Patterns amongst Competing Task Frequencies: Super-Linearities, and the Almond-DG Model

Danai Koutra[1,⋆], Vasileios Koutras[2],
B. Aditya Prakash[3], and Christos Faloutsos[1]

[1] Computer Science Dept., Carnegie Mellon Univ.
{danai,christos}@cs.cmu.edu
[2] Dept. of Accounting & Finance, Athens Univ. of Econ. & Bus.
vkoutras@aueb.gr
[3] Computer Science Department, Virginia Tech.
badityap@cs.vt.edu

Abstract. If Alice has double the friends of Bob, will she also have double the phone-calls (or wall-postings, or tweets)? Our first contribution is the discovery that the relative frequencies obey a power-law (sub-linear, or super-linear), for a wide variety of diverse settings: tasks in a phone-call network, like count of friends, count of phone-calls, total count of minutes; tasks in a twitter-like network, like count of tweets, count of followees etc. The second contribution is that we further provide a full, digitized 2-d distribution, which we call the Almond-DG model, thanks to the shape of its iso-surfaces. The Almond-DG model matches all our empirical observations: super-linear relationships among variables, and (provably) log-logistic marginals. We illustrate our observations on two large, real network datasets, spanning $\sim 2.2M$ and $\sim 3.1M$ individuals with 5 features each. We show how to use our observations to spot clusters and outliers, like, e.g., telemarketers in our phone-call network.

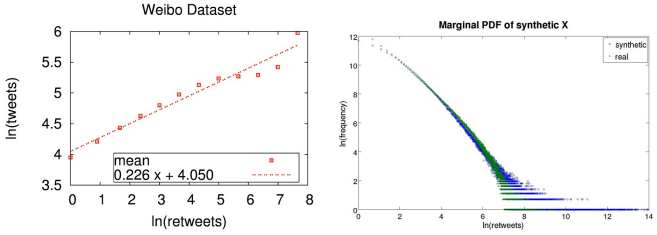
1 Introduction

If 'Alice' has 50 contacts and did 100 phonecalls to them, what should we expect for 'Bob', who has twice the contacts? One would expect a linear relationship (double the contacts, double the phonecalls). However, we show that in numerous settings, the relationship is a power law, being sub- or super-linear.

Useful as it may be for point estimates, the power-law relationship cannot give us any estimate for the variance. How would we model such joint distributions,

⋆ We would like to gratefully acknowledge the organizers of KDD Cup 2012 as well as Tencent Inc. for making the datasets available. Funding was provided by the U.S. Army Research Office (ARO) and Defense Advanced Research Projects Agency (DARPA) under Contract Number W911NF-11-C-0088. This work is also partially supported by an IBM Faculty Award and a grant from the VT College of Engineering. The content of the information in this document does not necessarily reflect the position or the policy of the Government, and no official endorsement should be inferred. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation here on.

J. Pei et al. (Eds.): PAKDD 2013, Part I, LNAI 7818, pp. 201–212, 2013.



(a) Super-linear Rel. Frequency (b) log-logistic marginals

Fig. 1. Illustration of super-linearity and goodness of our proposed ALMOND-DG. (a) Power-law relationship between count of tweets and count of retweets for each user in Tencent-Weibo (log-log scales). (b) Marginal PDF of the retweets, in log-log scales: real (in green triangles); generated by ALMOND-DG (in blue circles).

like, say, the number of contacts vs. number of phone-calls? What can we say about their marginals? They are definitely *not* Gaussian. Do they follow a Pareto (power-law) distribution in their marginals? What about the joint distribution?

The questions we want to answer here are:

1. **Q1: Patterns**: if 'Alice' executes task x (say, phone call) for n_x times, how many times n_y does she do task y (say, send an sms)?
2. **Q2: Distribution estimate**: What is the appropriate 2-d distribution to fit real, 2-d points (like, say <# tweets, # retweets> in the twitter setting)? Multivariate Gaussian fails miserably, due to heavy tails in real data.
3. **Q3: Practical use**: Can we answer "what-if" scenarios, and find anomalies?

Our contributions are exactly the answers to the above questions:

- **A1: Patterns:** We observe power law relationships between tasks competing for a person's resources (e.g., time).
- **A2: Distribution Estimate:** We propose the ALMOND-DG distribution, which uses the lesser-known tool of *copulas*, and has all the properties we observed in real data: the super-linear relationship, and also, log-logistic marginals (which are prevalent in many real-world datasets).
- **A3: Practical use:** Our ALMOND-DG fits several, diverse real datasets. We show how to spot outliers, and how to answer what-if questions.

We report results from two large, real, diverse network datasets. The first spans $\sim 3.1M$ users and is on a phone-call dataset; for each customer, we study the count of distinct contacts, phonecalls, text messages, and the total minutes. The second is from Tencent-Weibo network, a Chinese version of Twitter^{TM}, with count of tweets, re-tweets, followees etc. per user. Figure 1 illustrates our main ideas and discoveries: Figure 1(a) depicts the power-law relationship between count of tweets and count of retweets. Both axes are logarithmic; each red square is the conditional average of tweets, for the given count of retweets. The last few points are noisy, because of extreme-value effects (there are very few people with so many re-tweets, and they dominate the average). Figure 1(b) shows the marginal PDF of the retweets, again in log-log scales. The green triangles are the

real distribution, while the blue circles correspond to synthetic data, generated by our proposed ALMOND-DG. Notice that (i) the real distribution has a power-law tail, but it tilts in the beginning (top concavity) and (ii) our synthetic data fit well. Table 1 gives the major symbols we use and their definitions.

Table 1. Symbols and definitions

Symbol	Definition
$F_X(\cdot)$, $F_D(\cdot)$	cumulative distribution function (CDF) for: (a) random variable X or (b) distribution D (e.g., F_{LL} = CDF of log-logistic)
a_x, a_y	location parameter of log-logistic random variables X & Y
b_x, b_y	scale parameter of log-logistic random variables X & Y
$C(\cdot,\cdot)$	copula: 2 variable dependence function $[0,1] \times [0,1] \to [0,1]$
θ	parameter in Gumbel's copula that captures correlations between the random variables X & Y
SURF	Super-Linear Relative Frequency Observation
ALMOND	our 2-d continuous log-logistic distribution using Gumbel's copula
ALMOND-DG	our proposed distribution: the discretized and truncated version of ALMOND

2 Patterns and Observations

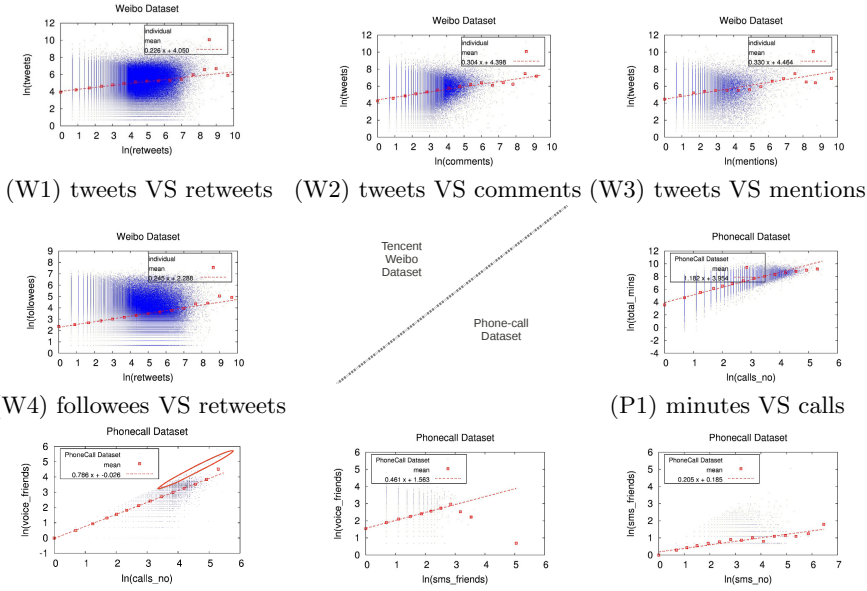
What happens to the number of tweets of a user if her re-tweets triple (Figure 1(a))? To answer such 'what-if' scenarios, we study two big, real networks, from which we extract 5 features for each user; each feature corresponds to the occurrences of a task:

- Tencent Weibo (W) [10]: one of the largest micro-blogging websites in China. For each of the ~ 2.2 million users we extracted five quantities: the number of her tweets, retweets, comments, mentions and followees.
- Phonecall dataset (P): phone-call records from a mobile provider in a big Asian city. For each of the ~ 3.1 million customers we obtained the number of her calls, messages, "voice" and "sms" friends, as well as the total minutes of her phone-calls.

In Fig. 2 we present the scatter plots of pairs of tasks. For each dataset we have $n = 5$ features. We are giving the plots for $n - 1$ pairs, instead of $\binom{n}{2}$, due to space limitations. Each user/customer is a blue point on the plane and is characterized by the number of times she did tasks 'A' and 'B'. All plots "suffer" from heavy over-plotting (not visible), especially for small values of occurrences. So, linear regression fails and we resort to the following solution: we group the points in logarithmic buckets and compute the mean (red points) of Y given X. The line $E[Y|X = x]$ is obtained by linear regression on the red points (ignoring the few last points, where the observations are extremely sparse, possibly due to the "horizon effect"). As we observe, in all cases the conditional expectation is a linear function of x. We call this sub- or super-linear relationship between the frequency of occurrence of the tasks SURF (Super-linear Relative Frequency).

Observation 1. *Deviations from the power-law pattern, as shown in Fig. 2(P2) are due to outliers, e.g. telemarketers.*

The customers within the red ellipse all have about 100 contacts, and 100 phone-calls, that is about one phone-call per contact. The rest of the population has many more phone-calls than contacts; thus, this difference in behavior leads to the suspicion that the former are telemarketers.



(W1) tweets VS retweets (W2) tweets VS comments (W3) tweets VS mentions

(W4) followees VS retweets (P1) minutes VS calls

(P2) voice friends VS calls (P3) voice VS sms friends (P4) sms friends VS sms

Fig. 2. SURF patterns in real datasets (log-log scale, "W": Weibo, and "P": Phonecall network): power-law relationship between competing tasks. In plot (P2), the 'anomalous' customers in the red ellipse deviate from the power-law pattern, having called each of their contacts only once; they are probably telemarketers.

3 Almond-DG Distribution

In order to model the observed patterns of the previous section, we need a probabilistic model; this model must satisfy the properties found in real 2-d distributions, including the important property found in the previous section, namely, the conditional average seems to follow a power law (Figure 1(a)). The rest of the paper focuses on two questions:

- **Q2.1**: Can we find additional properties of such 2-d (and more ambitiously, higher-d) distributions?
- **Q2.2**: Can we build a probabilistic model (i.e., find a 2-d PDF) that will fit most real clouds of points? It is clear that the multivariate Gaussian is heavily violated, even visually from Figure 1(b), and, as we show later, from the marginals of the x and y axis.

In summary, the answers are as follows:

- **A2.1**: Yes, the marginals of almost any attribute in our real datasets has a skewed distribution, and can be modeled well by a (truncated, digitized) log-logistic distribution.
- **A2.2**: Subject to log-logistic marginals, and power-law conditional averages, a possible candidate 2-d distribution is our proposed 'ALMOND' distribution (digitized and truncated).

We want a 2-d distribution whose iso-surfaces will resemble the ones in Figure 3. Since they have 'almond'-shape (see Fig. 3(c)), we name our proposed distribution as the 'ALMOND' distribution, and, after digitization and truncation, ALMOND-DG.

The final answer is given by the discretized form of Eq. (1) in p. 207, which we repeat here for convenience:

$$F_{ALM}(x, y; a_x, b_x, a_y, b_y, \theta) = e^{-\left(\left[\ln\left(1+(x/a_x)^{-b_x}\right)\right]^{\theta}+\left[\ln\left(1+(y/a_y)^{-b_y}\right)\right]^{\theta}\right)^{1/\theta}},$$

where θ captures the correlation between the random variables (attributes) X and Y, while a_x (a_y) and b_x (b_y) determine respectively the location ($\approx$ average/mode) and the spread ($\approx$ variance).

As we discuss next, this distribution has all the desired properties (the super-linear relationship, and provably, log-logistic marginals). The approach we follow for ALMOND-DG has two steps: (a) modeling the marginal (univariate) distributions of the tasks, and (b) combining them with the use of the so-called *copula*. Before we see the train of thought, we first give a property of the marginals (in response to question Q2, above), and then some definitions.

For Q1, given the overwhelming number of real-world (univariate) datasets that exhibit skewed distributions, one would expect that the marginals follow power-law or log-normal distributions. We found that this is *almost* true: an even better fit is provided by the so-called *log-logistic* distribution, which also accounts for the top concavity (see Fig. 1(b) p. 202, Fig. 4(a,b) p. 209, Fig. 5 p. 212). We proceed with some definitions.

3.1 Definitions

Log-logistic distribution. All the marginals we report match the so-called log-logistic distribution. Thus, we remind its definition next.

Definition 1 (CDF of log-logistic). *The log-logistic CDF is given by*

$$F_{LL}(x; a, b) = (1 + (x/a)^{-b})^{-1}, x \geq 0$$

where $a > 0$ is the scale parameter and $b > 0$ is the shape parameter.

By definition, a (continuous) random variable X follows the log-logistic distribution, $\mathcal{LL}(a, b)$, iff its logarithm $\ln X$ follows the logistic distribution $\mathcal{L}(\ln a, 1/b)$. Intuitively, the CDF of the logistic distribution is the sigmoid function – exactly

the one used for logistic regression, artificial neural networks, modeling product market penetration (Bass model), spread of epidemics (SI model), etc. A second property is that the odds-ratio of the log-logistic distribution, follows a power-law, and thus it is a straight line in log-log scales (see Fig. 5 in p. 212).

Moreover, the log-logistic distribution is related to the standard Pareto distribution: If $X \sim F_{LL}(x; 1, 1)$, then the shifted random variable $Z = X + 1$ follows the standard Pareto distribution.

Copulas for Modeling Dependence.
Our proposed 2-d log-logistic distribution (ALMOND-DG) is based on copulas. So, before we present the details of our distribution, we briefly give the main notions behind this powerful technique, which has been successfully used in survival models, financial risk management, and decision analysis.

In a nutshell, copulas are used for understanding and modeling dependence structures between random variables (e.g., $X = \#$ of phonecalls, $Y = \#$ of sms). More specifically, the copula function links the univariate margins (F_X, F_Y) with their full multivariate distribution. By construction, the latter (a) has the same marginals as X and Y, and (b) exhibits the correlation between them. Copulas have been proved very popular in statistical applications as they allow one to easily model and estimate the distribution of random vectors by estimating the marginals and copula separately. The major difference between the many parametric copula families cited in the literature, is their capability of assuming different dependence structures.

More formally, copulas are defined as follows:

Definition 2 (Copula). *Let X, Y be two random variables with marginal CDFs F_X and F_Y respectively. A copula $C(u, v)$ is a two-variable dependence function $C : [0, 1] \times [0, 1] \to [0, 1]$ that produces a joint CDF which captures the correlation between the F_X and F_Y variables, i.e., $F(x, y) = C(F_X(x), F_Y(y))$.*

The existence of such copula is guaranteed by Sklar's Theorem [17]. Note that if X, Y are independent variables, their joint CDF takes the form $F(x, y) = C(F_X(x), F_Y(y)) = F_X(x)F_Y(y)$. Hence the copula $C(u, v) = u\,v$ captures their independence.

The copulas are a powerful tool that can capture any type of dependence: positive, negative, none (independence). One of the prevailing families of copulas is the so-called Archimedean family; among its members we choose the so-called *Gumbel's* copula. It has been used successfully in several settings, for example, to model the dependence between indemnity payment (loss) and an allocated loss adjustment expense (e.g., lawyer's fees) in order to calculate reinsurance premiums [19]. Equally successfully, Gumbel's copula has been used to model the rainfall frequency as a joint distribution of rain characteristics (e.g., volume, peak, duration) [9]. Formally, it is defined as follows:

Definition 3 (Gumbel's Copula). *Gumbel's copula is given by $C(u, v) = e^{-\left[\phi(u)^\theta + \phi(v)^\theta\right]^{1/\theta}}$, where $\theta \geq 1$, and $\phi(t) = (-\ln t)^\theta$.*

3.2 Proposed Almond-DG Distribution

As we briefly mentioned in the previous section, the log-logistic distribution fits well the skewness of many real-world datasets. The reasons we pick the (digitized, truncated) log-logistic for the marginals, are the following:

- real data (#-mentions, #-phonecalls) have a linear log-odd plot (Fig. 5),
- the log-logistic is related to the Pareto distribution, and
- it captures the top concavity observed in numerous real-world 1-d datasets.

So, how can we model the distribution of pairs of tasks (e.g., number of likes and number of comments) competing for an individual's resources (e.g., time) now that we know that the distribution of each task is log-logistic or power-law-like? It turns out we have a lot of choices, since several 2d-logistic (not log-logistic) distributions with logistic marginals have been proposed in the literature. However, the majority of them (Malik and Abraham [12], Fang and Xu [7], etc) are not suitable in our case, since they are not flexible enough to capture the dependence between the variables that the real datasets exhibit.

Now we have all the ingredients (log-logistic marginals, a first step on how to combine them), and we only need to make sure that X and Y are positively correlated.

Notice that, in this section, we start from continuous distributions (like the log-logistic), and we use their digitized version ($floor()$), followed by truncation: whenever the result is "0" (say, zero phonecalls), we ignore it, since it won't register in our real datasets.

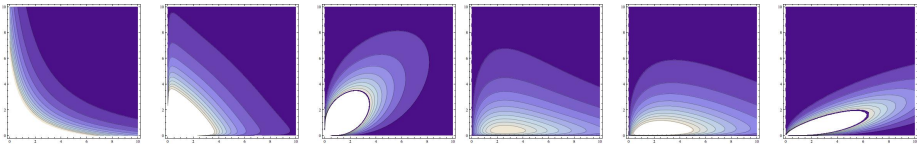
Consider two random variables X and Y (like 'number of phonecalls', and 'count of contacts' in our phonecall network). How can we model their joint PDF? Copulas provide a way to do that, when we know the marginals F_X and F_Y. Sklar's theorem [17] states that we can always find a 2-d function $C(u, v)$ that can model the joint CDF. We use Gumbel's copula with parameter θ, and log-logistic marginals with different parameters each. So, our proposed 'ALMOND' distribution has 5 parameters and is defined below:

Definition 4 ('Almond' Distribution). *The CDF of* ALMOND*, our proposed continuous 2-d log-logistic distribution, is given for $x, y \geq 0$ by (1), where θ is a parameter that captures the dependence between X and Y.*

$$F_{ALM}(x, y; a_x, b_x, a_y, b_y, \theta) = e^{-\left(\left[\ln\left(1+(x/a_x)^{-b_x}\right)\right]^{\theta}+\left[\ln\left(1+(y/a_y)^{-b_y}\right)\right]^{\theta}\right)^{1/\theta}} \quad (1)$$

For illustration purposes, in Fig. 3, we give some examples of contour plots of our ALMOND Distribution. The following observation is useful for estimating the parameter θ from the real data.

Observation 2. *For a pair of random variables (X, Y) that follows the* ALMOND *distribution, θ can be estimated by $\theta = (1 - \tau)^{-1}$, where τ is Kendall's tau rank correlation coefficient between X and Y. In practice, for efficiency, we use Spearman's coefficient ρ instead of τ.*



(a) $\theta = 1$ (b) $\theta = 1.67$ (c) $\theta = 3.33$ (d) $\theta = 1$ (e) $\theta = 1.26$ (f) $\theta = 3.33$

Fig. 3. Contour plots of the ALMOND distribution with parameters $a_x = a_y = 1$, $b_x = b_y = 1$ for (a)-(c); and $a_x = 6.5, a_y = 2.1$, $b_x = 1.6, b_y = 1.27$ – as in the "comments VS mentions" dataset – for (d)-(e)

Lemma 1. *The marginals of the* ALMOND *distribution are log-logistic distributions.*

Proof. By taking the limit of y to infinity, we obtain the marginal of X:

$$\lim_{y \to \infty} F(x, y) = F_X(x; a_x, b_x) = \frac{1}{1 + (x/a_x)^{-b_x}}$$

Hence, $X \sim \mathcal{LL}(a_x, b_x)$. Similarly, we can show that $Y \sim \mathcal{LL}(a_y, b_y)$. □

Observation 3 (Special case). *If $\theta = 1$, then $C(u, v) = uv$, and X, Y are independent log-logistic random variables. The CDF of* ALMOND *becomes then*

$$F(x, y; a_x, b_x, a_y, b_y) = \left(1 + (x/a_x)^{-b_x} + (y/a_y)^{-b_y} + (x/a_x)^{-b_x} (y/a_y)^{-b_y}\right)^{-1}.$$

The definition of our proposed digitized, truncated distribution follows:

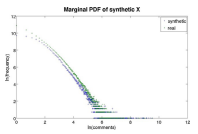
Definition 5 (Almond-DG Distribution). *If (X, Y) follows the* ALMOND *distribution, then the discrete bivariate random variable $(floor(X), floor(Y))$ given that $X \geq 1$ and $Y \geq 1$ follows the* ALMOND-DG *distribution.*

Essentially, ALMOND-DG is derived from ALMOND by discretizing its values and rejecting the pairs with either $X = 0$ or $Y = 0$.

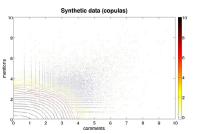
4 Goodness of Fit

In this section, we show the goodness of fit of our ALMOND-DG distribution in a qualitative manner. The interested reader may refer to the Appendix for information about the parameter fitting and generation of data following the ALMOND-DG distribution.

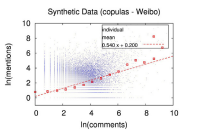
Note: Evaluating the goodness of fit for skewed distributions is a rather challenging task and the majority of methods seem to fail in real data. In [8], Johnson et al. explore several methods for evaluating the goodness of fit for *univariate* Pareto distributions with no clear winner. The difficulty in the evaluation increases even more in the case of bivariate distributions, which we are addressing



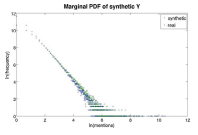
(a) Marginal PDF of comments (real: green / synthetic:blue)



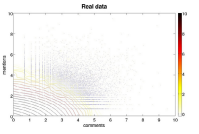
(c) Contour plot of synthetic data



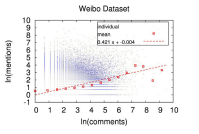
(e) Conditional means in synthetic data



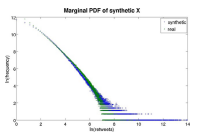
(b) Marginal PDF of mentions (real: green / synthetic:blue)



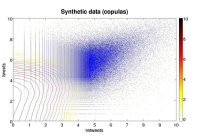
(d) Contour plot of real data



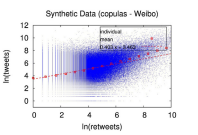
(f) Conditional means in real data



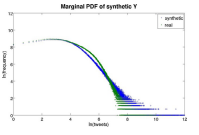
(a) Marginal PDF of tweets (real: green / synthetic:blue)



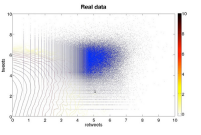
(c) Contour plot of synthetic data



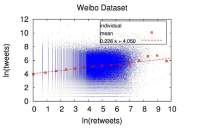
(e) Conditional means in synthetic data



(b) Marginal PDF of retweets (real: green / synthetic:blue)



(d) Contour plot of real data



(f) Conditional means in real data

Fig. 4. Goodness of fit of Almond-DG to the "mentions VS comments" (above) and "tweets VS retweets" datasets. (i) the log-logistic distribution fits well the marginals of the mentions and comments in the real Tencent Weibo network (plots a,b); (ii) the contour plots of the real (plots d) and synthetic 2-d datasets (plots c) have the same shape; (iii) both datasets obey the same power-law pattern, as shown in plots e,f.

in this paper. As we see in Fig. 4 , our proposed Almond-DG distribution fares pretty well. In the first column we see the marginal distributions of X (e.g., comments) and Y (e.g., mentions), i.e. $\ln(frequency)$. The green points represent the real data, while the blue points correspond to the distribution of the generated data with the estimated parameters (see Appendix).

Observation 4. *The real marginal distributions are captured well by our* Almond-DG *distribution, even when they are power-law-like.*

The second column of the figures holds the contour plots of the synthetic (c), and the real data points (d), while the last column shows the conditional means in synthetic (e) and real data (f).

Observation 5. *The contours of our truncated and digitized* ALMOND-DG *distribution with estimated parameters resemble the real contour lines.*

Observation 6. *The* SURF *pattern is preserved in the synthetic data; the real and synthetic data have similar power-law slopes.*

All in all, ALMOND-DG captures well the patterns found in the real-world data.

5 Related Work

Power Laws: Power laws have been discovered in numerous cases [3], often in conjunction with fractals and self-similarities [15]. Some of the most famous power laws are the Zipf distribution [20] and the Pareto distribution [14]. They have negative slopes, though. Power laws with positive slopes have also been discovered (length of coastlines, number of quad-tree blocks versus granularity [6]), and more recently in graphs ([11], [18], [13]). Akoglu et. al. [1] proposed the Triple Power Law (3PL), a bivariate distribution, to model reciprocity in phone-call networks. However, the model is bound to power-law distributions, which is not always the case in real networks.

Logistic and Log-Logistic Distributions: They have been studied extensively, in the continuous, univariate setting. The multivariate setting has been studied for the logistic distribution [12]. The discretized version of the univariate case has been shown to be a good fit for the duration of phonecalls by real users [4]. Earlier work [2] tried to fit discretized lognormals, or the so-called doubly-Pareto Lognormal [16].

However, none of the above articles provided a solution to our setting, namely, a 2-d distribution, with an explanation for the super-linearity we observe, and validation on several, diverse datasets.

6 Conclusions

The contributions are the answers to the questions we posed in the introduction: Q1: what can we say about the relative frequency of two tasks that compete for an individual's resources (e.g., phone calls vs. number of sms)? Q2: how can we model the corresponding 2-d clouds of points? Q3: how can we put our observations and developments to practical use.

Specifically, our contributions are:

1. **A1 [Patterns]:** Discovery of power law (SURF) in several, real, diverse network datasets, on most of their n-choose-2 pairs of attributes/tasks.
2. **A2 [Distribution Estimate]:** A new distribution, the ALMOND distribution, that describes well the skewed multivariate distributions and explains super-linearity, marginals and conditionals in real, diverse network datasets, on most of their n-choose-2 pairs of attributes/tasks.
3. **A3 [Practical Use]:** Illustration that ALMOND can be used for anomaly detection (Fig. 2(P2)), clustering and what-if scenarios.

References

1. Akoglu, L., Vaz de Melo, P.O.S., Faloutsos, C.: Quantifying reciprocity in large weighted communication networks. In: Tan, P.-N., Chawla, S., Ho, C.K., Bailey, J. (eds.) PAKDD 2012, Part II. LNCS (LNAI), vol. 7302, pp. 85–96. Springer, Heidelberg (2012)
2. Bi, Z., Faloutsos, C., Korn, F.: The "DGX" distribution for mining massive, skewed data. In: KDD (August. 2001)
3. Clauset, A., Shalizi, C.R., Newman, M.E.J.: Power-law distributions in empirical data. SIAM Rev. 51(4), 661–703 (2009)
4. 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 (LNAI), vol. 6323, pp. 354–369. Springer, Heidelberg (2010)
5. Embrechts, P., Lindskog, F., McNeil, A.: Modelling dependence with copulas and applications to risk management. In: Handbook of Heavy Tailed Distributions in Finance, pp. 331–385 (2003)
6. Faloutsos, C., Gaede, V.: Analysis of the z-ordering method using the hausdorff fractal dimension. In: VLDB (September 1996)
7. Fang, K.-T., Xu, J.-L.: A class of multivariate distributions including the multivariate logistic. Journal of Mathematical Research and Exposition 9, 91–98 (1989)
8. Johnson, N., Kotz, S., Balakrishnan, N.: Continuous Univariate Distributions, 2nd edn. Wiley (1995)
9. Karmakar, S., Simonovic, S.: Bivariate flood frequency analysis: Part 1. determination of marginals by parametric and nonparametric techniques. Journal of Flood Risk Management 1, 190–200 (2008)
10. KDD-Cup. Tencent Weibo Dataset (2012), `http://www.kddcup2012.org`
11. Leskovec, J., Kleinberg, J.M., Faloutsos, C.: Graphs over time: densification laws, shrinking diameters and possible explanations. In: KDD, pp. 177–187 (2005)
12. Malik, H.J., Abraham, B.: Multivariate logistic distributions. Annals of Statistics 1, 588–590 (1973)
13. McGlohon, M., Akoglu, L., Faloutsos, C.: Weighted graphs and disconnected components: patterns and a generator. In: KDD, pp. 524–532 (2008)
14. Pareto, V.: Oeuvres Completes. Droz, Geneva (1896)
15. Schroeder, M.: Fractals, Chaos, Power Laws: Minutes from an Infinite Paradise. W.H. Freeman and Company, New York (1991)
16. 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)
17. Sklar, A.: Fonctions de répartition à n dimensions et leurs marges. Publ. Inst. Statist. Univ. Paris 8, 229–231 (1959)
18. Tsourakakis, C.E.: Fast counting of triangles in large real networks without counting: Algorithms and laws. In: ICDM, pp. 608–617 (2008)
19. Valdez, E.A.: Understanding relationships using copulas. North American Actuarial Journal 2, 1–25 (1998)
20. Zipf, G.: Human Behavior and Principle of Least Effort: An Introduction to Human Ecology. Addison Wesley, Cambridge (1949)

Appendix: Fitting and Data Generation

Most of the estimation methods (e.g., MLE, MOM, etc.) fail when we have skewed, truncated, and digitized distributions with outliers. Fitting becomes even more difficult when fitting 2-d skewed distributions.

Fitting

We propose a fast method for fitting log-logistic (or power-law-like) data to the univariate log-logistic distribution. It is well known that if a random variable X follows the logistic distribution $\mathcal{L}(\mu, \sigma)$, then $-\ln\{odd\} = -\ln\left\{\frac{P(X\leq x)}{P(X>x)}\right\} = -\ln\left\{\frac{F_X(x)}{1-F_X(x)}\right\} = \frac{x-\mu}{\sigma}$.

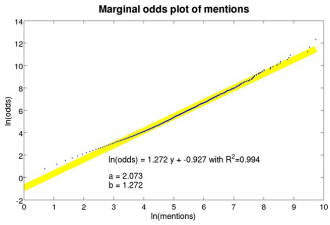


Fig. 5. "Log-odd plot" method for fitting

As depicted in Fig. 5, to fit the data via the "log-odd plot": (a) estimate the slope and intercept by applying linear regression on the "log-odd" plot, (b) solve $slope{=}\frac{1}{\sigma}$ and $intercept{=}-\frac{\mu}{\sigma}$ for μ and σ, and (c) compute the log-logistic parameters: $a_x = exp(\mu)$ and $b_x = \frac{1}{\sigma}$.

Generation of Synthetic Data

Step 1: Select $\theta = \frac{1}{1-\tau}$, where τ is the Kendall tau rank correlation coefficient.

Note: Since the computation of τ is prohibitive for the large datasets, being quadratic in the number of points, we use Spearman's ρ, which is a good substitute of τ, and robust to outliers in the data.

Step 2. Estimate the parameters using one of the traditional methods of parameter estimation (e.g., MLE, MOM) ; otherwise, use the "log-odd plot" method described above. In our experiments, we mainly used the MLE of the parameters.

Step 3: Generate two independent random variables s and u following the Uniform distribution $\mathcal{U}(0,1)$ and solve the equation $K(t) = u$ for t, where $K(t) = t - \frac{\phi(t)}{\phi'(t)} = t - \frac{1}{\theta} t \ln t$ (see Def. 3).

Step 4: Compute $x_1 = t^{s^{1/\theta}}$ and $y_1 = t^{(1-s)^{1/\theta}}$.

Note: Originally, $x_1 = \phi^{-1}(s\phi(t))$ and $y_1 = \phi^{-1}((1-s)\phi(t))$.

By starting from a general algorithm for copula-based data generation [5] and using the formulas related to Gumbel's copula, we obtain the formulas given in steps 3-5. Essentially, up to this point we have two uniform $\mathcal{U}(0,1)$ random variables correlated according to Gumbel's copula.

Step 5: Compute $x_0 = \ln a_x - \frac{1}{b_x} \ln\left(\frac{1}{x_1} - 1\right)$ and $y_0 = \ln a_y - \frac{1}{b_y} \ln\left(\frac{1}{y_1} - 1\right)$, which follow the logistic distribution $\mathcal{L}(\ln a_x, \frac{1}{b_x})$ and $\mathcal{L}(\ln a_y, \frac{1}{b_y})$ respectively.

Note: we use the fact that if $U \sim \mathcal{U}$, then $X = \mu - \sigma \ln(\frac{1}{U} - 1) \sim \mathcal{L}(\mu, \sigma)$.

Step 6: Find the "coupled", "digitized" log-logistic variables $x = floor\{e^{x_0}\}$ and $y = floor\{e^{y_0}\}$, and truncate them by keeping only the (x, y) pairs with $x > 0$ and $y > 0$.

Similar digitization process is traditionally practiced when one wants to shift from continuous to discrete power law distributions [3].