Quantifying Reciprocity in Large Weighted Communication Networks

Leman Akoglu[1,3], Pedro O.S. Vaz de Melo[2,3], and Christos Faloutsos[1,3]

[1] Carnegie Mellon University, School of Computer Science
[2] Universidade Federal de Minas Gerais
[3] iLab, Heinz College
{lakoglu,christos}@cs.cmu.edu, olmo@dcc.ufmg.br

Abstract. If a friend called you 50 times last month, how many times did you call him back? Does the answer change if we ask about SMS, or e-mails? We want to quantify reciprocity between individuals in weighted networks, and we want to discover whether it depends on their topological features (like degree, or number of common neighbors). Here we answer these questions, by studying the call- and SMS records of *millions* of mobile phone users from a large city, with more than 0.5 *billion* phone calls and 60 *million* SMSs, exchanged over a period of six months. Our main contributions are: (1) We propose a novel distribution, the *Triple Power Law* (3PL), that fits the reciprocity behavior of all 3 datasets we study, with a better fit than older competitors, (2) 3PL is parsimonious; it has only three parameters and thus avoids over-fitting, (3) 3PL can spot anomalies, and we report the most surprising ones, in our real networks, (4) We observe that the *degree* of reciprocity between users is correlated with their local topological features; reciprocity is higher among mutual users with larger local network overlap and greater degree similarity.

1 Introduction

One of the important aspects in human relations is the reciprocity, a.k.a. mutuality. Reciprocity can be defined as the tendency towards forming mutual connections with one another by returning similar acts, such as email and phone calls. In a highly reciprocal relationship both parties share equal interest in keeping up their relationship, while in a relationship with low reciprocity, one person is much more active than the other.

It is important to understand the factors that play role in the formation of reciprocity as there exists evidence that reciprocal relationships are highly probable to persist in the future [6]. Also, [18] shows that reciprocity related behaviors provide good features for ranking and classification based methods for trust prediction. Reciprocity plays other important roles in social and communication networks. For example, if the network supports a propagation process, such as spreading of viruses in email networks or spreading of information and ideas in social networks, then the presence of mutual links clearly speeds up the propagation. Non-existence of reciprocal links can also reveal unwanted calls and emails in spam detection.

Despite its importance, reciprocity has remained an under-explored dynamic in networks. Most work in network science and social network analysis focus on node level degree distributions [5,10,14], communities [20,22,19], and triadic relations, such as

P.-N. Tan et al. (Eds.): PAKDD 2012, Part II, LNAI 7302, pp. 85–96, 2012.

clustering coefficients and triangle closures [12]. The study of *dyadic* relations [26] and the related *bivariate* distributions they introduce is, however, mostly overlooked, and thus is the focus of this paper. Our motivation is grouped into two topics:

M1. Modeling bivariate distributions in real data: Two vital components of understanding data at hand are studying the simple distributions in it and visualizing it [27]. The study of reciprocity introduces bivariate distributions, such as the distribution $\Pr(w_{ij}, w_{ji})$ of edge weights on mutual edges, where association between *two* quantitative variables needs to be explored. A vast majority of existing work focus on *univariate* distributions in real data such as power-laws [8], log-normals [4], and most recently DPLNs [21], however the study of *multivariate* distributions has limited focus.

In addition, visualization of multivariate data in 2D is hard and often misleading due to issues regarding over-plotting. More importantly, mere visualization does not provide a compact data representation as opposed to data modeling. Summarization via aggregate functions such as the average or the median loses a lot of information and is also not representative, especially for skewed distributions as found in real data.

Models, on the other hand, provide compact data representations by capturing the patterns in the data, and are ideal tools for applications like data compression and anomaly detection.

M2. A weighted approach to reciprocity: Traditional work [11] usually study reciprocity on directed, *unweighted* networks as a *global* feature which is quantified as the ratio of the number of mutual links pointing in both directions to the total number of links. Defining reciprocity in such an unweighted fashion, however, prevents understanding the *degree* of reciprocity between mutual dyads. In a weighted network, even though two nodes might have mutual links between them, the skewness and the magnitude of the weights associated with these links would contain more information about how much reciprocity is really there between these nodes. For example, in a phone call network the reciprocity between a mutual dyad where the parties make 80%-20% of their calls respectively is certainly different than that of a mutual dyad with 50%-50% share of their calls. In short, edge weights are crucial to study reciprocity as a property of each dyad rather than as a global feature of the entire network and give more insight into the *level* of mutuality.

In this paper, we analyze phone call and SMS records of 1.87 million mobile phone users from a large city collected over six months. The data consists of over half a billion phone calls and more than 60 million SMSs exchanged. Our contributions are:

1. We observe similar bivariate distributions $\Pr(w_{ij}, w_{ji})$ of mutual edge weights in the communication networks we study. We propose the Triple Power Law (3PL) function to model this observed pattern and show that 3PL fits the real data with millions of points very well. We statistically demonstrate that 3PL provides better fits than the well-known Bivariate Pareto and Bivariate Yule distributions. We also use 3PL to spot anomalies, such as a pair of users with low mutuality where one of the parties makes 99% of the calls during the entire working hours, non-stop.
2. We use weighted measures of reciprocity in order to quantify the degree of reciprocal relations and study the correlations between reciprocity and local topological features among user pairs. Our results suggest that mutual users with larger local network overlap and higher degree similarity exhibit greater reciprocity.

2 Related Work

Bivariate Distributions in Real Data: A vast majority of existing work focus on *univariate* distributions in real data such as power-laws, Pareto distributions and so on [17]. For example, the degree distribution has been found to obey a power-law in many real graphs such as the Internet Autonomous Systems graph [10], the WWW link graph [1], and others [5,7]. 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 and -1.5 for blog posts [14].

A recent comprehensive study [8] 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 [2] 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.

Other deviations from power-laws continue. Discrete Gaussian Exponential (DGX) [4] was shown to provide good fits to distributions in a variety of real world data sets such as the Internet click-stream data and usage data from a mobile phone operator. Most recently, [21] studied several phone call networks and proposed a new distribution called the Double Pareto Log-Normal (DPLN) that was used to separately model the per-user number of call partners, number of calls and number of minutes. Other related work on explaining and modeling the behaviour of phone network users include [6,9,16,23].

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 co-variation between two quantitative variables x_1 and x_2. Association is based on how two variables simultaneously change together, for example the total number of calls with respect to the number of call partners of users.

Unlike univariate distributions, the multivariate distributions have mostly been studied theoretically in mathematics and statistics [3]. On the other hand, analysis of especially skewed multivariate distributions *in real data* has attracted much less focus. Existing work includes [28], which uses the bivariate log-normal distribution to describe the joint distributions of flood peaks and volumes, as well as flood volumes and durations. Also, [15] 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.

Reciprocity in Unweighted Networks: Previous studies usually consider reciprocity as a global metric of a given directed network where reciprocity is quantified as $r=\frac{L^{\leftrightarrow}}{L}$, the ratio of the number of mutual links $L^{\leftrightarrow}$ pointing in both directions to the total number of links L. By definition, r=1 for a purely bidirectional network (e.g. collaboration networks) and r=0 for a purely unidirectional network (e.g. citation networks).

There are two issues with this definition. First, it depends on the density of the network; reciprocity is larger in a network with larger link density. Second, this definition treats the graph as unweighted, and thus fails to quantify the *degree* of reciprocity between mutual dyads. [11] combines this classical definition with the network density into a single measure which tackles the first problem, however the new measure still remains a global, unweighted metric and does not allow to study the degree of reciprocity.

3 Data Description

In this work, we study anonymous mobile communication records of millions of users collected over a period of six months, December 1, 2007 through May 31, 2008. The data set contains both phone call and SMS interactions.

From the whole six months' of activity, we build three networks, CALL-N, CALL-D and SMS, in which nodes represent users and directed edges represent phone call and SMS interactions between these users. CALL-N is a who-calls-whom network with edge weights denoting (1) total number of phone calls, CALL-D is the same who-calls-whom network with edge weights denoting (2) total duration of phone calls (aggregated in minutes), and SMS is a who-texts-whom network with edge weights denoting (3) total number of SMSs. Table 1 gives the data statistics. Global unweighted reciprocity is r=0.84 for CALL, and r=0.24 for SMS.

Table 1. Data statistics. The number of nodes N, the number of directed edges E, and the total weight W in the mutual and non-mutual CALL and SMS networks.

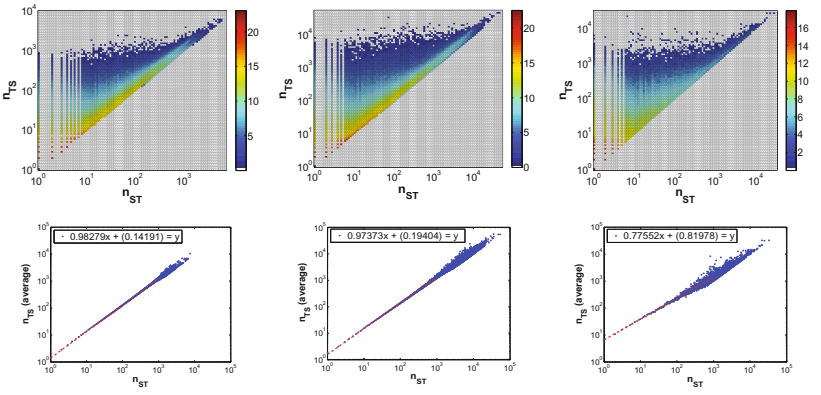
Network	N	E	W_N	$W_D(min)$	**Network**	N	E	W_{SMS}
CALL	1,87M	49,50M	483,7M	915x10^6	SMS	1,87M	8,80M	60,5M
CALL(mutual)	1,75M	41,84M	468,7M	885x10^6	SMS(mutual)	0,58M	2,10M	46,6M

4 Proposed Model: 3PL

Given a network of users with *mutual*, *weighted* edges between them, say CALL-N, and given two users i and j in the network, is there a relation between the number of calls i makes to j (w_{ij}) and the number of calls j makes to i (w_{ji})? In this section, we want to understand the association between the weights on the reciprocal edges in human communication networks and study their distribution $\Pr(w_{ij}, w_{ji})$ across mutual dyads. Since we study the pair-wise joint distribution, the order of the weights do not matter. Thus, to ease notation, we will denote the smaller of these weights as n_{ST} (for weight from Silent-to-Talkative) and the larger as n_{TS}, and will study $\Pr(n_{ST}, n_{TS})$.

Figure 1(top-row) shows the weights n_{TS} versus n_{ST} for all the reciprocal edges in (from left to right) CALL-N, CALL-D, and SMS. Each dot in the plots corresponds to a pair of mutual edges. Since there could be several pairs with the same (n_{ST}, n_{TS}) weights, the regular scatter plot of the reciprocal edge weights would result in over-plotting. Therefore, in order to make the densities of the regions clear, we show the *heatmap* of the scatter plots where colors represent the magnitude of volume (red means high volume and blue means low volume).

In Figure 1, we observe that most of the points are concentrated (1) around the origin and (2) along the diagonal for all three networks. Concentration around the origin, for example in CALL-N, suggests that the vast majority of people make only a few phone calls with $n_{ST}, n_{TS} < 10$, and much fewer people make many phone calls, which points to skewness. In addition, concentration along the diagonal indicates that mutual people call each other mostly in a balanced fashion with $n_{ST} \approx n_{TS}$. Notice that similar arguments hold for CALL-D and SMS.



(a) CALL-N (b) CALL-D (c) SMS

Fig. 1. (top-row) Scatter plot heatmaps: total weight n_{ST} (Silent to Talkative) vs the reverse, n_{TS}, in log scales. Visualization by scatter plots suffers from over-plotting. Heatmaps color-code dense regions but do not have compact representations or formulas. Figures are best viewed in color; red points represent denser regions. The counts are in $\log_2$ scale. (bottom-row) Aggregation by average: summarization and data aggregation, e.g. averaging, loses a lot of information.

Even though heatmaps reveal similar patterns in all the three networks, mere visualization does not provide compact representations for our data. One way to go around this issue is to do data summarization. For example, Figure 1(bottom-row) shows how n_{TS} changes with n_{ST} *on average*. The least square fit of the data points in log-log scales then provides a mathematical representation of the data. Data summarization by means of an aggregate function such as the average, however, loses a lot of information about the actual distribution: in our example, the slope of the least square fit in CALL-N is close to 1, which suggests that n_{TS} is equal to n_{ST} on average, and does not provide any information for the deviations. This issue arises mostly because aggregation by the average is not a good representative, especially for skewed distributions.

Given our observation that the distribution of reciprocal edge weights (n_{ST}, n_{TS}) follows a similar pattern across all three networks, how can we model the observed distributions? Since neither visualization nor aggregation qualify for compact data representation, we propose to formulate the distributions with the following bivariate functional form $\Pr(n_{ST}, n_{TS})$, which we call the Triple Power-Law (3PL) function.

Proposed Model 1 (Triple Power-Law (3PL)). *In human communication networks, the distribution* $\Pr(n_{ST}, n_{TS})$ *of mutual edge weights* n_{ST} *and* n_{TS} *(*n_{ST} *being the smaller) follows a Triple Power-Law in the following form*

$$\Pr(n_{ST}, n_{TS}; \alpha, \beta, \gamma) \quad \propto \boxed{\frac{n_{ST}^{-\alpha} n_{TS}^{-\beta} (n_{TS} - n_{ST} + 1)^{-\gamma}}{Z(\alpha, \beta, \gamma)}}, \alpha > 0, \beta > 0, \gamma > 0, \textit{ and}$$

$$n_{TS} \geq n_{ST} > 0,\ Z(\alpha, \beta, \gamma) = \textstyle\sum_{n_{ST}=1}^{M} \sum_{n_{TS}=n_{ST}}^{M} n_{ST}^{-\alpha} n_{TS}^{-\beta} (n_{TS} - n_{ST} + 1)^{-\gamma}.$$

where Z is the normalization constant and M is a very large integer.

Next we elaborate on the intuition behind the exponents α, β and γ.

Intuition behind the β Exponent: 3PL is the 2D extension of the "rich-get-richer" phenomenon; people who make many phone calls will continue making even more, and even longer ones, leading to skewed, power-law-like distributions. The β exponent is the skewness of the main component, the number n_{TS} of phone-calls from 'talkative' to 'silent'. High β means more skewed distribution; β=0 is roughly uniform distribution. As we show in Figure 1, there are many people who make only a few (and short) phone calls and only a few people who make many (and long) phone calls. Visually, the vast majority of people who make only a few phone calls are represented with the high density (dark red) regions around the origin in all three networks.

Intuition behind the α Exponent: Similarly, this indicates the skewness for n_{ST}, the number of silent-to-talkative phone-calls. High value of α means high skewness, while α close to zero means uniformity. Notice that $\alpha \approx 0$ for our real phone-call datasets (see Figure 2).

Intuition behind the γ Exponent: It captures the skewness in asymmetry. High γ means that large asymmetries are improbable. This is the case in all our real datasets. For example, in addition to the origin in Figure 1(a), the regions along the diagonal also have high densities. These regions correspond to mutual pairs with about equal interaction in both directions. This suggests that humans tend to reciprocate their communications. 3PL also captures this observation; notice that the probability is higher for n_{TS} close to n_{ST} and drops for larger inequality $(n_{TS} - n_{ST})$ as a power-law with exponent γ.

4.1 Comparison of 3PL to Competing Models

In this section, we compare our model with two well-known parametric distributions for skewed bivariate data, the Bivariate Pareto [13] and the Bivariate Yule [25]. Their functional forms are given as two alternative competitor models as follows.

Competitor Model 1 (Bivariate Pareto)

$$f_{X_1,X_2}(x_1, x_2) = k(k+1)(ab)^{k+1}(ax_1 + bx_2 + ab)^{-k-2}, x_1, x_2, a, b, k > 0.$$

Competitor Model 2 (Bivariate Yule)

$$f_{X_1,X_2}(x_1, x_2) = \frac{\rho_{(2)}(x_1 + x_2)!}{(\rho+1)_{(x_1+x_2+2)}}, x_1, x_2, \rho > 0; \alpha_{(\beta)} = \Gamma(\alpha+\beta)/\Gamma(\alpha), \alpha > 0, \beta \in R.$$

We use maximum likelihood estimation to fit the parameters of each model for each of our three networks. In Figure 2, we report the best-fit parameters as well as the corresponding data log likelihood scores (the higher, the better). Notice that for CALL-N and CALL-D the 3PL achieves higher data likelihood than both Bivariate Pareto and Bivariate Yule. On the other hand, for SMS, the data likelihood scores of all three models are about the same; with Bivariate Pareto giving a slightly higher score.

The simple sign of the difference between the log likelihoods (log likelihood ratio $\mathcal{R}$), however, does not on its own show conclusively that one distribution is better than the other as it is subject to statistical fluctuation. If its true value over many independent

data sets drawn from the same distribution is close to zero, then the fluctuations can easily change its sign and thus the results of the comparison cannot be trusted. In order to make a firm judgement in favor of 3PL, we need to show that the difference between the log likelihoods is sufficiently large and that it could not be the result of a chance fluctuation. To do so, we need to know the standard deviation σ on $\mathcal{R}$, which we estimate from our data using the method proposed in [24].

	CALL-N	CALL-D	SMS
Triple Power Law (3PL)			
α	1e-06	1e-06	0.8120
β	2.0703	1.8670	1.5896
γ	0.8204	0.9650	0.3005
Loglikelihood	-7.55e+07	-8.88e+07	-5.41e+06
Bivariate Pareto			
k	0.7407	0.7657	0.7862
a	0.2119	0.5723	0.7097
b	10e+05	1.25e+04	0.7553
Loglikelihood	-7.77e+07	-9.26e+07	-5.39e+06
z	803.73	975.75	-41.06
p	0	0	0
Bivariate Yule			
ρ	1.11e-16	5.55e-17	1e-06
Loglikelihood	-8.59e+07	-10.00e+07	-5.41e+06
z	2.14e+03	1.93e+03	1.49
p	0	0	0.03

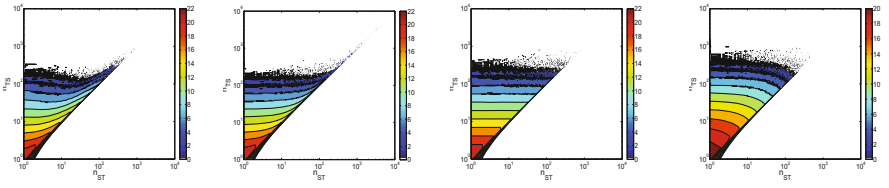
Fig. 2. Maximum likelihood parameters estimated for 3PL, Bivariate Pareto and the Bivariate Yule and data log-likelihoods obtained with the best-fit parameters. We also give the normalized log likelihood ratios z and the corresponding p-values. A positive (and large) z value indicates that 3PL is favored over the alternative. A small p-value confirms the significance of the result. Notice that 3PL provides significantly better fits to CALL and is as good as its competitors for SMS.

In Figure 2, we report the normalized log likelihood ratio denoted by $z = \mathcal{R}/\sqrt{2n}\sigma$, where n is the total number of data points (number of mutual edge pairs in our case). A positive z value indicates that the 3PL model is truly favored over the alternative. We also show the corresponding p-value, $p = erfc(z)$, where *erfc* is the complementary Gaussian error function. It gives an estimate of the probability that we measured a given value of $\mathcal{R}$ when the true value of $\mathcal{R}$ is close to zero (and thus cannot be trusted). Therefore, a small p value shows that the value of $\mathcal{R}$ is unlikely to be a chance result and its sign can be trusted.

Notice that the magnitude of z for CALL-N and CALL-D is quite large, which makes the p-value zero and shows that 3PL is a significantly better fit for those data sets. On the other hand, z is relatively much smaller for SMS, therefore we conclude that 3PL provides as good of a fit as its competitors for this data set. Note that the number of mutual edge pairs n in SMS ($\approx$1 million) is much smaller compared to that of the call networks ($\approx$21 million) (Table 1). It is worth emphasizing that difference, because the bivariate pattern of reciprocity might reveal itself better in larger data sets, and it would be interesting to see whether 3PL provides a better fit for SMS when more data samples become available.

Next, we demonstrate also visually that 3PL provides a better fit to the real data than its competitors. To this end, having estimated the model parameters for all three models, we generated synthetic data sets with the same number of samples as in each of our networks. We show the corresponding plots for CALL-N in Figure 3 (a) for real

data, and synthetic data generated by (b) 3PL, (c) Bivariate Pareto, and (d) Bivariate Yule. We notice that the simulated data distribution from 3PL looks more realistic than its two competitors. Similar results for CALL-D and SMS are omitted for brevity.



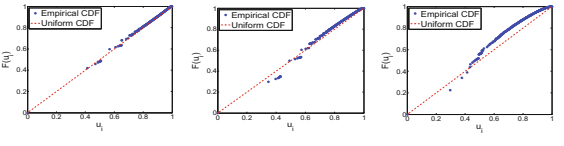
(a) CALL-N (real data) (b) 3PL (synthetic) (c) Biv. Pareto (synthetic) (d) Biv. Yule (synthetic)

Fig. 3. Contour-maps for the scatter plot n_{TS} versus n_{ST} in CALL-N (a) for real data, and synthetic data simulated from (b) 3PL, (c) Bivariate Pareto and (d) Bivariate Yule functions using the best-fit parameters. Notice that synthetic data generated by 3PL looks more similar to the real data than its competitors also visually. Counts are in $\log_2$ scale. Figures are best viewed in color.

4.2 Goodness of Fit

The likelihood ratio test is used to compare two models to determine which one provides a *better* fit to a given data. However, as we mentioned in the previous section, it cannot directly show when both competing models are poor fits to the data; it can only tell which is the least bad. Therefore, in addition to showing that 3PL provides a better (or as good) fit than its two competitors, we also need to demonstrate that it indeed provides a good fit itself.

A general class of tests for goodness of fit work by transforming the data points $(x_{1,i}, x_{2,i})$ according to a cumulative distribution function (CDF) F as $u_i = F(x_{1,i}, x_{2,i})$ for $\forall i, 1 \leq i \leq n$. One can show that if F is the correct CDF for the data, u_i should be *uniformly* distributed (derivation follows from basic probability theory). That is, if the CDF $\hat{F}$ estimated from our model is approximately correct, the empirical CDF of the $\hat{u}_i = \hat{F}(x_{1,i}, x_{2,i})$ should be approximately a straight line from $(0, 0)$ to $(1, 1)$.



(a) CALL-N (b) CALL-D (c) SMS

Fig. 4. Distribution of $\hat{u}_i = \hat{F}(x_{1,i}, x_{2,i})$ for all data points i according to cumulative distribution function (CDF) $\hat{F}$ estimated from our 3PL model. An approximately uniform distribution of $\hat{u}_i$ shows that 3PL provides a good fit to real data.

For each of our three data sets, we generate synthetic data drawn from our 3PL function with the corresponding estimated best-fit parameters. Then, we compute $\hat{u}_i = \hat{F}(x_{1,i}, x_{2,i})$ for all the data points in each of the data sets, where $\hat{F}$ is the estimated CDF from each synthetic data. In Figure 4, we show the CDF of $\hat{u}_i$ as well as the CDF for a perfect uniform distribution. Notice that the distribution of $\hat{u}_i$ is almost uniform for CALL-N and CALL-D, and quite close to the uniform for SMS. This corroborates our case that our model provides a good approximate to the correct CDF of our data sets, and thus indeed provides a *good* fit.

4.3 3PL at Work

There exist at least three levels at which we can make use of parametric statistical models for real data: (1) *as data summary*: compact mathematical representation, data reduction; (2) *as simulators*: generative tools for synthetic data; (3) *in anomaly detection*: probability density estimation.

In Figure 5(a), we show top 100 pairs in CALL-D with lowest 3PL likelihood (marked with triangles). Figure 5(b) shows the local neighborhood of one of the pairs, say A and B (marked with circles in (a)). We notice low mutuality; A initiated 99% of the calls in return to less than 2 hours total duration of calls B made. Further inspection revealed constant daily activity by A, including weekends, with about 7 hours call duration per day on average, starting at around 9am in the morning until around 5-8pm in the evening. It is also surprising that all these calls are addressed to the same contact, B. While for privacy reasons, we cannot fully tell the scenario behind this behavior, this proves to be an interesting case for the service operator to further look into. Other interesting anomalous observations are omitted for brevity.

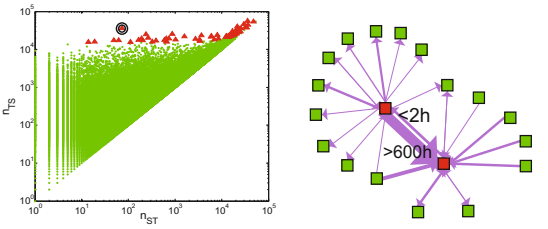


Fig. 5. (a) Least likely 100 points by 3PL (shown with triangles). (b) Local neighborhood of one mutual pair detected as an outlier (marked with circles). Edge thickness is proportional to edge weight.

5 Reciprocity and Local Network Topology

Given that person i calls person j w_{ij} times and person j calls person i w_{ji} times, what is the *degree* of reciprocity between them? In this section, we discuss several *weighted* metrics that quantify reciprocity between a given mutual pair. Later, we study the relationship between reciprocity among mutual pairs and their topological similarity.

5.1 Weighted Reciprocity Metrics

Three metrics we considered in this work to quantify the "similarity" or "balance" of weights w_{ij} and w_{ji} are (1) *Ratio* $r = \frac{\min(w_{ij},w_{ji})}{\max(w_{ij},w_{ji})} \in [0,1]$, (2) *Coherence* $c = \frac{2\sqrt{w_{ij}w_{ji}}}{(w_{ij}+w_{ji})} \in [0,1]$ (geometric mean divided by the arithmetic mean of the edge weights), and (3)*Entropy* $e = -p_{ij}\log_2(p_{ij}) - p_{ji}\log_2(p_{ji}) \in [0,1]$, where $p_{ij} = \frac{w_{ij}}{(w_{ij}+w_{ji})}$ and $p_{ji} = 1 - p_{ij}$. All these metrics are equal to 0 for the (non-mutual) pairs where one of the edge weights is 0, and equal to 1 when the edge weights are equal. Although these metrics are good at capturing the *balance* of the edge weights, they fail to capture the *volume* of the weights. For example, human would score $(w_{ji}$=100, w_{ij}=100) higher than $(w_{ji}$=1, w_{ij}=1), whereas all the metrics above would treat them as equal.

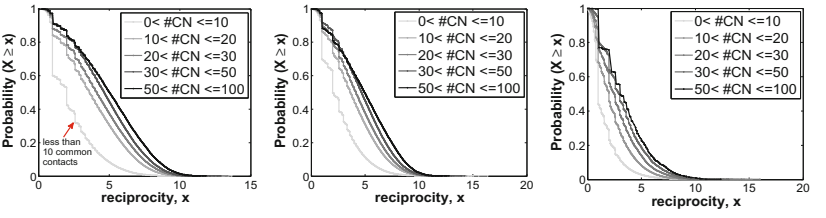
Therefore, we propose to multiply these metrics by the logarithm of the total weight, such that the reciprocity score consists of both a "balance" as well as a "volume" term. In

the rest of this section, we use the *weighted* ratio $r_w = \frac{\min(w_{ij},w_{ji})}{\max(w_{ij},w_{ji})} \log(w_{ij}+w_{ji})$ as the reciprocity measure in our experiments. The results are similar for the other weighted metrics, c_w and e_w.

5.2 Reciprocity and Network Overlap

Here, we want to understand whether there is a relation between the local network overlap (local density) and reciprocity between mutual pairs. Local network overlap of two nodes is simply the number of common neighbors they have in the network.

In Figure 6, we show the cumulative distribution of reciprocity separately for different ranges of overlap. The figures suggest that people with more common contacts tend to exhibit higher reciprocity, both in their SMS and phone call interactions.



(a) CALL-N (b) CALL-D (c) SMS

Fig. 6. Complementary cumulative distribution of reciprocity for different ranges of local network overlap (number of Common Neighbors). Notice that the more the number of common contacts, the higher the reciprocity.

5.3 Reciprocity and Degree Similarity

Next, we investigate the relation between the degree similarity (degree assortativity) and reciprocity. In Figure 7, we show the heatmap for the average reciprocity among pairs with respective degrees d_i and d_j for CALL-N (similar figures for other networks are omitted for brevity). The heatmap plot suggests that two people with more similar number of contacts exhibit larger reciprocity; notice the increase in reciprocity with increasing d_j for fixed d_i (from bottom to diagonal, towards degree similarity) and then the drop from diagonal to the right, towards degree dissimilarity.

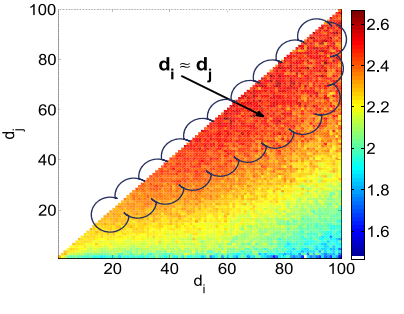


Fig. 7. Average reciprocity among dyads with degrees (d_i, d_j) in CALL-N

6 Conclusions

In this paper, we analyze more than 0.5 *billion* phone call and 60 *million* SMS records of millions of mobile phone users over six months; and study reciprocity; the distribution

and strength of mutual relations in weighted human communication networks. Our main contributions and findings are the following:

- **Patterns in joint pdf $\Pr(\mathbf{w_{ij}}, \mathbf{w_{ji}})$:** We find that the joint distribution $\Pr(w_{ij}, w_{ji})$ of the weights on mutual edges in mobile communication networks of users follow a bivariate pattern for all three types of weights; number of phone calls, duration of phone calls and number of SMSs. More specifically, the data points concentrate (1) around the origin as well as (2) along the diagonal in the scatter plot of w_{ij} versus w_{ji}. Observation (1) suggests a power-law like distribution in the amount of interactions; e.g., many people with few calls and only a few people with many calls. Observation (2) indicates that human communications are mostly reciprocal.
- **New model (3PL) for the joint pdf $\Pr(\mathbf{w_{ij}}, \mathbf{w_{ji}})$:** We propose the Triple Power Law (3PL) bivariate function to model this joint distribution. Our goodness of fit tests show that 3PL can model the observed distributions with more than 20 *million* mutual edge pairs quite well. We statistically demonstrate that it provides better fits than two other well-known bivariate distributions for skewed data, the Bivariate Pareto and the Bivariate Yule.
- **3PL at work:** 3PL provides a compact as well as a sparse data representation with only three parameters. We also show how to exploit 3PL to detect anomalies. Our case studies successfully reveal suspicious mutual interactions that agree with human intuition.
- **Weighted reciprocity:** Lastly, we take a weighted network approach and use weighted metrics to quantify the *degree* of reciprocity in human interactions. We observe that reciprocity is higher (1) for mutual pairs with larger local network overlap, that is, people with more common friends; and (2) for mutual pairs with larger degree-similarity, that is, people with similar number of contacts.

Acknowledgements. Research was sponsored by the National Science Foundation under Grant No. IIS1017415 and the Army Research Laboratory under Cooperative Agreement No. W911NF-09-2-0053. It is continuing through participation in the Anomaly Detection at Multiple Scales (ADAMS) program sponsored by the U.S. Defense Advanced Research Projects Agency (DARPA) under Agreements No. W911NF-11-C-0200 and W911NF-11-C-0088. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Laboratory, of the National Science Foundation, of the U.S. Government, or any other funding parties. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation here on.

References

1. Albert, R., Barabasi, A.-L.: Emergence of scaling in random networks. Science, 509–512 (1999)
2. Amaral, L.A.N., Scala, A., Barthélémy, M., Stanley, H.E.: Classes of small-world networks. Proceeding of the National Academy of Sciences (2000)
3. Arnold, B.C.: Bivariate distributions with pareto conditionals. Statistics & Probability Letters 5(4), 263–266 (1987)

4. Bi, Z., Faloutsos, C., Korn, F.: The "DGX" distribution for mining massive, skewed data. In: KDD (2001)
5. Broder, A., Kumar, R., Maghoul, F., Raghavan, P., Rajagopalan, S., Stata, R., Tomkins, A., Wiener, J.: Graph structure in the web: experiments and models. In: WWW (2000)
6. Cesar, C.R.-S., Hidalgo, A.: The dynamics of a mobile phone network. Physica A: Statistical Mechanics and its Applications 387(12), 3017–3024 (2008)
7. Chakrabarti, D., Zhan, Y., Faloutsos, C.: R-MAT: A recursive model for graph mining. In: SDM (2004)
8. Clauset, A., Shalizi, C.R., Newman, M.E.J.: Power-law distributions in empirical data. SIAM Rev. 51(4), 661–703 (2009)
9. Eagle, N., Pentland, A., Lazer, D.: Inferring social network structure using mobile phone data. Proceedings of the National Academy of Sciences (PNAS) 106, 15274–15278 (2009)
10. Faloutsos, M., Faloutsos, P., Faloutsos, C.: On power-law relationships of the internet topology. In: SIGCOMM, pp. 251–262 (August-September 1999)
11. Garlaschelli, D., Loffredo, M.I.: Patterns of Link Reciprocity in Directed Networks. Phys. Rev. Lett. 93, 268701 (2004)
12. Granovetter, M.: The strength of weak ties. Amer. Jour. of Sociology 78, 1360–1380 (1973)
13. Kotz, S., Balakrishnan, N., Johnson, N.L.: Continuous multivariate distributions, 2nd edn. Models and Applications, vol. 1 (2000)
14. Leskovec, J., McGlohon, M., Faloutsos, C., Glance, N., Hurst, M.: Cascading behavior in large blog graphs: Patterns and a model. In: SDM (2007)
15. Nadarajah, S.: A bivariate pareto model for drought. Stochastic Environmental Research and Risk Assessment 23, 811–822 (2009)
16. Nanavati, A.A., Gurumurthy, S., Das, G., Chakraborty, D., Dasgupta, K., Mukherjea, S., Joshi, A.: On the structural properties of massive telecom call graphs: findings and implications. In: CIKM 2006, pp. 435–444. ACM, New York (2006)
17. Newman, M.E.J.: Power laws, Pareto distributions and Zipf's law. Contemporary Physics 46(5), 323–351 (2005)
18. Nguyen, V.-A., Lim, E.-P., Tan, H.-H., Jiang, J., Sun, A.: Do you trust to get trust? a study of trust reciprocity behaviors and reciprocal trust prediction. In: SDM, pp. 72–83 (2010)
19. Nussbaum, R., Esfahanian, A.-H., Tan, P.-N.: Clustering social networks using distance-preserving subgraphs. In: ASONAM (2010)
20. Satuluri, V., Parthasarathy, S.: Scalable graph clustering using stochastic flows: applications to community discovery. In: KDD, pp. 737–746 (2009)
21. Seshadri, M., Machiraju, S., Sridharan, A., Bolot, J., Faloutsos, C., Leskovec, J.: Mobile call graphs: beyond power-law and lognormal distributions. In: KDD, pp. 596–604 (2008)
22. Tantipathananandh, C., Berger-Wolf, T., Kempe, D.: A framework for community identification in dynamic social networks. In: KDD, pp. 717–726 (2007)
23. Vaz de Melo, P.O.S., Akoglu, L., Faloutsos, C., Loureiro, A.A.F.: Surprising Patterns for the Call Duration Distribution of Mobile Phone Users. In: Balcázar, J.L., Bonchi, F., Gionis, A., Sebag, M. (eds.) ECML PKDD 2010, Part III. LNCS, vol. 6323, pp. 354–369. Springer, Heidelberg (2010)
24. Vuong, Q.H.: Likelihood ratio tests for model selection and non-nested hypotheses. Econometrica 57, 307–333 (1989)
25. Xekalaki, E.: The bivariate yule distribution and some of its properties. Statistics 17(2), 311–317 (1986)
26. Xiang, R., Neville, J., Rogati, M.: Modeling relationship strength in online social networks. In: WWW, pp. 981–990 (2010)
27. Yang, X., Asur, S., Parthasarathy, S., Mehta, S.: A visual-analytic toolkit for dynamic interaction graphs. In: KDD, pp. 1016–1024 (2008)
28. Yue, S.: The bivariate lognormal distribution to model a multivariate flood episode. Hydrological Processes 14, 2575–2588 (2000)