Hibernating Process: Modeling Mobile Calls at Multiple Scales

Siyuan Liu*, Lei Li†, Ramayya Krishnan*
*Carnegie Mellon University, †University of California, Berkeley
{siyuan, leili, rk2x}@cmu.edu

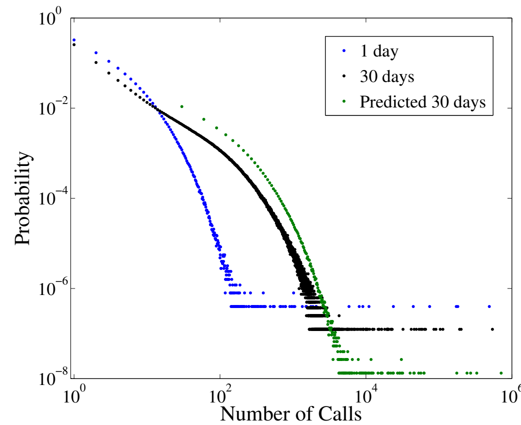
***Abstract*—Do mobile phone calls at larger granularities behave in the same pattern as in smaller ones? How can we forecast the distribution of a whole month's phone calls with only one day's observation? There are many models developed to interpret large scale social graphs. However, all of the existing models focus on graph at one time scale. Many dynamical behaviors were either ignored, or handled at one scale. In particular new users might join or current users quit social networks at any time. In this paper, we propose Hibernating Process (HiP), a novel model to capture longitudinal behaviors in modeling degree distribution of evolving social graphs. We analyze a large scale phone call dataset using HiP, and compare with several previous models in literature. Our model is able to fit phone call distribution at multiple scales with 30% to 75% improvement over the best existing method on each scale.**

***Keywords*-Mobile phone call graph, churning behavior, heavy tailed distribution, non-parametric model.**

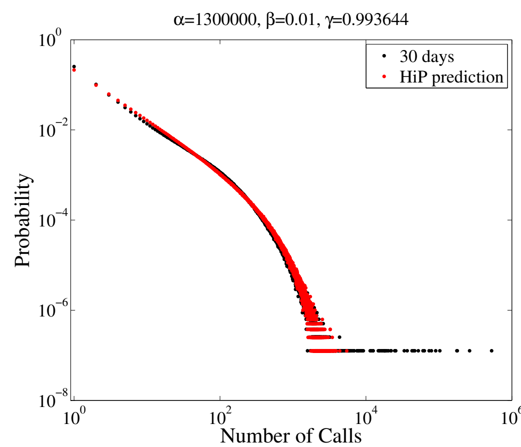
I. Introduction

Large scale social graphs emerge in many online and offline communication networks. As one specific social graph, mobile phone call networks receive more and more attentions recently [1], [2], [3], [4], [5], [6]. It is one of the key challenge to detect the latent structure and patterns hidden in the graph. One aspect is to examine the degree distribution of the nodes in the graph, which is previously discovered to follow a heavy-tailed distribution [7], [8]. To large extend, many patterns have been discovered about large social networks [9], [10], [11], [12]. For example, the degree of nodes often exhibits a certain heavy-tailed distribution (i.e. richer-get-richer). Various models have been proposed to model the behavior and in particular the degree distributions in a social network, such as generalized Pareto [12], [10], DPLN [13], [14], Lognormal [3], and Loglogistic [15], [4]. However, often those models focus on behaviors at one scale (e.g. the friends within a month), and assume the same pattern exist in all scale. However, we observe that often at different scales the macro behavior might be visually different. For example, the number of connections one made during a day might distribute differently from that of one month (this observation is reported in Figure 1(a)).

Therefore, can we predict a network's behaviors (e.g. the number of calls in a month) with or without observing one day's behavior? One naïve solution is "multiplying" by the scale factor. To illustrate the challenge here, we show that simply multiplying does not work in real data. Figure 1(a) shows distributions of the number of calls in one day and



(a) Mismatch of simple prediction by multiplying.



(b) HiP prediction for a month.

Figure 1. (a): Distribution of the number of calls for one day (in blue), one month (in black), and the predicted 30 days from one day's data (in green) by the simple multiplication rule. Note that simply multiplying one day data does NOT produce prediction for one month data. (b): The predicted distribution of the number of calls by HiP (in red) matches the real data of one month (in black). $N = 396486732$, $\alpha = 1300000$, $\beta = 0.01$ and $\gamma = 0.993644$. HIP generates number of phone calls completely from scratch (day zero) using the given parameters.

one month. A simple rule of thumb to predict a month from single day is to simply multiply the number of phone calls in a month is just that of one day by 30 (Eq.(1)).

$$\log p_{month}(x) = \log p_{day}(e^{\log x - \log 30}) - \log 30 \quad (1)$$

where x is number of calls, p_{day} is the probability density function (PDF) of one day's phone calls, and p_{month} is PDF of one month's. Visually it is equivalent to the moving of day's PDF curve (the blue line in Figure 1(a)) to rightward and downward by certain amount. However, such simple prediction does not produce a distribution close to reality as shown in Figure 1(a). Note that in Figure 1(a) the simulated 30 days' PDF has a big difference from the real 30 days' PDF, not only the shape of the curve, but also the placement. Previous studies show that at certain specific scale several models fit well, such as generalized Pareto [12], [10], Lognormal [3], Loglogistic [15], [4] and DPLN distribution [13], [14]. To date, there is no study applying these successful models on multiple scales. It is therefore previously unclear whether they would work or not at different scales. Our study however gives negative result (see experiment section for more details). As a comparison, without going to details, our proposed method HiP does surprisingly well in predicting phone call distribution with the input of the total number of calls in the month. Note HiP generates the number of calls for 30 days completely from scratch (day zero) using only three additional parameters (Figure 1(b))!

In our work, we aim to understand the process from quantitative change to qualitative change in social graphs (in particular mobile phone call graph) and then propose a model to capture desirable behaviors at different scales. Our proposed method is inspired by an interesting observation, called *churning* behavior, which is less noticed and studied in the previous social network models. After examining three large scale mobile social graphs from three different countries, with millions of mobile phones, billions of call records, as long as one year's duration, we obtain a surprising finding: the nodes in the mobile call graphs not only accumulate edges along the time, but also drop or retire from the graphs, which we define as "churning". The "churning behavior" may be caused by the quitting of users, burner users (prepaid short time users), inactive users or tourists in the city, etc. Given the limited time period of mobile social graph, it is very hard for us to retrieve who is really churning and going to churn. Apparently, mobile operators are very interested in the churning behavior, since they simply want to retain as more phones in their network and expect them to call as many as possible for more profit.

The contributions of this paper are below. First, we discover surprising patterns of churning nodes in social graphs. Second, we devise a novel model, Hibernating Process (HiP), for describing the dynamic evolving behavior in large scale graphs, and give the interpretation of the underlying generation process. As shown in Figure 1(b), the prediction by our model for one month is reported. There are only four numerical inputs to HiP, while it does surprisingly well and is able to generate about 400 million calls that matches real data. At last but not the least, we evaluate HiP on a large scale phone call dataset. HiP produce distributions of phone calls very close to real data in every cases and on multiple scales. Our method can achieve 75% improvement than the previous models.

II. Related work

Social graph mining: Rodriguez et. al [16] tried to infer networks of diffusion and influence. Leskovec et. al [17], [11], [18], [19], [20] studied graph over time by densification laws, shrinking diameters and possible explanations, and provided a graph generator based on a forest fire spreading process to study the graph evolution. McGlohon et. al [21] studied patterns in weighted graphs and proposed a generator. There are a category of recent studies on social networks evolution and growth [22], [23], [24], [25], [26], [27], [28]. Vaz de Melo et al. [4] investigated the patterns for the call duration distribution of mobile phone users, and proposed a Truncated Lazy Contractor distribution to fit the call duration distribution, which is a truncated version of log-logistic distribution. Seshadri et al. [13] observed some distributions (of number of calls, distinct call partners, nd total talk time), and proposed a Double Pareto LogNormal distribution to fit the data. In [5], the authors proposed a log-normal distribution to fit the call duration. In [29], the authors found that the call duration neither exponentially nor log-normally distributed, and the distribution has a semi-heavy tail, which asks for a more heavy-tailed distribution. In our work, we investigate the graph generated from mobile phone networks at multiple scales, and propose the non-parametric fitting and prediction method.

Parametric models for social network: The stable distribution family is sometimes referred to as the Lévy alpha-stable distribution [30]. The normal distribution is one family of stable distributions [31]. Reed et al. [14], Clauset et al. [8], and Newman [12] studied the heavy tail distribution, and proposed the distribution function, approximation method and generation mechanism. Fofack et al. [9] and Nolan [10] studied the tail behavior, modes, modeling and accurate computation way of stable distribution. In our work, we propose a non-parametric model to fit and predict in different time scale data, which cannot be accomplished by the above parametric models.

Non-parametric models for social network: Non-parametric model is a distribution-free method (e.g., Chinese Restaurant Process, India Buffet Process and Yule–Simon distribution), which does not rely on assumptions that the data are drawn from a given probability distribution. Blei et al. [32] presented the nested Chinese restaurant process and showed that this stochastic process can be used as a prior distribution in a Bayesian nonparametric model of document collections. Chen et al. [33] proposed a nonparametric Bayesian contextual focused topic model. All these methods model the birth process and accumulation of degree or weights, and produce richer-get-richer phenomenon. Our proposed model falls into this category, while we explicitly model the "churning" behaviors in communication network.

III. Mobile Phone Call Datasets

In our work, our study is based on a large scale mobile phone data set which is collected from a large city in country 1. The size of the city is around 8700 km^2. In this city, there are 2.5 million mobile phones and 15 million records per day. The size of the raw data set that we collected from 1^{st} January, 2008 to 31^{st} December, 2008, is around 0.7 Terabytes.

The data set is generated from the *Call Detail Record* (CDR) which is the information related to mobile phone communication, such as caller ID, callee ID, call start time and call duration. In the following study, to make our method and findings clear, we illustrate our work by one attribute of the mobile phone communication data, that is the number of calls. The number of calls is defined as the total number of calls per user in a given time interval. Based on a set of call records, we can construct a *mobile phone call graph*: each node is a mobile phone number and each edge is a phone call. The number of calls distribution is the the data distribution of all the users' call in a given time period (time scale). Time scale is defined as a time period that we observe the accumulated data. For example, a day means we observe the data by one day as the time unit. Our method is obviously not limited to the distribution of the number of calls, and can be easily adapted to other attributes.

In this paper, we emphasize that our interest is in aggregating statistical analysis and therefore, we do not study any particular individual's calling behavior. In order to preserve the user privacy and anonymity, data that could identify users, e.g., the phone numbers, is not utilized in this study.

IV. Proposed Model: HiP

In this section we propose a model for interpreting the increase and flattening of the number of phone calls. We describe how the number of phone calls for each user accumulate over time. Our goal is to develop a model that can generate phone calls that match the observed behaviors in real mobile phone call data. We also provide a fast generation algorithm and the parameter estimation for Hibernating Process (HiP).

Our proposed Hibernating Process (HiP) maintains a set of phones, which can be further divided into two subsets: active set (AS), and hibernated set (HS). The phones in AS can start a call, while those in the hibernation set will keep all the past calls but never start additional ones. HiP models three processes: the growing process of the size of AS, the accumulation process of calls from phones in AS, and the transition process of phones from AS to HS. Once a phone reaches HS, it will not change its state. More formally, HiP describes the following calling procedure:

1) Initially all phones make zero phone call, and both AS and HS are empty;
2) Randomly pick up a phone number, make a phone call and put this phone number to AS;

Algorithm 1: HiP

```
Input: N, α, β, γ
Output: A list F of pairs, ⟨c_i, f_i⟩ representing f_i
        phones making c_i calls
H ← ∅; A ← {1 → 1};
d ← 1; W ← {1 ← 1}; C ← 1;
for n ← 2 to N do
    generate random value r ~Uniform(0, 1);
    if r < (α+β·d)/(n+α) then
        increment A[1], W[1] and d by 1;
    else
        generate random value s ~Uniform(0, C);
        foreach ⟨c, f⟩ ∈ A ∧ s > 0 do
            s ← s − W[c];
            if s < 0 then
                increment A[c + 1] by 1;
                decrement A[c] by 1;
                increment W[c + 1] by c + 1;
                decrement W[c] by c;
    C ← C + 1;
    generate random value r ~Uniform(0, 1);
    if r > γ then
        generate random value s ~Uniform(0, d);
        foreach ⟨c, f⟩ ∈ A ∧ s > 0 do
            s ← s − f;
            if s < 0 then
                decrement A[c] and d by 1;
                decrement C and W[c] by c;
                increment H[c] by 1;
foreach c ∈ Keys of A ∪ H do  F[c] ← A[c] + H[c];
```

3) From then on, n-th time tick:
   - with probability $\frac{\alpha+\beta\cdot number\_distinct\_phone}{n+\alpha}$, there will be a new phone joining AS and making one call;
   - otherwise, with probability proportional to the number of calls for each phone already in AS, one phone will make one call;
4) Every time after making the i-th call, with probability $1-\gamma$, one random phone from AS will *hibernate* and move to HS (hence its number of calls will never increase);

From the above process, we can obtain the following results.

Lemma 1: At time tick n, the probability of retaining all phone is AS is γ.

Lemma 2: At time tick n, suppose the size of AS is d_n, the probability of phone i ∈AS hibernates is $\frac{1-\gamma}{d_n}$.

Lemma 3: If $\beta = 0$ and $\gamma = 1$, the expected number of distinct phones for N calls from HiP is $\sum_{i=1}^{N} \frac{\alpha}{\alpha+i-1}$.

How do we generate phone calls using the process (Hibernating Process) describe above? A naïve approach would be simulating the process one by one and keep the number of calls for each phone. However, it is extremely slow when we have a huge number of phone calls. The total number of calls from or to a mobile network in a month can reach 400 million in a normal month, and 1.17 billion in a quarter. The number of distinct phones in a month can also reach 8 million. With such big numbers it is inefficient to simulate using the straightforward approach. Hence, we provide a fast generation algorithm as below.

The full algorithm is described in Algorithm 1. It accepts four inputs: the total number of calls N and three model parameters α, β, and γ.

There are a few algorithmic techniques to achieve high performance, namely

- Using two hash tables A (for active set) and H (for hibernated set) to record the number of phones who made c calls, denoted as $f_{n,c}$;
- Maintaining partially computed probability and accumulative probabilities using binary indexed trees (or Fenwick tree [34]) and updating incrementally in Step 9 and 20 of Alg.1. Note that updating the accumulative sums of weights in W using binary indexed trees cost $O(\log |W|)$;

V. Experimental Results and Analysis

A. Experiment Setup

We observe mobile phone data at four different time scales: day data, one week data, one month data, and three months data and find out a clear pattern which is very different from the pattern we showed in Figure 1(a), that is, the longer time period data cannot be generated from the shorter time period data by the intuitive method in Eq. 1. But our non-parametric model, Hibernating Process, is able to not only fit different time scales data, but also have the power to conduct prediction from the shorter time period data (e.g., day data) to the longer time period data (e.g., one month data or longer).

We select 1^{st} day, 1^{st} week, and the whole month, as the representative of day, week, and month respectively.

- **Day**: A full 24 hours data.
- **Week**: 1101 to 1107 is the first week of November.
- **Month**: By default, monthly data means 30 days data in the following.

B. Goodness-of-fit

For testing a hypothesis whether a distribution function fits a given data, there are Kolmogorov-Smirnov (KS) test, Berk-Jones test, score test and their integrated versions [35], [36]. Kolmogorov-Smirnov score quantifies a distance between the empirical distribution function (ECDF) of the sample and the cumulative distribution function of the reference distribution, or between the empirical distribution functions of the two samples, as a sequence. Intuitively, it can be used to test whether a sample comes from the same distribution (KS-score is 0 in the ideal case). KS score is defined as follows.

$$S_x = \sup_x |F_n(x) - F(x)| \tag{2}$$

where $F_n(x)$ is ECDF of n samples.

In addition, we can visualize the fitting by plotting the PDF against the real data. In our case, since the number of calls has a long tail, we could plot them in log-log scale. Another way to visualize is through odds ratio as a measure of a fitting to real data. We will see both plots in the experiment section. The better fitting, the closer of the two curves.

C. Churning Behaviors Analysis

Newcomer and Churner: We examine the number of people join and leave the network to exam whether there exists a pattern of retiring in the mobile phone social network. We count the number of comer and churner based on following definitions:

- Newcomer: the customer ID which does not appear from day i-30 to day i-1, but appears in day i (say, appears in day i but not in any of the 30 days before day i)
- Churner: the customer ID which appears in day i but not in day i+1 to day i+30 (say, appears in day 1 but not in any of the 30 days after day i)

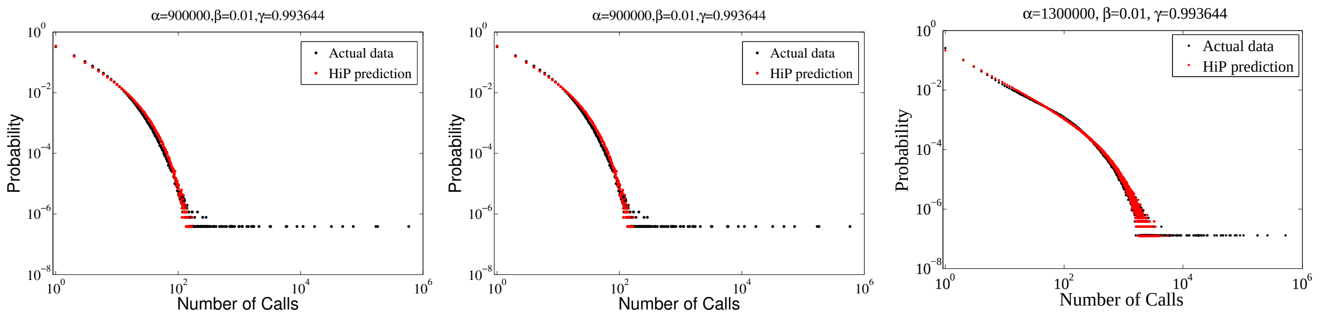
We notice from our data set that there is a relative constant daily in-and-out in the social network. For the time period we exam, the number of newcomers is slightly larger than the churner. But for other time period, the situation could be different. At the same time, even for a given time period, there is slight chance that the number of newcomers is the same as the number of churners. Thus, the churners' impact or change to the data is hard to be eliminated by the newcomers. Therefore, we design HiP model, with parameters α and β for new phones joining the network, and a parameter γ representing the effect of mobile phone customers leaving the network.

D. Comparison Study

We compare our model, HiP, with a set of parametric models, Lognormal, GP, DPLN and Loglogistic, by real mobile phone call data.

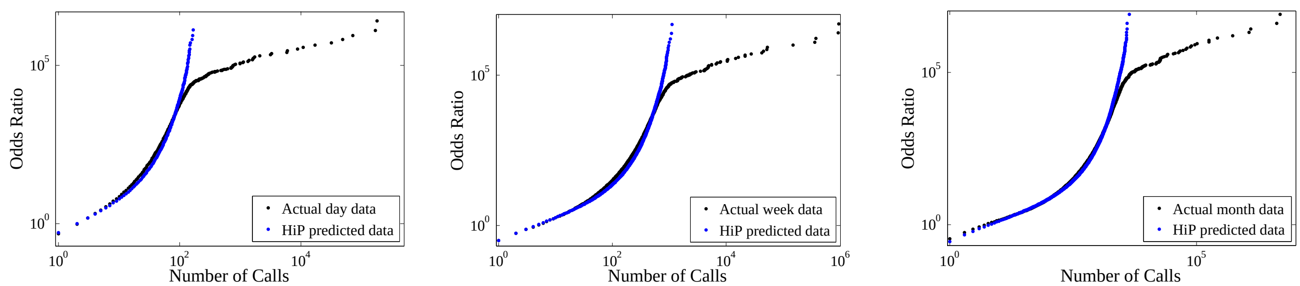
1) Hibernating Process: We evaluate HiP on day, week and month data as shown in Figure 2. Apparently our HiP model fits the real data distribution very well: not only for one day data and one week data, but also for one month data.

In Figure 3, we report the odds ratio curve of HiP for the day data, week data and month data. Odds ratio is defined



(a) Number of calls for a day (b) Number of calls for a week (c) Number of calls for a month

Figure 2. HiP prediction of number of calls daily, weekly and monthly. HiP (in red) matches real data (in black) at all time scales. Notice that HiP estimates its parameter from only two values: the total count of calls and distinct phones.



(a) Daily data (b) Weekly data (c) Monthly data

Figure 3. Odds ratio curve of HiP prediction (in blue) for one day, one week and one month data. HiP almost overlap with the real in every case.

as $ratio(x) = \frac{CDF(x)}{1-CDF(x)}$. The results conclude that HiP is a good fit for different time-scales.

In Table I, we report the KS score of HiP to fit day data, week data and month data. Our method can achieve 75% improvement than the previous models. The analysis of the fitting results is provided in the following subsection, compared with a set of parametric models, Lognormal, GP, DPLN and Loglogistic (our model HiP can fit the real data much better than Lognormal, GP, DPLN and Loglogistic).

2) Parametric models: To investigate the goodness-of-fit of four distributions at multiple scale data, we conduct KS test, and the results are reported in Table I. Our method can achieve 75% improvement than the previous models. For the day data, we run 60 days' simulation, and take a mean of each KS score. For the week data, we run 30 weeks' simulation, and take a mean of each KS score. For the month data, we run 12 months' simulation, and then take a mean of each KS score. The KS score confirms that HiP is much better than the other four models in fitting to the actual data, which means HiP can model the real generation process of the actual mobile social graph at different time scales.

Table I
GOODNESS-OF-FIT (KS SCORE, LOWER IS BETTER)

Method	Day	Week	Month
Lognormal	0.1881	0.1298	0.1347
DPLN	0.3272	0.1298	0.2544
Loglogistic	0.1762	0.1238	0.1442
GP	0.2232	0.1398	0.1467
HiP	**0.0432**	**0.0817**	**0.0940**
HiP improvement	75.5%	34.0%	30.2%

VI. CONCLUSIONS

In this paper, we revealed a very interesting phenomenon hidden in mobile social networks, that is, churning behaviors. Newcomers and churners in together maintain the dynamics in a mobile social graph. Based on such observations, we proposed Hibernating Process (HiP), a novel non-parametric model for the accumulating process of phone calls. HiP was able to predict the number of phone calls in mobile network at multiple scale (daily, weekly, monthly). We applied HiP on three different mobile social network data and those at three scales, with constantly high quality matches, visually and quantitatively. Our HiP improves prediction by 30% to 75% over the best existing method on each single scale.

Our findings and model may well extend to other social networks with such transient behaviors, which is a promising direction for future research.

REFERENCES

[1] L. A. Adamic and B. A. Huberman, "The web's hidden order," *Commun. ACM*, vol. 44, pp. 55–60, September 2001.

[2] S. Aral and D. Walker, “Identifying influential and susceptible members of social networks,” *Science*, vol. 337, July 2012.

[3] D. Chakrabarti and C. Faloutsos, “Graph mining: Laws, generators, and algorithms,” *ACM Comput. Surv.*, vol. 38, June 2006.

[4] P. O. S. V. De Melo, L. Akoglu, C. Faloutsos, and A. A. F. Loureiro, “Surprising patterns for the call duration distribution of mobile phone users,” in *ECML-PKDD*, 2010.

[5] J. Guo, F. Liu, and Z. Zhu, “Estimate the call duration distribution parameters in gsm system based on k-l divergence method,” in *IEEE WiCOM*, 2007.

[6] S. Keshav, “Why cell phones will dominate the future internet,” *SIGCOMM Comput. Commun. Rev.*, vol. 35, April 2005.

[7] M. Faloutsos, P. Faloutsos, and C. Faloutsos, “On power-law relationships of the internet topology,” in *ACM SIGCOMM*, 1999.

[8] A. Clauset, C. R. Shalizi, and M. E. J. Newman, “Power-law distributions in empirical data,” *SIAM Rev.*, vol. 51, November 2009.

[9] H. Fofack and J. Nolan, “Tail behavior, modes and other characteristics of stable distributions,” *Extremes*, vol. 2, pp. 39–58, 1999.

[10] J. P. Nolan, *Stable Distributions - Models for Heavy Tailed Data*. Boston: Birkhauser, 2011.

[11] J. Leskovec, J. Kleinberg, and C. Faloutsos, “Graph evolution: Densification and shrinking diameters,” *ACM Trans. Knowl. Discov. Data*, vol. 1, March 2007.

[12] M. E. J. Newman, “Power laws, pareto distributions and zipf’s law,” *Contemporary Physics*, vol. 46, May 2007.

[13] M. Seshadri, S. Machiraju, A. Sridharan, J. Bolot, C. Faloutsos, and J. Leskove, “Mobile call graphs: beyond power-law and lognormal distributions,” in *ACM SIGKDD*, 2008.

[14] W. J. Reed and M. Jorgensen, “The double pareto-lognormal distribution: A new parametric model for size distributions,” *Communications in Statistics - Theory and Methods*, vol. 33, 2004.

[15] Z. Bi, C. Faloutsos, and F. Korn, “The ”dgx” distribution for mining massive, skewed data,” in *ACM SIGKDD*, 2001.

[16] M. Gomez Rodriguez, J. Leskovec, and A. Krause, “Inferring networks of diffusion and influence,” in *ACM SIGKDD*, 2010.

[17] J. Leskovec, J. Kleinberg, and C. Faloutsos, “Graphs over time: densification laws, shrinking diameters and possible explanations,” in *ACM SIGKDD*, 2005.

[18] J. Leskovec and C. Faloutsos, “Scalable modeling of real graphs using kronecker multiplication,” in *ICML*, 2007.

[19] J. Leskovec, D. Chakrabarti, J. Kleinberg, C. Faloutsos, and Z. Ghahramani, “Kronecker graphs: An approach to modeling networks,” *J. Mach. Learn. Res.*, vol. 11, Mar. 2010.

[20] J. Leskovec, L. Backstrom, R. Kumar, and A. Tomkins, “Microscopic evolution of social networks,” in *ACM SIGKDD*, 2008.

[21] M. McGlohon, L. Akoglu, and C. Faloutsos, “Weighted graphs and disconnected components: patterns and a generator,” in *ACM SIGKDD*, 2008.

[22] J. Huang, Z. Zhuang, J. Li, and C. L. Giles, “Collaboration over time: characterizing and modeling network evolution,” in *WSDM*, 2008.

[23] B. Ribeiro and D. Towsley, “Estimating and sampling graphs with multidimensional random walks,” in *IMC*, 2010.

[24] C.-T. Li and S.-D. Lin, “Social flocks: a crowd simulation framework for social network generation, community detection, and collective behavior modeling,” in *ACM SIGKDD*, 2011.

[25] R. Patro, G. Duggal, E. Sefer, H. Wang, D. Filippova, and C. Kingsford, “The missing models: a data-driven approach for learning how networks grow,” in *ACM SIGKDD*, 2012.

[26] E. Zheleva, H. Sharara, and L. Getoor, “Co-evolution of social and affiliation networks,” in *ACM SIGKDD*, 2009.

[27] J. Kunegis, D. Fay, and C. Bauckhage, “Network growth and the spectral evolution model,” in *ACM CIKM*, 2010.

[28] R. Kumar, J. Novak, and A. Tomkins, “Structure and evolution of online social networks,” in *ACM SIGKDD*, 2006.

[29] D. Willkomm, S. Machiraju, J. Bolot, and A. Wolisz, “Primary users in cellular networks: A large-scale measurement study,” in *IEEE DySPAN*, 2008.

[30] D. Montgomery and G. Runger, *Applied Statistics and Probability for Engineers*. Wiley India Pvt. Ltd., 2007.

[31] Z. Wu, X. Li, R. Husnay, V. Chakravarthy, B. Wang, and Z. Wu, “A novel highly accurate log skew normal approximation method to lognormal sum distributions,” in *IEEE WCNC*, 2009.

[32] D. M. Blei, T. L. Griffiths, and M. I. Jordan, “The nested chinese restaurant process and bayesian nonparametric inference of topic hierarchies,” *J. ACM*, vol. 57, no. 2, Feb. 2010.

[33] X. Chen, M. Zhou, and L. Carin, “The contextual focused topic model,” in *ACM SIGKDD*, 2012.

[34] P. M. Fenwick, “A new data structure for cumulative frequency tables,” *Software: Practice and Experience*, vol. 24, pp. 327–336, 1994.

[35] A. Koning and L. Peng, “Goodness-of-fit tests for a heavy tailed distribution,” Erasmus University Rotterdam, Econometric Institute, Econometric Institute Report, Nov 2005.

[36] M. C. Bryson, “Heavy-tailed distributions: Properties and tests,” *Technometrics*, vol. 16, no. 1, 1974.