Reciprocity and Related Bivariate Patterns in Large Phone and SMS Networks

Leman Akoglu
CMU, SCS, and iLab
lakoglu@cs.cmu.edu

Pedro O.S. Vaz de Melo
NTT
olmo@dcc.ufmg.br

Christos Faloutsos
CMU, SCS, and iLab
christos@cs.cmu.edu

ABSTRACT

In a mobile communication network, if user i calls/texts user j n times, what can we say about the number of times j calls/texts i? Also, given a user has k contacts, can we say anything about the total number/duration of his/her phone-calls/SMSs? In this work, we study two real communication networks of millions of users registered to an anonymous mobile phone company in a large city: a Mobile Call Graph(MCG) and a Mobile Text Graph(MTG) and answer these questions.

Such real graphs were found to exhibit many common patterns for example in degree distribution of nodes. Those patterns are usually modeled by *univariate* distributions such as power laws, log-normals, double Pareto log-normals, etc. In this paper, we take one step ahead and study *bivariate* distributions. Our main contributions are: (1) we observe bivariate patterns in (a) the joint distribution of weights on reciprocated edges, $Prob(w_{ij}, w_{ji})$; and (b) the joint distribution of degree and strength of nodes, $Prob(k, s)$ in the MCG and MTG. We observe that these patterns show skewed characteristics and so summarization by the mean and the median often give different, even conflicting insights. Instead, (2) we propose two bivariate functions, namely the Triple Power-Law(3PL) and the Cascaded Log-Normal(CLN), to fit the observed patterns in (a) and (b), respectively. We show how to estimate parameters of the proposed functions and that our fitting of parameters can model the real distributions in MCG and MTG with up to 2M nodes and 42M edges well.

1. INTRODUCTION

Looking at the data is a vital part of understanding it. In data mining, many researchers usually start analyzing their data by visualizing it. For example for graph data, visualization of the nodes and the edges in different layouts could be very helpful in understanding the global structure and the high-level connectivity in general. In fact, data summarization and visualization is a widely studied research area by itself [7, 14, 27, 28, 29].

Another basic step to understand the data in hand is to study the simple distributions in it. For example, one could look at the degree distribution of nodes in a graph, or the distribution of city populations in a country to get insights about how a particular quantity (degree, population, etc.) is distributed among data (nodes, cities, etc.). Such distributions in real data were found to obey several parametric *univariate* distributions such as power-laws [9], log-normals [5], and recently new distributions such as DPLNs [30]. The study of *multivariate* distributions in real data, on the other hand, has very limited focus (Section 2).

In this paper, we take one step ahead and study *bivariate* patterns in two real mobile communication networks of millions of users; a Mobile phone-Call Graph(MCG) and a Mobile Text(SMS) Graph (MTG) (Section 3). Our main focus is in the study of *reciprocity* and *edge weights*. Informally, we give answers to the following questions: In a mobile phone-call graph of users, if user i calls user j n times, what can we say about the number of times j calls i? Also, given a user with k contacts, what can we say about the total number/duration of his/her phone-calls? How about in a mobile SMS graph? More formally, (how) can we model the *joint* distribution of (1) the weights (w_{ij}, w_{ji}) on reciprocal edges (e_{ij}, e_{ji}) between pairs of nodes (i, j); and (2) the degree and strength (k, s) of nodes? (here, the strength of a node is defined as the total weight of the edges attached to it).

The motivation behind this work is as follows:

- Characterization of bivariate distributions in real data is limited (Section 2.2).
- Visualization is usually misleading due to over-plotting.
- Summarization by average or median values is not representative for skewed distributions and often the two give different insights (Section 4).
- A better representation and formulation of multivariate data is fundamental.

Main contributions of this work are the following:

1. We propose the Triple Power-Law(3PL) to formulate the distribution of nodes with given degree k and strength s;
2. We propose the Cascaded Log-Normal(CLN) to formulate the distribution of reciprocal edge pairs (e_{ij}, e_{ji}) with given respective weights (w_{ij}, w_{ji}).
3. We show how to estimate parameters for the proposed functions and that our proposed functions and fitting of parameters model the real distributions in the MCG and MTG well (Section 5).

Additional contributions are:

1. We observe that the bivariate distributions in MCG and MTG are very skewed and so 2D visualization could be misleading due to high over-plotting (Section 4.1).
2. We show that there exists a *conundrum* between mean and median values in summarization and that for skewed distributions the two give different results (Section 4.2).

KDD '10, July 25–28, 2010, Washington, DC, USA.

2. BACKGROUND AND RELATED WORK

In this section, we provide background on several skewed univariate and bivariate distributions observed in real graphs such as power-laws and log-normal distributions. We also give a brief survey of prior work on mobile communication networks.

2.1 Univariate Distributions in Real Data

A quantity x follows a power-law if it is drawn from a probability distribution $p(x) \propto x^{-\alpha}$, where α is called the power law exponent. Many patterns regarding power-laws have been found to occur in real graphs and social networks. For example, the degree distribution obeys a power-law in many real graphs from a large variety of domains such as the Internet Autonomous Systems graph [13], the WWW link graph [3, 6], several phone-call graphs [1, 2], and many more [8, 16, 18, 22]. Additional power laws seem to govern the popularity of posts in citation networks, which drops over time, with power law exponent of -1 for paper citations [26] or -1.5 for blog posts [19].

A recent comprehensive study [9] on power-law distributions in empirical data shows that while power-laws exist in many graphs, deviations from a pure power-law are also observed. Those deviations usually appear in the form of exponential cut-offs and log-normals. Similar deviations were also observed in [4] where the electric power-grid graph in a specific region in California as well as airport networks were found to exhibit power-law distributions with exponential cut-offs. Also, [25] observe that subsets of the WWW, for example university homepages, deviate significantly from a power-law distribution.

Discrete Gaussian Exponential(DGX) [5] was also shown to provide good fits to distributions in a variety of real world datasets such as the Internet click-stream data and usage data from a mobile phone operator. Most recently, [30] studied several phone-call networks and proposed a new distribution called the Double Pareto Log-Normal(DPLN) for the per-user number of call partners, number of calls and number of minutes.

2.2 Bivariate Distributions in Real Data

While univariate distributions are used to model the distribution of a specific quantity x, for example the number of calls of users, bivariate distributions are used to model the association and covariation between two quantitative variables x and y. Association is based on how two variables simultaneously change together, for example the number of calls w.r.t. the number of call partners of users.

Unlike univariate distributions, the study of multivariate distributions has been limited to theoretical analysis of such distributions in mathematics and statistics. On the other hand, multivariate analysis *in real data* has much less focus. [32] uses the bivariate log-normal distribution to describe the joint distributions of flood peaks and volumes, and flood volumes and durations. Also, [20] studies the drought in the state of Nebraska and models the duration and severity, proportion and inter-arrival time, and duration and magnitude of drought with bivariate Pareto distributions.

2.3 Study of Mobile Phone Graphs

Social networks of mobile phone users have been previously studied in the literature. For example Onnela et. al. [23, 24] study the local and global structure of a large communication network and show that there exists coupling between interaction strengths and local neighborhoods of individuals. Nanavati et. al. [21] study the structure and the global shape of four geographically disparate mobile call graphs and propose the Treasure-Hunt model to fit their observations. [10] analyze the number and size of the triangles and maximal cliques that users participate in phone networks and find patterns in their distributions. [31] study the formation of social communities in temporal telecommunications records. Finally, Eagle et. al. [11, 12] study massive amounts of mobile phone records and infer social structure and behavior of users by their mobile phone interactions.

3. DATA DESCRIPTION

In this work, we studied anonymous mobile communication records of millions of users over a time period of six months. The dataset not only contains phone-call but also SMS interactions. The data spans from December 1, 2007 through May 31, 2008 (183 days). It contains all the interactions between the within-network users (actual customers) as well as incoming/outgoing interactions from/to out-of-network users.

The data is in the form of *callerID, calleeID, date/time-of-call/text, duration (in seconds, only for phone-calls)*. From the whole six months' of activity, we built two graphs in which nodes represent users and directed edges represent (phone-call and SMS) interactions between these users. We call the who-calls-whom graph as the MCG (for Mobile Call Graph) and the who-texts-whom graph as the MTG (Mobile Text Graph).

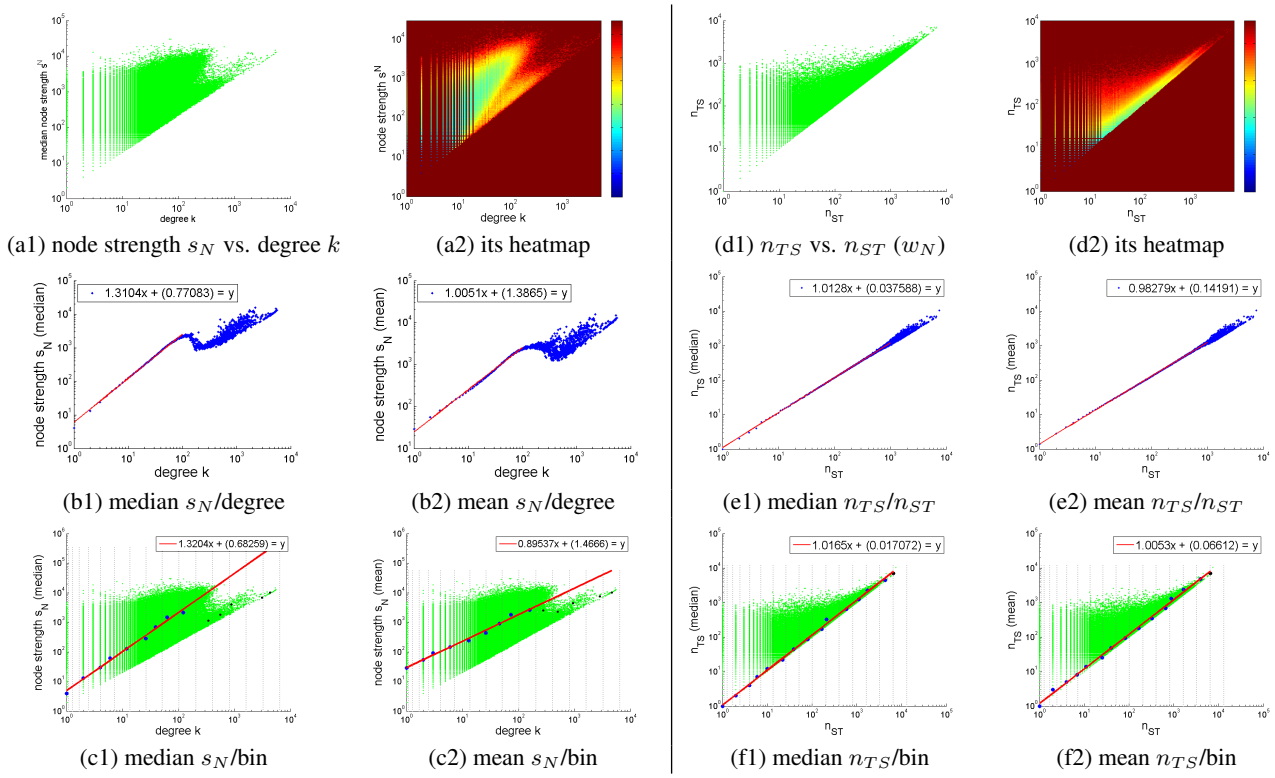
By construction, our graphs are weighted. Here, we consider two types of weights on the arcs e_{ij}: (1) total number of phone-calls w_N, similarly total number of SMSs w_{SMS}; and (2) total duration of phone-calls w_D from node i to j (only for MCG, aggregated in 10 minutes).

There are two choices that we made to construct the MCG and MTG. Firstly, since only the activity of within-network customers can be tracked, our data does not contain all the interactions of out-of-network users. Thus, as we have only partial information about the out-of-network users, we decided to include only the within-network customers and the interactions among themselves in our graphs. Secondly, our graphs include only reciprocated edges between nodes. The reason is, as [23] point out in their study of a similar network, we consider only *mutual* interactions between node i and node j to be "real". That is, we believe that two nodes *actually* interact only if for example i calls j *and* j also calls i. We will denote the smaller weight on the reciprocal edges as n_{ST} and the larger weight as n_{TS}, (S for Silent and T for Talkative).

To give a sense of the scale of the data we studied, we show the number of customers N (with at least one contact), the number of directed interactions E_{dir}, the total number of phone-calls/SMSs, $W_{N,SMS}$ and the total duration of phone-calls W_D, for both the mutual and non-mutual MCG and MTG in Table 1 (note that we performed experiments only on the mutual networks as we discussed earlier). Notice that the MTG shrinks considerably when only mutual edges are considered, whereas MCG remains almost intact. This shows that the vast majority of SMS interactions occur non-mutually suggesting for spam messages.

Network	N	E_{dir}	$W_{N,SMS}$	$W_D(sec.s)$
MCG(non-mutual)	1,87M	49,50M	483,7M	5,49x10^{10}
MCG(mutual)	1,75M	41,84M	468,7M	5,31x10^{10}
MTG(non-mutual)	1,87M	8,70M	119,5M	N/A
MTG(mutual)	0,53M	1,99M	91,8M	N/A

Table 1: The number of nodes N, the number of edges E, and the total weight W in the mutual and non-mutual MCG and MTG. Note the sharp drop in the size of the MTG when only mutual edges are considered.



(a1) node strength s_N vs. degree k (a2) its heatmap (d1) n_{TS} vs. n_{ST} (w_N) (d2) its heatmap

(b1) median s_N/degree (b2) mean s_N/degree (e1) median n_{TS}/n_{ST} (e2) mean n_{TS}/n_{ST}

(c1) median s_N/bin (c2) mean s_N/bin (f1) median n_{TS}/bin (f2) mean n_{TS}/bin

Figure 1: (top row) Visualization by scatter plots loses information due to over-plotting. 2D heatmaps recover the missing information, highlighting the density of regions (figures best viewed in color, dark blue is denser, dark red is sparser). (bottom two rows) Conundrum between mean and median: median suggests super-linear, whereas mean suggests (sub-)linear growth for (left) strength s_N w.r.t. degree k; and (right) edge weight n_{TS} w.r.t. n_{ST}, where weights denote the number of phone-calls w_N in MCG.

4. OBSERVATIONS AND ISSUES

In this section we introduce the problem we are looking at and give the motivation behind our work.

Our main goal in this paper is to analyze the covariation between pairs of features (F_1, F_2) of nodes and edges in our mobile graphs MCG and MTG. That is, we want to understand how feature F_1 changes when feature F_2 changes among nodes/edges. In particular, we studied *two* feature pairs.

1. **f(strength s, degree k) among nodes:** Here, we want to understand how the strength s (total weight) changes with increasing degree k among nodes in the MCG and MTG. In other words, we study how the total number of phone-calls s_N, SMSs s_{SMS} and the total duration s_D a user spends on the phone is affected by the number of his/her contacts.
2. **f(n_{ST}, n_{TS}) among reciprocal edges:** Here, we analyze how the larger weight n_{TS} is affected by a change in the smaller weight n_{ST} among mutual edges in the MCG and MTG. In other words, we want to understand how the number of phone-calls/SMSs from node i to j is related to those from node j to node i.

4.1 Over-plotting in visualization

Figure 1(a1) shows the scatter plot of the strength s versus degree k for all the nodes in the MCG, where the strength of a user denotes the total number of his/her phone-calls s_N. Each green dot in the plot corresponds to a node in the graph. As one can imagine, there exists a lot of over-plotting in the shown figure and therefore the densities of regions are not clear. To alleviate this problem, one can instead look at the 3D histogram plots or 2D heatmaps where colors represent the magnitude of volume. In Figure 1(a2), we show the heatmap for the same plot. We observe that there are a lot of points around the origin which points to a majority of small degree nodes with correspondingly small total weights in the graph.

Similarly, Figure 1(d1) shows the weights n_{TS} versus n_{ST} for all the reciprocal edges in the MCG, where weights denote the total number of phone-calls w_N. Each green dot in the plot corresponds to a pair of mutual edges. Again, there is over-plotting in the shown figure and therefore the densities of regions are not clear. In Figure 1(d2), we show the heatmap for the same plot. Notice that most of the edges are concentrated (1) around the origin and then (2) close to the diagonal. (1) suggests that most of the edges have small weights whereas fewer edges have very high weights, which points to skewness. In addition, (2) indicates that it is highly probable that i and j call each other in a more balanced fashion, that is, around the same number of times.

To summarize, in order to understand the distribution of points in 2D, one should consider over-plotting and treat it carefully.

4.2 Conundrum between mean and median

In order to formulate the bivariate distributions we mentioned in the previous section, for example how the node strength changes with degree, one can compute an aggregate function (e.g. the average) of the node strengths of the nodes with a given degree. Figure 1(b1) and Figure 1(b2) show the median and the mean node strength s_N for a given degree k, respectively. The number of nodes with degree higher than 100 is very small and therefore we treat the points beyond that value on the x-axis as noise and observe

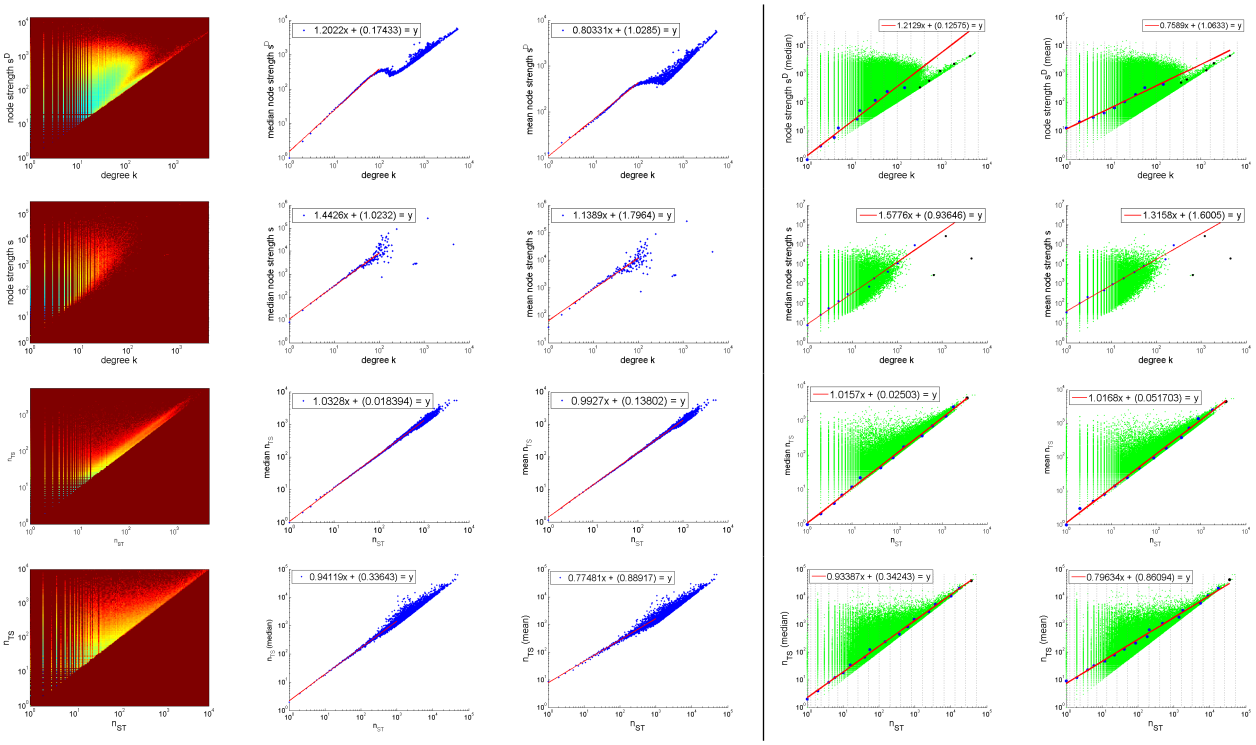


Figure 2: (left) Conundrum between the median and the mean for (from top to bottom) total duration s_D vs. degree k (with respective slopes 1.20,0.80), number of SMSs s_{SMS} vs. degree k (1.44,1.13), n_{TS} vs. n_{ST} (w_D) (1.03,0.99), and n_{TS} vs. n_{ST} (w_{SMS}) (0.94,0.77). (right) Logarithmic binning also gives similar results. See text for details.

that a Least Squares(LS) line can be fit to the rest of the points. The line fit to (x, y) points on log-log scales suggests a power-law relation in the form of $y \propto x^{\alpha}$.

Another method to summarize 2D data is to divide the data points into chunks after logarithmically binning the x-axis [22] and computing the mean/median of the points in each bin (this is to account for the sparsifying nature of the data with increasing x). In Figures 1(c1,2;f1,2), dashed vertical lines show the boundaries of each bin. The blue dots correspond to the mean/median values computed over the points in each bin. Again, the red lines show the LS fits to the blue dots.

The main point we want to make here is that, the power-law exponent α is greater than 1 for the median fit (Fig.1(b1,c1)), which suggests a super-linear growth in strength with increasing degree, whereas it is less than 1 (or around 1) for the mean fit (Fig.1(b2,c2)), which on the contrary suggests a (sub-)linear growth. More intuitively, mean fit indicates that a user with higher number of friends spends less time per friend on average than a user with fewer friends. On the other hand, median fit suggests that the more friends one talks to, super-linearly more time s/he will spend on the phone in total. In addition, we observe a similar conflict on the reciprocal edge weights w_N (Figure 1(e1,e2) and (f1,f2)), though the power-law exponents are relatively close in this case.

We note that the same conundrum between the mean and the median also occurs in the MTG and when the weights are taken to be the total duration of phone-calls w_D in MCG. See Figure 2.

All in all, we observed that there exist bivariate patterns in the mobile communication graphs MCG and MTG in (1) the distribution of the strengths s_N, s_D, s_{SMS} and the degree among nodes; and (2) the distribution of the edge weights w_N, w_D, w_{SMS} among reciprocal edges. We also showed that summarizing these distributions by aggregation could give different insights, in particular, that the mean and the median fall into conundrum.

5. BIVARIATE PATTERNS IN MCG AND MTG

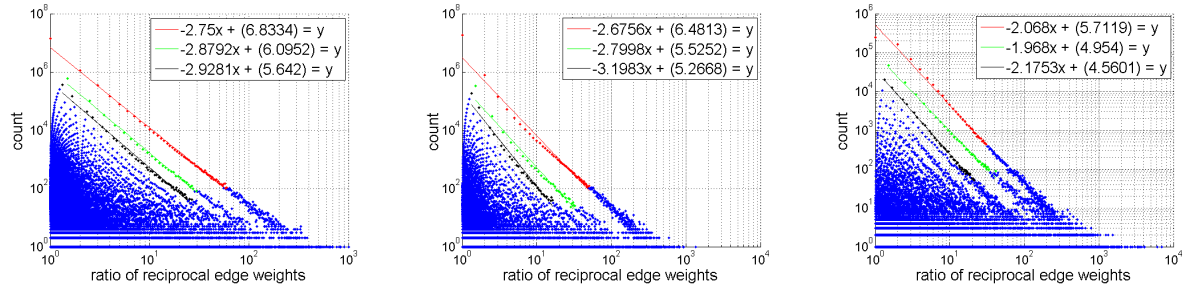
In this section, we delve more into the details of the observed patterns. As an alternative to summarization/aggregation, we propose bivariate functions to fit these observed distributions. We also provide parameter fitting routines to the proposed functions.

5.1 Patterns in Reciprocal Edge Weights

Given a network of customers with *mutual*, *weighted* edges between them, we want to understand the association between the weights on the reciprocal edge pairs. In other words, given two nodes i and j, say in MCG, is there a relation between the number of calls i makes to j (n_{ij}) and the number of calls j makes to i (n_{ji})? To ease notation, we will denote the smaller of these weights as n_{ST} (for Silent-to-Talkative), and the larger as n_{TS}.

Figure 3 shows the distribution of weight ratios $\frac{n_{TS}}{n_{ST}}$ of all reciprocal edge pairs (a) in MCG with weights w_N as number of phone-calls, (b) in MCG with weights w_D as total duration aggregated in 10 minutes, and (c) in MTG with weights w_{SMS} as number of SMSs. We observe that the distribution of the weight ratios in all three cases follow *layers* of power-laws. Fitting LS lines to the top three so-called "layers" of points in log-log scales, we notice that the power-law fits have similar exponents with shifted intercepts –many $1, 2, 3, \ldots$; fewer $1.5, 2.5, 3.5, \ldots$; even fewer $1.33, 1.66, 2.33, \ldots$; and so on.

In order to visualize the relation between n_{ST} and n_{TS}, we plot the bivariate data points in a scatter plot in Figure 5(left), using 2D heatmap, and contour plots for weights (from top to bottom) w_N, w_D and w_{SMS}, respectively. We notice that in all three types of



(a) w_N (b) w_D (c) w_{SMS}

Figure 3: Distribution of the ratio of weights on reciprocal edges $\frac{n_{TS}}{n_{ST}}$ follows "layers" of power-laws with similar exponents for all three types of weights (a) number of phone-calls w_N, (b) duration of phone-calls w_D, and (c) number of SMSs w_{SMS}.

weights, majority of the points are concentrated around the origin. Also, while the volume decreases with increasing n_{TS} for a fixed n_{ST}, the volume seems to increase with increasing n_{ST} for a fixed n_{TS} –hence the higher volume along the diagonal.

To further analyze the co-variation between n_{ST} and n_{TS}, we looked at the distributions $P(n_{TS}|n_{ST} = x)$ and $P(n_{ST}|n_{TS} = y)$ for x and y up to 100. One can imagine this as looking at the distribution of points on vertical and horizontal "slices". For brevity, we show the distributions for x ={2,8}; and for y ={80,100} in Figure 4(a) and (b) for w_N and w_D, respectively (all the rest look similar). Here, we observe that $P(n_{TS}|n_{ST} = x)$ follows a power-law distribution with an "elbow" shape. That is, we can fit *two* power-laws; one for the relatively smaller values of x and another for the rest with an often *larger* exponent (blue and red lines in the same figures). We mark the "elbow" point with a vertical black dashed line. On the other hand, we observe that $P(n_{ST}|n_{TS} = y)$ decreases slowly with small oscillations, and then follows a power-law with a negative exponent.

For w_{SMS} the above argument holds with an interesting deviation. In Figure 4(c), we plot $P(n_{TS}|n_{ST} = x)$ for x ={1,2}, and notice that the data follows two trends of power-laws. We realize that here, one power-law can be fit to n_{TS} of *even* values (green dots) and another with a larger exponent to the *odd* values (blue dots). This indicates that the weights on edges in the MTG are more likely to be *even*. Although we do not have a clear explanation for this, we think that when it comes to SMS, users might be communicating in a 4-way handshake procedure, that is, Question-Answer-ACK-ACK; which accounts for two messages per user at a time. On the other hand, $P(n_{ST}|n_{TS} = y)$ follows a similar pattern as above and decreases slowly with small oscillations for even n_{ST} with no clear increase as with w_N and w_D, but is very noisy for odd values.

Given the above observations, we want to model $P(n_{ST}, n_{TS})$. Here, one can think of using well-known parametric distributions from statistics, such as the bivariate Pareto [17] and the bivariate log-normal [15] distributions. There are several bivariate Pareto distributions in the statistics literature. Among these, the simplest has the joint probability density function(pdf) specified by

$$f_{X,Y}(x, y) = k(k + 1)(ab)^{k+1}(ax + by + ab)^{-k-2}$$

for $x > 0, y > 0, a > 0, b > 0$ and $k > 0$. However, looking at the pdfs of these and many more bivariate distributions, we notice that none of them can be used to model the distributions we observe here. The reason is that in such distributions, $f_{X,Y}(x, y)$ always decreases by increasing either of the variables x and y. On the other hand in our case, when y is fixed, $f_{X,Y}(x, y)$ should be increasing after x exceeds a certain point. (See Figure 4(right)).

Weight type	β	α	γ	RSS
w_N	1.3796	0.4924	1.6152	4.6065e-04
w_D	2.0507	1.3341	1.7879	0.0015
w_{SMS}^{ee}	1.4	1.4428	0	0.0046
w_{SMS}^{eo}	1.0144	0.5147	0.2251	8.7056e-05
w_{SMS}^{oe}	1.0858	0.3304	0.4700	1.6677e-04
w_{SMS}^{oo}	0.6273	0.3333	0.2093	3.1207e-04

Table 2: LS parameters fit to the 3PL function $f(n_{ST}, n_{TS}) \propto n_{ST}^{-\alpha} n_{TS}^{-\beta}(n_{TS} - n_{ST})^{-\gamma}$ and Residual Sum of Squares (RSS) error for weights w_N, w_D and w_{SMS}.

As we cannot model the observed patterns by well-known bivariate parametric distributions, we propose to formulate these distributions with a what we call Triple Power-Law (3PL) function.

MODEL 1 (TRIPLE POWER-LAW (3PL)). *In phone and SMS networks, the number of mutual edge pairs with weights n_{ST} and n_{TS} (number of phone-calls/SMSs or total duration of phone-calls) on each reciprocal edge (n_{ST} being the smaller of the two) follows a Triple Power-Law in the form of*

$$\begin{aligned} P(n_{ST}, n_{TS}) &= P(n_{ST})P(n_{TS}|n_{ST}) \\ &\propto \boxed{n_{ST}^{-\alpha} n_{TS}^{-\beta}(n_{TS} - n_{ST})^{-\gamma}}. \end{aligned}$$

for $\alpha > 0$, $\beta > 0$, and $\gamma > 0$. Here, we integrate our observations that both n_{ST} and n_{TS} are power-law distributed (first and second terms). We also observed that $P(n_{TS}|n_{ST})$ obeys a power-law, by definition starting at n_{ST} ($n_{TS} \geq n_{ST}$)(third term). Second and third terms also account for the "elbow" shape we observed for $P(n_{TS}|n_{ST})$, to which two separate power-laws can be fit.

Finally, to find the fitting parameters to the 3PL model for the real distributions in MCG and MTG, we used the optimization toolbox in Matlab where we defined the objective function to be the Residual Sum of Squares (RSS) to be minimized with constraints $\alpha > 0$, $\beta > 0$, and $\gamma > 0$. We show the estimated parameters for all three graphs as well as the RSS errors in Table 2. Note that we had to estimate parameters for the distribution of all the cross-product of even and odd weights in the MTG, w_{SMS}^{ee}, w_{SMS}^{eo},w_{SMS}^{oe} and w_{SMS}^{oo}.

Having estimated the model parameters we generated synthetic samples with the same number of edge pairs for each graph. We show the corresponding plots for the synthetic graphs in Figure 5(right) for (from top to bottom) w_N, w_D and w_{SMS}, respectively. Also see Figure 6(top) for the contour plots of the distribution of edges in MTG^{ee}, MTG^{eo},MTG^{oe} and MTG^{oo}. We show the

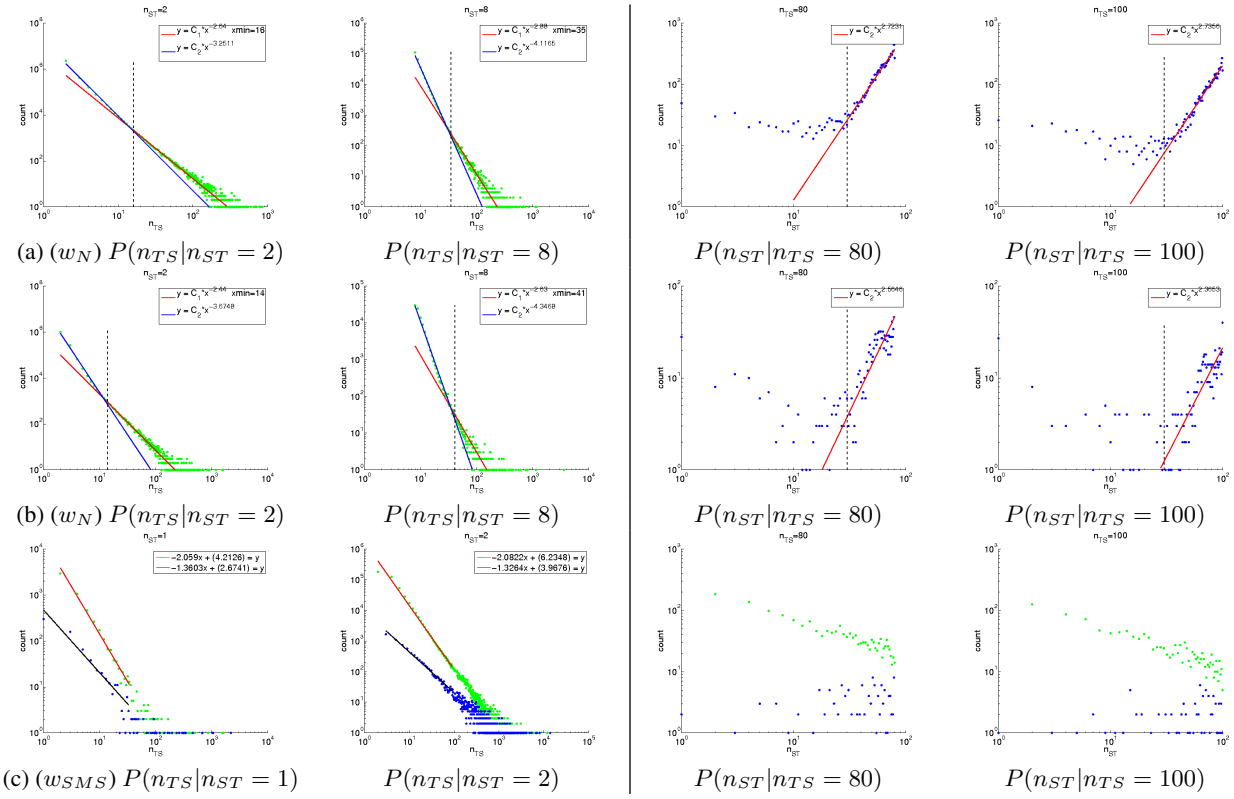


Figure 4: (left) Power-law(PL) fits to the distribution of n_{TS} given n_{ST}={2,8} (vertical "slices") for (a) w_N and (b) w_D. Notice the "elbow" shape to which two separate power-laws are fit. (c) Distribution of n_{TS} given n_{ST}={1,2} for w_{SMS}. Even values (green) are more probable than odd values (blue) with a smaller PL exponent. (right) Distribution of n_{ST} given n_{TS}={80,100} (horizontal "slices"). Notice that $P(n_{ST}|n_{TS})$ decreases slowly up to a point (marked with dashed lines), after which it starts increasing.

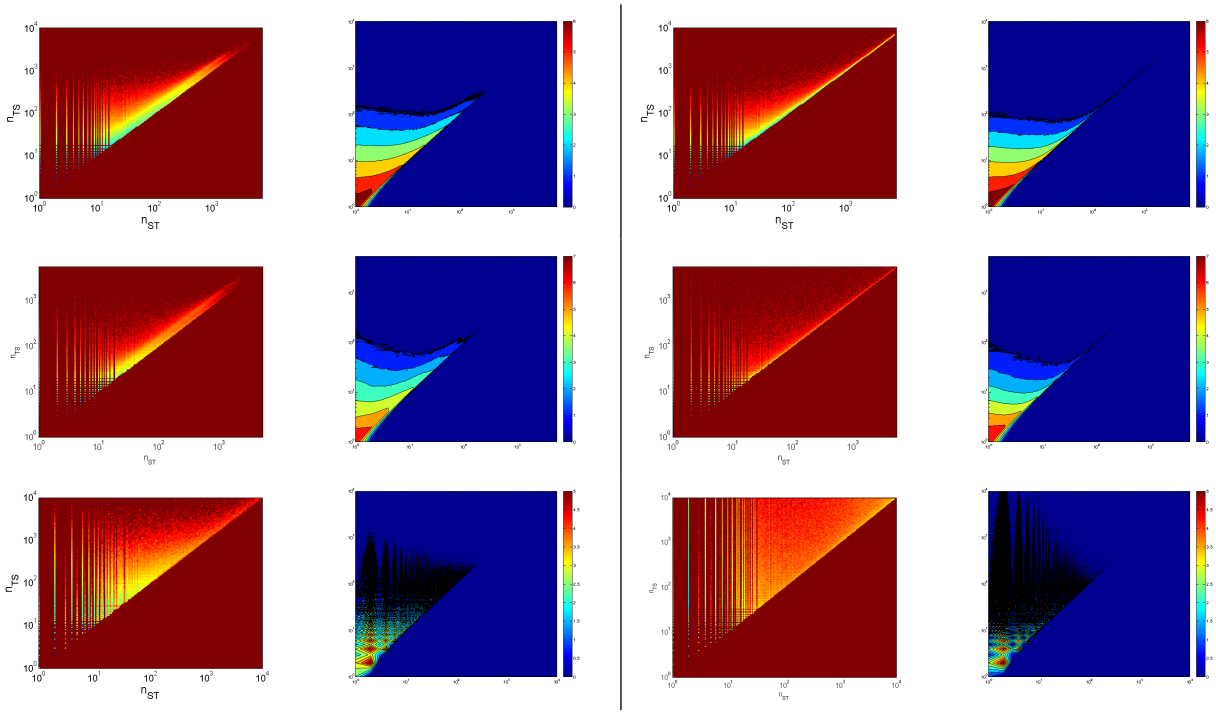


Figure 5: (left) n_{TS} versus n_{ST} for (from top to bottom) w_N, w_D, and w_{SMS}. (right) Synthetic data generated after fitting 3PL parameters to the real data for the same figures. Figures are best viewed in color.

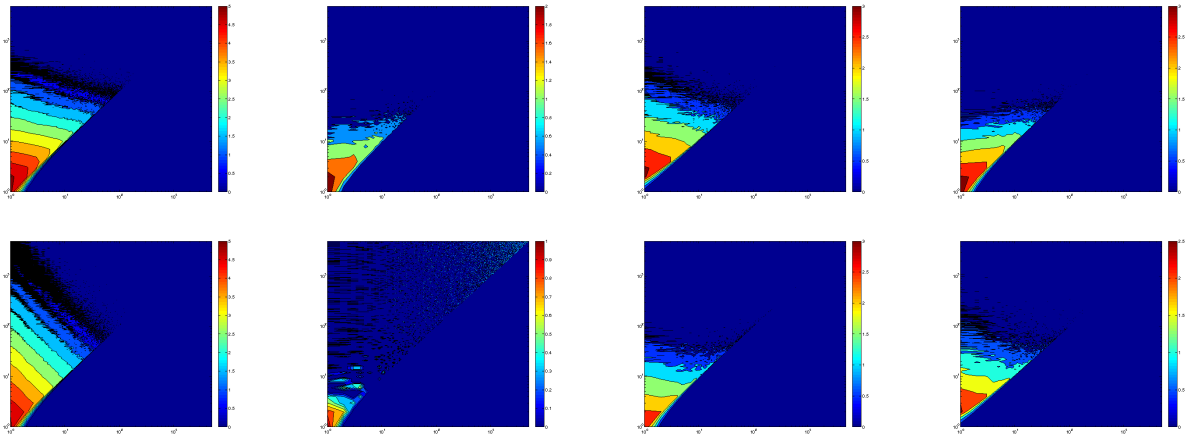


Figure 6: (top) Contour plots for the distribution of edges w.r.t. n_{TS} vs. n_{ST} for (from left to right) w^{ee}_{SMS}, w^{oo}_{SMS},w^{oe}_{SMS} and w^{eo}_{SMS}. (bottom) Synthetic data generated after fitting 3PL parameters to the real data for the same figures. Figures are best viewed in color.

synthetic data generated in the bottom row of the same figure. Notice that the fit to w_{SMS} is comparatively more noisy. This is because the number of edge pairs in MTG ($\sim$2M) over which we estimated the parameters is much smaller compared to that of MCG ($\sim$42M) (Table 1).

5.2 Patterns in Node Degree and Strengths

In this section we study the association between the degree of a node and its strength. The question we want to answer is 'how does the strength of a node change with changing degree?' We show the scatter plot of strength s versus degree k for all the nodes in MCG in Figure 9, (a,d) using 2D heatmap, (b,e) contour plots and (c,f) 3D surface plots for weights s_N and s_D, respectively. We observe that the distribution of nodes with respect to their degree and strengths follow a similar pattern for both types of weights; (a similar pattern holds also for s_{SMS} but we omit the related plots for brevity).

Given the observed patterns in MCG and MTG, we want to model $P(k, s)$. In the following, we describe a framework for modeling and parameter fitting of the observed patterns.

As in the previous section, we start by looking at the strength distribution of the nodes with a specific degree, that is $P(s|k = x)$, for x up to 100. We show the distributions for x ={1,5,10,20} in Figure 7 for (top) s_N and (bottom) s_D. Here, we observe that $P(s|k)$ follows a log-normal distribution with mean μ and standard deviation σ that change with degree k. To further understand how the μ and σ change with increasing k, we plot the estimated μ and σ values versus k in Figure 8(b,c) for (top) s_N and (bottom) s_D, respectively. We see that we can fit least-squares curves to model the changing value of μ and σ as a function of k. We will denote the value of μ for a specific k as $f_\mu(k)$ and that of σ as $f_\sigma(k)$. In Figure 8(a), we see that the distribution of degrees in MCG, that is $P(k)$, also obeys a log-normal distribution. All in all, we can model $P(k, s) = P(k)P(s|k)$ with a what we call Cascaded Log-Normal function.

MODEL 2 (CASCADED LOG-NORMAL (CLN)). *In phone and SMS networks, the number of nodes with degree k and strength s follows a Cascaded Log-Normal in the form of*

$$P(k,s) = P(k)P(s|k) \propto \boxed{LN(k,\mu_k,\sigma_k)LN(s-k,f_\mu(k),f_\sigma(k))}.$$

where μ_k and σ_k denote the mean and the variance of the log-normal distribution of degrees k (Figures 8(a)) and $f_\mu(k)$ and $f_\sigma(k)$ denote the mean and variance of the log-normal distributions of strengths s given a particular degree k, which changes as a function of k (Figure 8(b,c)).

Given the above CLN function with corresponding parameters, we generated synthetic graphs with the same number of nodes for each graph we studied. We show the related plots for the synthetic graphs in Figure 9(left) for s_N and (right) for s_D, next to each figure. We note that the synthetic distributions could model the real distributions qualitatively well.

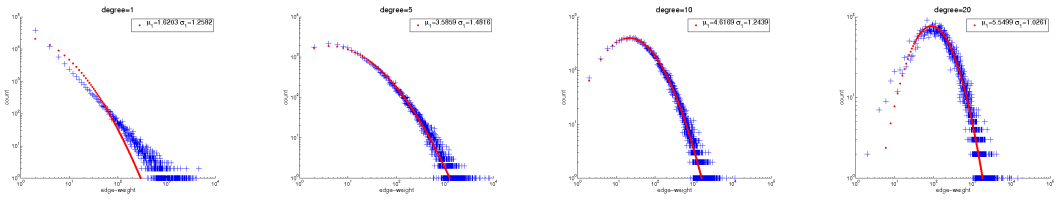
6. CONCLUSIONS

In this paper, we analyzed reciprocity and weights in both phone (MCG) and SMS (MTG) networks of millions of mobile phone users in a large city. The three major conclusions of our work are the following:

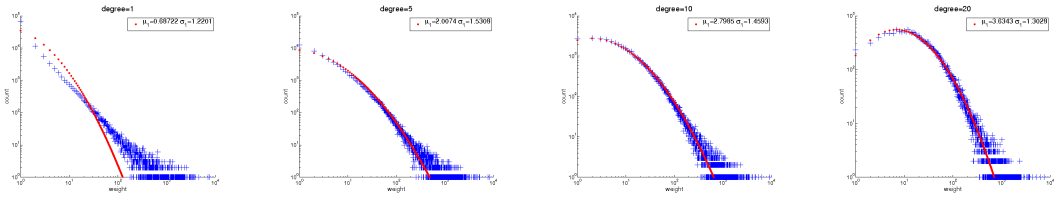
1. We found bivariate patterns in (1) the distribution of the weights on reciprocal edges; and (2) the distribution of the degree and strengths of nodes in MCG and MTG for all three types of weights w_N, w_D and w_{SMS}.
2. We proposed two bivariate functions (1) the 3PL to model the joint distribution $P(k, s)$ of nodes with degree k and strength s; and (2) the CLN to model the joint distribution $P(n_{ST}, n_{TS})$ of reciprocal edge pairs (e_{ST}, e_{TS}) with respective weights (n_{ST},n_{TS}). We further described a parameter fitting routine for the proposed functions.
3. We showed that our proposed functions and fitting of parameters could model the real distributions in MCG and MTG up to 2M nodes and 42M mutual edges well.

Additional contributions can be listed as follows:

1. We observed that 2D visualization of skewed bivariate distributions in real data could be misleading due to high overplotting; and
2. We showed the conundrum between the mean and the median values in the MCG and MTG which give different results and therefore provide different, even conflicting insights.

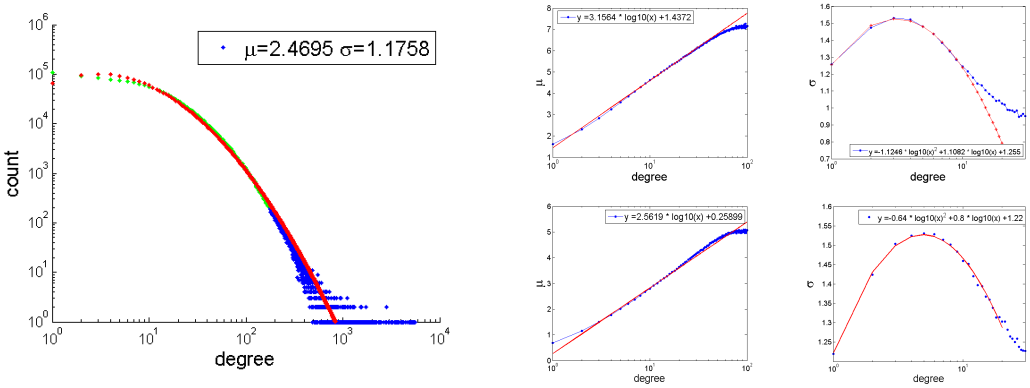


(a) s_N



(a) s_D

Figure 7: Log-normal fits to the distribution of node strengths given degree={1,5,10,20} for (top) s_N, and (bottom) s_D. Notice that μ increases, while σ increases and then decreases with increasing k.



(a) degree distribution of nodes in MCG (b) μ vs degree (c) σ vs degree

Figure 8: (a) $P(k) \propto$ LN(2.46, 1.17). (b) μ and (c) σ of $P(s|k)$ changes with k similarly for (top) s_N and (bottom) for s_D.

7. REFERENCES

[1] J. Abello, A. L. Buchsbaum, and J. Westbrook. A functional approach to external graph algorithms. In *ESA*, pages 332–343, 1998.

[2] W. Aiello, F. Chung, and L. Lu. A random graph model for massive graphs. In *STOC '00: Proceedings of the thirty-second annual ACM symposium on Theory of computing*, pages 171–180, New York, NY, USA, 2000. ACM.

[3] R. Albert and A.-L. Barabasi. Emergence of scaling in random networks. *Science*, pages 509–512, 1999.

[4] L. A. N. Amaral, A. Scala, M. Barthélémy, and H. E. Stanley. Classes of small-world networks. In *Proceeding of the National Academy of Sciences*, 2000.

[5] Z. Bi, C. Faloutsos, and F. Korn. The "DGX" distribution for mining massive, skewed data. *KDD*, Aug. 2001. Runner up for Best Paper Award.

[6] A. Broder, R. Kumar, F. Maghoul1, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, and J. Wiener. Graph structure in the web: experiments and models. In *WWW Conf.*, 2000.

[7] D. Carr, A. Olsen, and D. White. Hexagon mosaic maps for the display of univariate and bivariate geographical data. *Cartograph. Geograph. Information Systems*, 19:228–231, 1992.

[8] D. Chakrabarti, Y. Zhan, and C. Faloutsos. R-MAT: A recursive model for graph mining. *SIAM Int. Conf. on Data Mining*, Apr. 2004.

[9] A. Clauset, C. R. Shalizi, and M. E. J. Newman. Power-law distributions in empirical data. *SIAM Rev.*, 51(4):661–703, 2009.

[10] N. Du, C. Faloutsos, B. Wang, and L. Akoglu. Large human communication networks: patterns and a utility-driven generator. In *KDD '09*, pages 269–278, New York, NY, USA, 2009. ACM.

[11] N. Eagle. Behavioral inference across cultures: Using telephones as a cultural lens. *IEEE Intelligent Systems*, 23:62–64, 2008.

[12] 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.

[13] M. Faloutsos, P. Faloutsos, and C. Faloutsos. On power-law relationships of the internet topology. *SIGCOMM*, pages 251–262, Aug-Sept. 1999.

[14] H. Fuchs, M. Levoy, and S. M. Pizer. Interactive

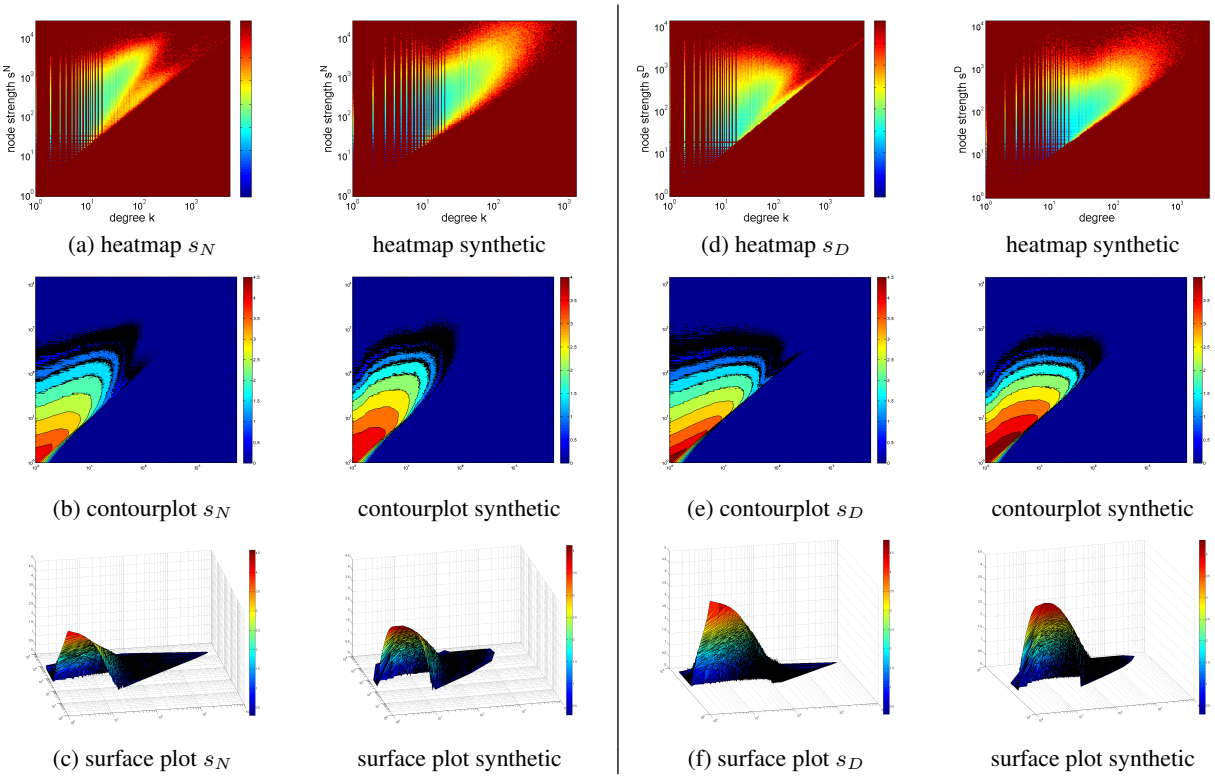


Figure 9: Distribution of nodes w.r.t. their strength s and degree k shown with (a,d) 2D heatmaps, (b,e) contour plots and (c,f) 3D surface plots. Next to each figure is the corresponding synthetic data generated after fitting CLN parameters to real data.

visualization of 3d medical data. *IEEE Computer*, 22(8):46–51, Aug. 1989.

[15] A. J and B. JAC. The lognormal distribution. 1957.

[16] J. Kleinberg, S. R. Kumar, P. Raghavan, S. Rajagopalan, and A. Tomkins. The web as a graph: Measurements, models and methods. In *Proceedings of the International Conference on Combinatorics and Computing*, 1999.

[17] S. Kotz, N. Balakrishnan, and N. L. Johnson. Continuous multivariate distributions, volume 1, models and applications, 2nd edition. 2000.

[18] J. Leskovec, J. Kleinberg, and C. Faloutsos. Graphs over time: densification laws, shrinking diameters and possible explanations. In *Proc. of ACM SIGKDD*, pages 177–187, Chicago, Illinois, USA, 2005. ACM Press.

[19] J. Leskovec, M. McGlohon, C. Faloutsos, N. Glance, and M. Hurst. Cascading behavior in large blog graphs: Patterns and a model. In *Society of Applied and Industrial Mathematics: Data Mining*, 2007.

[20] S. Nadarajah. A bivariate pareto model for drought. *Stochastic Environmental Research and Risk Assessment*, 23:811–822, Aug. 2009.

[21] A. 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 *CIKM '06*, pages 435–444, New York, NY, USA, 2006. ACM.

[22] M. E. J. Newman. Power laws, pareto distributions and zipf's law, December 2004.

[23] J.-P. Onnela, J. Saramaki, J. Hyvonen, G. Szabo, M. A. de Menezes, K. Kaski, A.-L. Barabasi, and J. Kertesz. Analysis of a large-scale weighted network of one-to-one human communication, 2007.

[24] J. P. Onnela, J. Saramaki, J. Hyvonen, G. Szabo, D. Lazer, K. Kaski, J. Kertesz, and A. L. Barabasi. Structure and tie strengths in mobile communication networks, 2006.

[25] D. M. Pennock, G. W. Flake, S. Lawrence, E. J. Glover, and L. C. Giles. Winners don't take all: Characterizing the competition for links on the web. *Proceedings of the National Academy of Sciences*, 99(8):5207–5211, 2002.

[26] S. Redner. Citation statistics from more than a century of physical review, Oct 2004.

[27] P. Reilly. Data visualization in archeology. *IBM Systems Journal*, 28(4):569–579, 1989.

[28] J. F. Rodrigues, Jr., H. Tong, A. J. M. Traina, C. Faloutsos, and J. Leskovec. Gmine: a system for scalable, interactive graph visualization and mining. *VLDB*, pages 1195 – 1198, 2006.

[29] D. Scott. Multivariate density estimation: Theory, practice, and visualization. 1992.

[30] M. Seshadri, S. Machiraju, A. Sridharan, J. Bolot, C. Faloutsos, and J. Leskove. Mobile call graphs: beyond power-law and lognormal distributions. In *KDD '08*, pages 596–604, New York, NY, USA, 2008. ACM.

[31] Q. Ye, B. Wu, L. Suo, T. Zhu, C. Han, and B. Wang. Telecomvis: Exploring temporal communities in telecom networks. In *ECML PKDD '09*, pages 755–758, 2009.

[32] S. Yue. The bivariate lognormal distribution to model a multivariate flood episode. *Hydrological Processes*, 14:2575–2588, Oct. 2000.