right\} = p^{k-1}(n)(1-p(n))^2$ for $n \geq k+2$ and $i = 1, 2, ..., n-k-1$. Here the equality holds almost surely as in all equations below. Due to dependence on l_0 and possibly l_{n+1} the expression for $\mathbb{E}\left\{F_0^k \middle| N = n\right\}$ is slightly different: $\mathbb{E}\left\{F_0^k \middle| N = n\right\} = p^{k-1}(n)(1-p(n))$ for $n > k$ and $\mathbb{E}\left\{F_0^k \middle| N = n\right\} = p^{k-1}(n)$ for $n = k$. Similarly, due to dependence on l_{n+1} we obtain that $\mathbb{E}\left\{F_{n-k}^k \middle| N = n\right\} =$

$p^{k-1}(n)(1-p(n))$ for $n \geq k+1$.

Step 2: We proceed to derive an expression for $\mathbb{E}\{S_k|\, N=n\}$. We note that $\mathbb{E}\{S_k|\, N=n\} = 0$ for $n < k$. For $n \geq k$ it holds that $S_k = \sum_{i=0}^{n-k} F_i^k$, hence, $\mathbb{E}\{S_k|\, N=n\} = \sum_{i=0}^{n-k} \mathbb{E}\left\{F_i^k \middle|\, N=n\right\}$. Substituting the expressions for $\mathbb{E}\left\{F_i^k \middle|\, N=n\right\}$ and rearranging the terms we obtain Statement 1.

Step 3: Similarly to Step 2, $\mathbb{E}\left\{S_k^2 \middle|\, N=n\right\} = 0$ for $n < k$. For $n \geq k$, it holds that

$$\mathbb{E}\left\{S_k^2 \middle|\, N=n\right\} = \sum_{i=0}^{n-k}\sum_{j=0}^{n-k} \mathbb{E}\left\{F_i^k F_j^k \middle|\, N=n\right\}. \quad (2)$$

In order to conclude the proof of the statement we need to calculate $\mathbb{E}\left\{F_i^k F_j^k \middle|\, N=n\right\}$ for different values of i and j. We note that $\mathbb{E}\left\{(F_i^k)^2 \middle|\, N=n\right\} = \mathbb{P}\left\{F_i^k=1 \middle|\, N=n\right\} = \mathbb{E}\left\{F_i^k \middle|\, N=n\right\}$ for all $i = 0,1,...,n-k$, hence was calculated in the proof of Statement 1. For $0 < j-i < k+1$, the events $F_j^k = 1$ and $F_i^k = 1$ cannot occur together as the former requires $l_{j-1} = 0$ and the latter requires $l_{j-1} = 1$, therefore, $\mathbb{E}\left\{F_i^k F_j^k \middle|\, N=n\right\} = 0$. For $j-i = k+1$ the events $F_j^k = 1$ and $F_i^k = 1$ share the requirement that $l_{j-1} = 0$, hence, $\mathbb{E}\left\{F_i^k F_j^k \middle|\, N=n\right\} = \mathbb{P}\left\{F_i^k=1 \middle|\, N=n\right\}\mathbb{P}\left\{F_j^k=1 \middle|\, N=n\right\} / (1-p(n))$. For $j-i > k+1$, the events $F_j^k = 1$ and $F_i^k = 1$ are independent, hence

$$\mathbb{E}\left\{F_i^k F_j^k \middle|\, N=n\right\} = \mathbb{E}\left\{F_i^k \middle|\, N=n\right\}\cdot\mathbb{E}\left\{F_j^k \middle|\, N=n\right\}.$$

Substituting to 2 yields Statement 2.

Step 4: We further note that $\mathbb{E}\{S|\, N=n\} = \mathbb{E}\left\{\sum_{k=K}^{n} S_k \middle|\, N=n\right\}$ and obtain Statement 3 by substituting result of Statement 1 and rearranging the terms.

Step 5: The second moment of S can be calculated using an argument similar to the proof of Statement 2. We note that

$$\mathbb{E}\left\{S^2 \middle|\, N=n\right\} = \sum_{k=K}^{n}\sum_{o=K}^{n}\sum_{i=0}^{n-k}\sum_{j=0}^{n-o} \mathbb{E}\left\{F_i^k F_j^o \middle|\, N=n\right\}. \quad (3)$$

We can obtain a closed–form expression for $\mathbb{E}\left\{S^2 \middle|\, N=n\right\}$ by calculating $\mathbb{E}\left\{F_i^k F_j^o \middle|\, N=n\right\}$ for different values of k, o, i and j and substituting these values to equation 3. However, the resulting expression contains a large number of terms and virtually unusable, therefore we take a slightly different approach.

Let $H(m,p(n))$ be the number of out–stars emerging among calls $\{c_i\}_{i=n-m}^{n}$ given that either $n-m=1$ or calls c_{n-m-1} and c_{n-m} are distanced by more than T minutes, i.e., $l_{n-m-1} = 0$. Let $\overline{H}(m,p(n)) = \mathbb{E}\{H(m,p(n))\}$.

It can be easily seen that $\mathbb{E}\{S|\, N=n\} = \overline{H}(n-1,p(n))$ and $\mathbb{E}\left\{S^2 \middle|\, N=n\right\} = \mathbb{E}\left\{H^2(n-1,p(n))\right\}$. We note that $\mathbb{E}\{H(m;p(n))\} = 0$ for all $m < K-1$. For $m = K-1$, the only realization in which an out–star can emerge is that if intervals between all calls are less than T, namely, $l_i = 1$ for all $i = n-m,...,n-1$, hence $\mathbb{E}\{H(m;p(n))\} = p^{K-1}(n)$.

Now consider $m \geq K-1$. We define events $E_{i,m}$ for $i = 1,..,m+1$ in the following way: $E_{1,m} = \{l_{n-m} = 0\}$; $E_{i,m} = \cap_{j=n-m}^{n-m+i-2}\{l_j = 1\}\cap\{l_{n-m+i-1} = 0\}$ for $i = 2,...,m$; and $E_{m+1,m} = \cap_{j=n-m}^{m}\{l_j = 1\}$. It can be easily seen that $\mathbb{P}\{E_{i,m}\} = (1-p(n))p^{i-1}(n)$ for all $i = 1,..,m$ and $\mathbb{P}\{E_{m+1,m}\} = p^m(n)$. We note that these events partition the sample space, namely, they are disjoint and their union covers the sample space. Hence, by the complete probability formula:

$$\mathbb{E}\{H(m,p(n))\} = \sum_{i=1}^{m+1} \mathbb{E}\{H(m,p(n))|\, E_{i,m}\}\,\mathbb{P}\{E_{i,m}\}. \quad (4)$$

For $i < K$ it holds that $\mathbb{E}\{H(m,p(n))|\, E_{i,m}\} = \overline{H}(m-i,p(n))$, for $K \leq i \leq m$ (if exist) it holds that

$$\mathbb{E}\{H(m,p(n))|\, E_{i,m}\} = \overline{H}(m-i-1,p(n)) + 1,$$

finally, for $i = m+1$ it holds that $\mathbb{E}\{H(m,p(n))|\, E_{i,m}\} = 1$. Substitution of these observations to (4) yields (1). We thus obtained an additional formula for $\mathbb{E}\{S\}$, namely,

$$\mathbb{E}\{S\} = \sum_{n=K}^{\infty} P_{DGX}\{n|a,b\}\overline{H}(n-1;p(n)).$$

We now apply the same argument to $\mathbb{E}\left\{H^2(m,p(n))\right\}$. Let $D(m,p(n)) = \mathbb{E}\left\{H^2(m,p(n))\right\}$. By the the complete probability formula:

$$D(m,p(n)) = \sum_{i=1}^{m-1} \mathbb{E}\left\{H^2(m,p(n))\middle|\, E_{i,m}\right\}\mathbb{P}\{E_{i,m}\}. \quad (5)$$

For $i < K$ it holds that $\mathbb{E}\left\{H^2(m,p(n))\middle|\, E_{i,m}\right\} = D(m-i,p(n))$, for $i = m+1$ it holds that $\mathbb{E}\left\{H^2(m,p(n))\middle|\, E_{i,m}\right\} = 1$, finally, for $K \leq i \leq m$ (if exist) it holds that

$$\begin{aligned}\mathbb{E}\left\{H^2(m,p(n))\middle|\, E_{i,m}\right\} &= \\ \mathbb{E}\left\{(1+H(m-i-1,p(n))^2\middle|\, E_{i,m}\right\}& \\ = 1 + 2\overline{H}(m-i-1,p(n)) + D(m-i-1,p(n)).&\end{aligned}$$

Substitution of these observations to (5) concludes the proof of Statement 4. □

As mentioned, Proposition 1 considers a single gossip–leader. The following corollary generalizes its result to multiple gossip–leaders and provides an estimate to the size distribution of generated out–stars.

COROLLARY 2.
1. Denote by $\widetilde{S_k}$ the number of out–stars of size k generated by $|G|$ gossip–leaders. Denote by $\widetilde{S}$ the total number of out–stars generated by $|G|$ gossip–leaders. Then $\mathbb{E}\left\{\widetilde{S_k}\right\} = |G|\,\mathbb{E}\{S_k\}$, $\mathbb{E}\left\{\widetilde{S_k}^2\right\} = |G|^2\,\mathbb{E}\{S_k^2\}$, $\mathbb{E}\left\{\widetilde{S}\right\} = |G|\,\mathbb{E}\{S\}$ and $\mathbb{E}\left\{\widetilde{S}^2\right\} = |G|\,\mathbb{E}\left\{S^2\right\}$.

2. The ratio $\widetilde{S_k}/\widetilde{S}$ converges to $\mathbb{E}\{S_k\}/\mathbb{E}\{S\}$ almost surely as $|G| \to \infty$[1].

Proof: See [1] for the complete proof. □

2.2 Information Flow Tree Model

The second model considers paths in which the gossip propagates from the source subscriber to all the others. Namely, it generates the directed graph $\mathcal{G} = (V,E)$ in which the set V of nodes contains all subscribers that eventually received the information and the set E of edges contains all directed edges $n_1 \to n_2$ so that the subscriber n_1 was chronologically

[1]Hence, when the number of gossip–leaders is large enough, the fraction $\mathbb{E}\{S_k\}/\mathbb{E}\{S\}$ can be used as an estimate for the empirical frequency of out–stars of size k among all out–stars.

the first to pass the information to n_2. We refer to the node n^r that initiated the information propagation as the *root*.

We consider only the first time the information is received by a node, hence the in–degree of each node is 1, except for the root n^r that has a degree of 0. It can be easily seen from the definition of the graph that the information flow graph is a directed tree.

This model requires the following parameters as its input:(1) the set of subscribers N; (2) parameters a_0, b_0 of the root degree distribution; and (3) parameters $a_{h',r'}$ and $a_{h',r'}$ for $h' = 1, 2$ and $r' = 1, 2, 3$ of degree distribution of other nodes.

The information flow tree is generated in a layer-by-layer fashion. We begin from the layer 0 that contains a single node which is the root. At each stage we go over all nodes in the current layer, generate their out-degrees and place the corresponding number of nodes in the next layer. This procedure is repeated for the next layer as long as it is non-empty. The probability distribution of the degree r of the root is given by $p_0(\cdot) = DGX(a_0, b_0)(\cdot)$. The degree d of a node in layer $h \geq 1$ is generated through a two step process. We begin by generating an auxiliary variable d' according to a probability distribution $DGX(a_{h',r'}, b_{h',r'})$, where

$$h' = \min\{h, 2\} \text{ and } r' = \min\{r, 3\}. \tag{6}$$

We then set $d = d' - 1$. The "truncation" of h and r was observed empirically and the variable d' is introduced in order to give a positive probability to the event $d = 0$. The identities of the nodes are generated according to a uniform distribution over N without repetitions. We note that the resulting tree may contain less than M subscribers, in this case, the process may be repeated from the start. Here is the complete description of the model:

1. Generate the identity n^r of the root according to uniform distribution over N. Generate the degree r of the root according to $DGX(a_0, b_0)$ and generate r distinct subscribers $\{n_i\}_{i=1}^r$ according to a uniform distribution over $N \setminus \{n^r\}$.
2. Let $L_1 = \{n_1, ..., n_r\}$, $L_2 = \{\}$ and $h = 1$. Update the sets of nodes and edges as follows: $V = \{n^r\} \cup L_1$ and $E = \cup_{i=1}^r \{n^r \to n_i\}$.
3. For each subscriber $n \in L_h$ do the following:
   - Generate the auxiliary variable d' of the current subscriber according to $DGX(a_{h',r'}, b_{h',r'})$, where h' and r' are defined in (6). Set the degree d of the current node n to $d = d' - 1$. If $d = 0$ then continue to the next node in L_h.
   - If $d > 0$ then generate d distinct subscribers $\{n_i\}_{i=1}^d$ for $i = 1, .., d$ according to a uniform distribution over $N \setminus V$. Add new nodes to the next layer, namely, set $L_{h+1} \leftarrow \cup_{i=1}^d \{n_i\} \cup L_{h+1}$. Update the sets of nodes and edges as follows: $V \leftarrow \cup_{i=1}^d \{n_i\} \cup V$ and $E \leftarrow \cup_{i=1}^d \{n \to n_i\} \cup E$.
4. Set $h \leftarrow h+1$ and $L_{h+1} \leftarrow \{\}$. Go to step 3 if the new current set L_h is not empty.

We divide the generated trees into the following classes based on their topology.
(1) *Topology 1:* Pure stars, namely, trees in which the root has an out–degree of at least $M - 1$ and rest of the nodes are leafs, i.e., have an out–degree of 0;
(2) *Topology 2:* Trees in which the out–degree of the root is 1, the out–degree of its only child is at least $M - 2$ and the rest of the nodes are leafs.
(3) *Topology 3:* Trees in which the out–degree of the root is at least $M - 2$, exactly one of its children has an out–degree of 1 and the rest of the nodes are leafs.
(4) *Topology 4:* Strings, namely, trees in which all the nodes has an out–degree of 1 except for a single node which is leaf.
(5) *Topology 5:* The out–degree of the root is at least 2, out–degrees of at least two additional nodes is exactly 1 and the rest of the nodes are leafs.
(6) *Topology 6:* The rest of the trees.
GPCs of topologies 1 to 4 are illustrated in Figure 1. The significance of this specific topologies was observed empirically by applying clustering techniques to topologies of information flow in the gossip propagation events in the data.

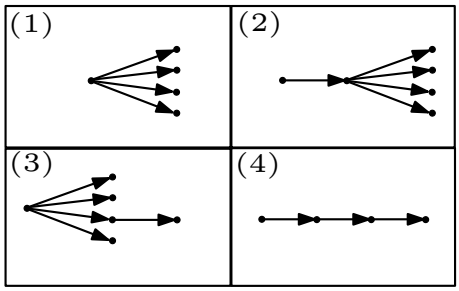

Figure 1: Topologies 1 to 4.

We proceed to investigate the properties of some of these topologies. Assume that $M \geq 4$ and consider a tree generated by the Information Flow Tree Model with parameters as above. Let N denote the total number of nodes in the tree and let E_i for $i = 1, 2, .., 4$ denote the event that the generated tree is of Topology i. The following proposition assesses one of the basic properties of these topologies, namely, the distribution $\mathbb{P}\{N = \cdot | E_1\}$ of sizes of GPCs in each topology.

PROPOSITION 3. *The size distribution of GPCs of Topologies 1 to 4 is given by the following expressions.*
(1) For Topology 1:

$$\mathbb{P}\{N = n | E_1\} = p_0(n-1)\left(p_{1,3}(0)\right)^{n-1} / \mathbb{P}\{E_1\} \mathbb{I}_{\{n \geq M\}},$$

where $\mathbb{P}\{E_1\} = \sum_{r=M-1}^{\infty} p_0(r)\left(p_{1,3}(0)\right)^r$.
(2) For Topology 2:

$$\mathbb{P}\{N = n | E_2\} = p_0(0) p_{1,1}(n-2)\left(p_{2,1}(0)\right)^{n-2} / \mathbb{P}\{E_2\},$$

where $\mathbb{P}\{E_2\} = \sum_{d=M-2}^{\infty} p_0(0) p_{1,1}(d)\left(p_{2,1}(0)\right)^d$.
(3) For Topology 3:

$$\mathbb{P}\{N = n | E_3\} = \frac{p_0(n-2)\left[d p_{1,3}(1)\left(p_{1,3}(0)\right)^{n-3}\right] p_{2,3}(0)}{\mathbb{P}\{E_3\}},$$

where $\mathbb{P}\{E_3\} = \sum_{d=M-2}^{\infty} p_0(d)\left[d p_{1,3}(1)\left(p_{1,3}(0)\right)^{d-1}\right] p_{2,3}(0)$.
(4) For Topology 4:

$$\mathbb{P}\{N = n | E_4\} = \left(p_{1,3}(1)\right)^{n-M}\left(1 - p_{1,3}(1)\right),$$

where $\mathbb{P}\{E_4\} = p_0(1) p_{1,1}(1)\left(p_{1,3}(1)\right)^{M-3} p_{1,3}(0)/(1-p_{1,3}(1))$.

Proof: These results can be easily shown by straightforward calculations. □

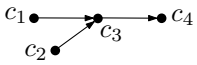

Figure 2: Illustrative T-causal line graph.

3. IDENTIFICATION OF GOSSIP PROPAGATION EVENTS

As mentioned, we assume that gossip spreads rapidly in the sense that once a subscriber receives it, he will either transfer it further to at least one subscriber within T minutes or will refrain from transferring it again. Once the subscriber finishes the transfer, it either transfers this gossip to an additional subscriber within the following T minutes or halts spreading it.

Equivalently, we consider a call c to be involved in gossip propagation if one of the following two conditions holds: (1) this call is the first call in gossip propagation; or (2) at least one of the subscribers was involved in gossip–propagation call in the last T minutes. Naturally, identification of a gossip propagation event requires finding all calls involved in it.

This observation is made rigorous in the following way. We say that call $c_j = \langle n_1^j, n_2^j, t_{st}^j, t_{end}^j \rangle$ is *T-causally connected* to call $c_i = \langle n_1^i, n_2^i, t_{st}^i, t_{end}^i \rangle$ if the following conditions hold: (1) these calls share at least one common subscriber; (2) c_i precedes c_j; and (3) the time interval between two calls is smaller than T. Namely, it holds that $\{n_1^i, n_2^i\} \cap \{n_1^j, n_2^j\} \neq \emptyset$ and $0 \leq t_{st}^j - t_{end}^i \leq T$.

Let $c_i = \langle n_1^i, n_2^i, t_{st}^i, t_{end}^i \rangle$ for $i = 1, .., k$ be a set of calls numbered in chronological order, namely, $t_{st}^i \leq t_{st}^{i+1}$ for all i. We call these calls *T-causally connected set* if every call c_j, $j > 1$, is *T-causally connected* to an earlier call c_i.

We define *Gossip Propagation Component* (GPC) as a *T-causally connected set* of calls that contains at least K calls and at least M distinct subscribers.

Our algorithm for identification of GPCs relies on a *T-causal line graph* that is defined in the following way.

DEFINITION 4. *Let $\mathcal{C}$ be a set of calls. The directed graph $G = (V, E)$ is a T-causal line graph corresponding to the set C if it has a vertex $v \in V$ for each call $c \in \mathcal{C}$ and a directed edge $v_i \rightarrow v_j$ for each pair of corresponding calls c_i and c_j so that c_j is causally connected to c_i.*

We extract GPCs by applying the DFS algorithm (e.g., see Cormen et.al. [3]) to the T-causal line graph and identifying each DFS tree that is large enough. We note that the same call may be partitioned to GPCs in several different ways. For instance consider the partition of the calls whose T-causal line graph is depicted in Figure 2 for $K = N = 4$. This CDR can be partitioned to DFS trees in several ways, including, $\{(c_1, c_3, c_4), c_2\}$, $\{(c_2, c_3, c_4), c_1\}$ and $\{(c_4), (c_3), (c_2), (c_1)\}$. We can not distinguish between the first two options as both of them may transfer information. However, the last decomposition is problematic because it breaks this call sequence to groups of sizes which are below the size threshold. We avoid such decompositions heuristically by choosing the starting point for the next DFS tree to be the chronologically earliest call that was not used is existing DFS trees. See Algorithm 1 for the complete details.

The GPCs provide us with the list of calls involved in the specific gossip propagation event. We proceed to extract the information paths from this list, namely, the routes in which information propagates from one subscriber to another. As mentioned in Section 2.2, the information paths graph is a tree, hence it is fully described by the identify of its root and parent-child relations of other nodes.

Algorithm 1 GPC identification

Input: Call data records, T
Build T-causal line graph $G = (V, E)$ that corresponds to the given CDR.
Initialize all nodes in G as unvisited.
repeat
  Choose the unvisited node with the smallest topological index.
  Run the DFS visiting algorithm starting from this node.
  Identify the obtained DFS tree as a GPC if it contains at least K calls and at least M subscribers.
until all nodes are visited

Our method proceeds as follows. The caller of the chronologically first call is assumed to be the origin of the gossip or root. The callee of this call is the child of the root. We note that both of these nodes posses the information after the first call. We continue by going over all other calls in a GPC according to a chronological order. In each new call either only one of the subscribers possesses the information or both of them. The latter case is irrelevant to information paths hence discarded. In the former case, the subscriber that receives the gossip is added as a child of the subscriber.

4. ANALYSIS OF REAL–WORLD DATA

Our experiments were performed on data from a large cellular operator. Specifically, we focused on calls logged by this operator in a city with population of over 2 million people in 24 days out of period from February 10th, 2008 to March 9, 2008. Each call is described by a Call Data Record (CDR), which contains its start and end times, as well as the obfuscated identity of the involved subscribers. Naturally, we have a record of a call only if either the caller or the callee used services of the analyzed operator. In cases where both subscribers were mobile, call details also contain the general geographical locations of the subscribers at the beginning of the call, given through cell identifiers.

The raw data contains more than 50 million calls involving 5.4 million distinct subscribers, out which approximately 2.07 million belong to the analyzed operator. The remaining subscribers belong to other service providers. Therefore we have only partial information on their calls. In the preprocessing phase we filtered out calls of all subscribers that are involved in more than 200 calls a day. We assumed that these subscribers were public service providers (police, etc.) and commercial payphones, which are less of interest to us in our investigation of information propagation events. There were detected less than 1000 subscribers of the above type and they produced overall of less than 2.5 million calls in the designated period.

This section proceeds as follows. In Section 4.1 we discuss the general properties of GPCs found in the data. Section 4.2 outlines our methodology for estimation of model parameters. Sections 4.3 and 4.4 discuss the goodness of fit of the Day Generation Model and the Information Flow Tree model respectively. Finally, in Section 4.5 we produce evi-

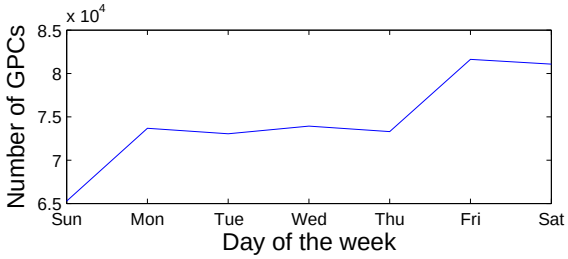

Figure 3: Number of GPCs by weekdays.

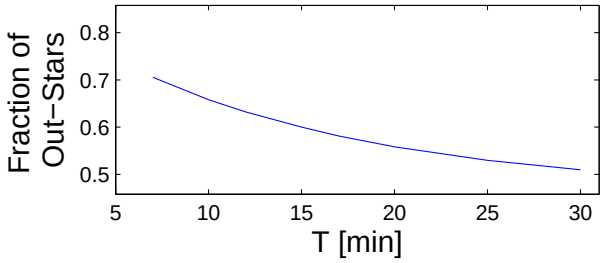

Figure 4: Fraction of out–stars among all GPCs.

dence that some of the GPCs transfer geographical data.

4.1 Structural Properties of GPCs

In this section we describe properties of GPCs found by our algorithm. We focus on GPCs that contain at least 4 calls that involve at least 4 unique subscribers. We further let the maximal time T between two consequent calls involved in propagation of a gossip be 20 minutes. This specific selection is arbitrary; furthermore, as indicated below, any value of T in the interval 7 minutes to 30 minutes would yield similar results. However, we note that the median time between two calls of the same subscriber is 35 minutes, hence it is not a typical behavior of a subscriber to produce several calls with a time interval smaller than 20 minutes between each other.

We emphasize that GPCs found by our algorithm need not have a single subscriber (gossip–leader) that calls all other subscribers, they may have a different structure. In fact, the found GPCs can be divided into the following three groups:

1. *Out–Star.* These GPCs have a dominant gossip–leader, namely, a subscriber that calls at least 80 percent of involved subscribers.
2. *In–Star.* This is an "inverted version" of the previous structure. Here, at least 80 percent of subscribers call one central node to (probably) get some information from him. A structure like this can be created by calls to some information center, e.g., traffic information service.
3. *Non–Star.* GPCs that do not fit in the previous two groups.

Figure 3 shows the average number of GPCs in each day of the week. As this figure shows, the number of GPCs depends on the day of the week: it has its smallest value on Sunday, has similar higher values on Monday to Thursday, and has its highest values on Friday and Saturday. This may due to higher levels of social activities towards the weekend.

The fraction of GPCs of each of the three types depends on the parameter T and does not vary greatly from day to day. For instance, Figure 4 depicts the fraction of out-stars among all GPCs as a function of T for February, 11 2008. Analyzing the identified GPCs, we find that over 50 percent of GPCs for $T \leq 30$ minutes are, in fact, out–stars. This lends support to the use of gossip–leader models such as the ones analyzed in this paper.

4.2 Estimation of Model Parameters

We begin by estimating the parameters for the Day Generating Model. We begin by calculating the empirical distribution of the number of calls made by gossip–leaders during a single day. The parameters of the DGX distribution used to generate the number of calls of a gossip–leader are found by fitting the empirical curve by DGX distribution in a log–log scale. The obtained DGX distribution fits with $\mathcal{R}^2 = 0.99$ which corresponds to a very good fit (see Shao [12] for the definition of $\mathcal{R}^2$).

The parameters α_a, α_b, β_a, β_b, γ_a and γ_b are estimated in the following two steps. In step 1, we consider a span of possible numbers N of calls made by a gossip–leader. For each value n of N we consider only gossip–leaders that make *exactly* n calls and calculate the empirical probability of time intervals between calls for these subscribers only. Using the same method as above we fit the DGX distribution to this empirical distribution. We denote the estimated parameters by a_n and b_n. In step 2, we fit series a_n and b_n with a quadratic function of n.

We proceed to estimate the parameters for the Information Flow Tree Model. We begin by calculating the empirical degree distribution of the root. We further estimate the empirical degree distribution $p_{r,h}(\cdot)$ of other nodes given the degree r of the root and node's depth h. We note that for $r > 3$ and any h the distributions $p_{r,h}(\cdot)$ and $p_{3,h}(\cdot)$ are similar, hence it suffices to consider only $r = 1, 2, 3$. Further, for any $h > 2$ and any r the distributions $p_{r,h}(\cdot)$ and $p_{r,2}(\cdot)$ are also similar, hence it suffices to consider $h = 1, 2$. The parameters of the model are estimated by fitting the DGX distribution to the corresponding empirical distribution. We find that DGX fits the degree distribution of the root with $\mathcal{R}^2 = 0.91$ and it fits distributions $p_{r,h}(\cdot)$ for $r = 1, 2, 3$ and $h = 1, 2$ with R-squared of at least 0.97.

In order to avoid overfitting we estimate the model parameters and check the quality of predictions made by our models on a data from separate days. However, it is important that these days would have a similar levels of social activity, therefore these days should be either both working days or both weekend days.

4.3 The fit of the Day Generating Model to the data

The property of out–stars we are interested in predicting the most is the distribution of their sizes. Figure 5 depicts the empirical distribution of sizes out–stars in terms of calls and in term of unique subscribers involved for $T = 20$ minutes. We note that the measured distribution of out–star sizes in terms of number of unique subscribers is significantly different from the other distributions. This discrepancy is attributed to the fact that the gossip–leader calls some of the subscribers more than once.

We proceed by generating out–stars according to the Day

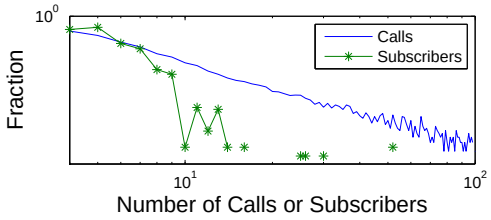

Figure 5: Distribution of Out-Stars Sizes

Generating Model using 40 thousand gossip leaders. The empirical distribution of sizes of the obtained out–stars fits the distribution in the real–world data with $\mathcal{R}^2 = 0.88$.

The threshold T is a parameter of the definition of out–stars and it is of interest to understand the dependence of the number of out–stars on T. The Day Generation Model allows us to predict the dependence between the number of out-stars and the value of parameter T. We note that the total number of out–stars is proportional to the size $|G|$ of the set of gossip–leaders, hence we normalize the simulated number of out–stars so it would equal to the measured one for $T = 20$ minutes. Figure 6 depicts the ratio of the normalized number of out stars predicted by the Day Generation Model and the number of out–stars found in the data. We note that for $7 \leq T \leq 30$ minutes we predict the number of out–stars with the precision of 5 percent.

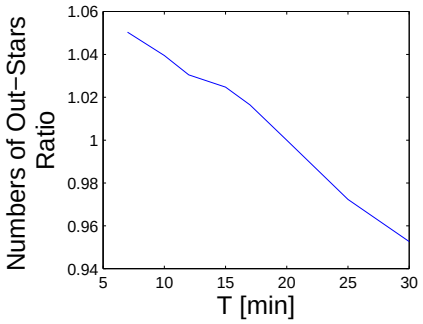

Figure 6: Ratio of number of out-stars found in the data to the normalized number of out-stars predicted by the Day Generation Model.

4.4 The fit of the Information Flow Tree Model to the data

In this section we focus on the quality of prediction of topological properties of information paths underlying the GPCs. In particular we consider topologies introduced Section 2.2.

We begin by checking the quality of prediction of fraction of GPCs belonging to each topology both generated and found in real–world data. In Table 1 listed the average fraction of GPCs of each topology over 17 working days in data and the average fraction of GPCs in generated data over 100 runs. The confidence of the values is measured by standard deviation time three. We note that the quality of prediction for Topologies $1-3$ and 5 is relatively high.

We further checked the quality of prediction of size and height distributions of all GPCs and GPCs of a specific topology. The generated distributions fit the data with R-squared of at least 0.93.

Topology	Data	Simulated
Topology 1	0.33 ± 0.02	0.29 ± 0.004
Topology 2	0.138 ± 0.01	0.102 ± 0.002
Topology 3	0.18 ± 0.01	0.189 ± 0.004
Topology 4	0.04 ± 0.004	0.011 ± 0.002
Topology 5	0.12 ± 0.01	0.126 ± 0.003
Topology 6	0.2 ± 0.02	0.281 ± 0.005

Table 1: Average fraction of GPCs belonging to each topology $\pm$ standard deviation times three.

4.5 Geographic evidence for information propagation

In this section we describe an experiment that provides evidence that GPCs represent information diffusion, by observing the relationship between the geospatial behavior of subscribers and the appearance, or lack therefore, of GPCs.

We consider 85,000 pairs p_i of subscribers so that there is exist a day on which they both appear in the same GPC. For each pair p_i, we choose randomly a single day D_i on which they both appear in the same GPC (there may be more than one day) and a single day $\tilde{D}_i$ on which they do not appear in the same GPC. We measure the probability of these subscribers to make or receive calls from the same location on D_i and on $\tilde{D}_i$.

The description of each call includes the number of cell in which s was located at the beginning of the call. The median radius of cells is only 0.7 kilometers, hence it defines rather accurately the geographic location of the subscriber at the time of the call. We can obtain an information about movements of the subscribers during days D_i and $\tilde{D}_i$ by considering all calls it participated in during the corresponding day. Let a_i and $\tilde{a}_i$ denote the number of cells **both** subscribers in p_i visited on the days D_i and $\tilde{D}_i$, respectively. Using Wilcoxon signed-rank test (e.g., see Shao [12]), we tested whether a_i and $\tilde{a}_i$ can be generated by the same probability distribution. The result was that the probability for this happening by chance were $p < 10^{-10}$. It can also be clearly seen from Figure 7 that the average value of a_i is larger than the average value of $\tilde{a}_i$. Namely, the appearance in the same GPCs increase the chances that two subscribers will visit the same geographic location during the day.

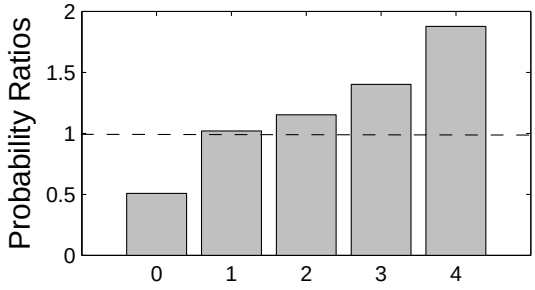

Figure 7: Ratio of the probability of the number of common cells for the day with common GPC to probability for the day without common GPC.

5. CONCLUSIONS

Identifying information propagation events makes it possible to track an important aspect of social interaction between subscribers in a mobile telephone network. Beyond the sociological interest, it may serve, alongside with identification of information propagation leaders, as an important preliminary step in choosing target audiences for a marketing campaign. We note that our choice of influential subscribers is based on dynamical social interactions rather than on topology of the underlying static graph of social connections. We believe that this approach can be extended to other social media. The identity of gossip–leaders may also prove to be a useful feature in various behavior prediction systems, e.g., in churn predictors.

In this work we focused on a rapid mode of information propagation - which we termed gossip. We provided two generative models: one for the temporal evolution of gossip propagation events and another for the topology of information propagation paths. These models are *local* in the sense that they describe the behavior of a small number of subscribers rather than of the whole graph of calls. We further introduced an algorithm for identification of gossip–propagation components, namely, sets of calls involved in gossip propagation. We applied our algorithm to large–scale real–world data and showed that our models provide a good description for the properties of a significant fraction (over 50% for one model and over 85% for another) of gossip-propagation components.

Further work of immediate interest includes finding a unifying model that would explain both the temporal evolution of the gossip propagation and the topology of the underlying information paths; modeling of an inter–day behavior of gossip–propagation leaders; incorporating the existence of the social circle in our analysis; and leveraging geographical information to get additional insights on the social structure of mobile call networks.

6. REFERENCES

[1] On–line appendix. `http://webee.technion.ac.il/people/kirilld/Preprints/KDD10_supplementary_material.pdf`.

[2] Z. Bi, C. Faloutsos, and F. Korn. The "DGX" distribution for mining massive, skewed data. In *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 17–26. ACM New York, NY, USA, 2001.

[3] T. Cormen, C. Leiserson, R. Rivest, and C. Stein. *Introduction to algorithms*. The MIT press, 2001.

[4] N. Eagle, A. Pentland, and D. Lazer. Inferring social network structure using mobile phone data. volume 106, pages 15274–15278, 2009.

[5] J. Goldenberg, S. Han, D. Lehmann, and J. Hong. The role of hubs in the adoption process. *Journal of Marketing*, 73(2):1–13, 2009.

[6] J. Goldenberg, B. Libai, S. Moldovan, and E. Muller. The NPV of bad news. *International Journal of Research in Marketing*, 24(3):186–200, 2007.

[7] S. Hill, D. Agarwal, R. Bell, and C. Volinsky. Building an effective representation for dynamic networks. *Journal of Computational and Graphical Statistics*, 15(3):584–608, 2006.

[8] S. Hill, F. Provost, and C. Volinsky. Network-based marketing: Identifying likely adopters via consumer networks. *Statistical Science*, pages 256–276, 2006.

[9] M. Jackson. *Social and Economic Networks*. Princeton University Press, Princeton and Oxford, 2008.

[10] A. Nanavati, S. Gurumurthy, G. Das, D. Chakraborty, K. Dasgupta, S. Mukherjea, and A. Joshi. On the structural properties of massive telecom call graphs: findings and implications. In *Proceedings of the 15th ACM international conference on Information and knowledge management*, page 444. ACM, 2006.

[11] Y. Richter, E. Yom-Tov, and N. Slonim. Predicting customer churn in mobile networks through analysis of social groups. In *To Appear in Proceedings of the 2010 SIAM International Conference on Data Mining (SDM 2010)*, 2010.

[12] J. Shao. *Mathematical Statistics*. Springer, New York, 2nd edition, 2003.

[13] F. Vega-Redondo. *Complex Social Networks*. Cambridge University Press, New York, 2007.