Bivariate Heavy Tailed Distribution Generator Using Poisson Processes: Models, Analysis and Real World Data

Shan Lu, Gennady Samorodnitsky, Weibo Gong, Bo Jiang, Jieqi Kang, Don Towsley*

Abstract

In this paper, we propose two types of 2D PCSDE model to study the generative mechanism of correlated power law distributions in real datasets. The model of type I generates unstable fractional asymptotic dependence coefficient. This model well explained the data observed in some social networks. The asymptotic dependence coefficient of model of type II can be arbitrarily parameterized between 0 to 1. This model is meaningful in providing suggestions to new generative models for complex networks.

1 Introduction

Power law distributions has been observed in variety of natural and man-made phenomenons [1]. Natural phenomenons reach a variety of fields like sizes of earthquakes, firing pattern in neural networks, etc.; other phenomenons better to be included in man-made categories are corpus of natural language, income ranks, degree distribution of complex networks, etc. Motivated by the mystery of power law distribution, researchers are looking for explanations that brings power law to different areas. In [2], Schwab, *et al.* proposed a general model without fine-tuning to produce Zipf's law in neuron systems. Barabási-Albert model (B-A model) is a scale-free network generative mechanism based on preferential attachment [3]. Researchers expend B-A model to directed network generative model to explain the existence of correlated in-degree and out-degree distribution in World-Wide-Web [4, 5, 6]. In addition to the above models targeting specific area existing power law,

*To be submitted September 2015. Acknowledgement: This work is partly supported by Army Research Office grant W911NF1210385.

some work tries to trace all the phenomenons back to a common cause, like the work in [7, 8, 9, 10]. Reed and Huges' basic idea in [8] is that power law exists when an exponentially growing process is stopped at at an exponentially distributed time. In their following paper in 2003 [9] , they demonstrated that this idea can be applied to explain power law behaviors in biology and growing networks.

In [11], Bo *et al.* interpret the basic idea in [8] in a different way by using Stochastic Differential Equations driven by Poisson counter (PCSDE). The steady state density of the PCSDE model shows lower tail or upper tail power law behavior. A simple model produces an upper tail power law distribution is as following, $dX_t = \beta X_t dt + (\epsilon - X_{t-})dN_t$, with $\beta, \epsilon > 0$, and N is a Poisson process with intensity λ. The steady state distribution follows $f_X(x) = Cx^{-(1+\frac{\lambda}{\beta})}, x \geq \epsilon$.

We show that this model can be used to describe the expected degree growth in growing network. In the generative algorithms, normally the network is growing by adding new nodes and new degrees in each step and when adding new degrees, the nodes are selected with preferential attachment. Assuming that the expected number of nodes added in each step is EN and the expectation of the total degree added in each step is EM. The total degrees added split into the part associated with the new nodes EM_1 and the part added to the nodes in the original graph with preferential attachment EM_2. Assuming that the expected number of nodes in the network grows exponentially with rate λ, the life time of the nodes follow exponential distribution with rate λ. We prove that the expected degree of a node in the network grows exponentially with rate $\beta = \frac{EM_2}{EM}\lambda$. The PCSDE model provide a directed explanation to the power law behavior in network generative models with preferential attachment. The exponent of B-A model in [3] can be estimated using PCSDE model to be 3 under the fact that $EM_2 = \frac{1}{2}EM$.

1D PCSDE model can be used to explain power law behaviors in undirected network. Actually, independent or correlated two-dimensional power law distribution has been widely observed in empirical data, like citation network, social network [12]. Traditional multivariate Pareto Distribution [13] has the same exponents for all univariate margins, which is not suitable to fit real world data with different marginal exponents. In [13], Asimit *et al.* designed a new type of multivariate Pareto distribution with arbitrarily parameterized margins. We developed a basic 2D SDE formulation to generate similar multivariate power law distribution. However, this model generates asymptotic independence distribution regardless of the parameters.

In this paper, we are seeking for modifications to the basic 2D PCSDE model to generate asymptotic dependence. The new models has the following

meanings: (1) help understanding the causes of correlated power law behaviors in empirical data; (2) predict network involvement with finite observed data; (3) it is quite natural to develop new network generative models in the spirit of the new features in the modified models.

The paper is organized as follows. In Section 2, we present some 2D power law distributions in empirical data and review the basic 2D SDE model; Section 3 is our first attempt to generate asymptotic dependence by introducing in Markov on-off process, the modification of type I. Since the model proposed can not produce asymptotic dependence in an elegant way, we proposed another model in Section 4. In section 4, the modification of type II is discussed. The last section concludes the paper.

2 2D Power Law Data and 2D PCSDE Model

2D power law distribution exists in some directed social networks, like friendship connections between users in Youtube and Flickr website [12]. As shown in Figure 1, the two variables in-degree and out-degree in the two datasets are obviously correlated. The empirical correlation coefficients of the two datasets are 0.9492 and 0.7558. In this paper, we attempts to explain the existence of 2D correlated power law distributions in real world data. Our first try is the following 2D PCSDE model with a shared Poisson Counter.

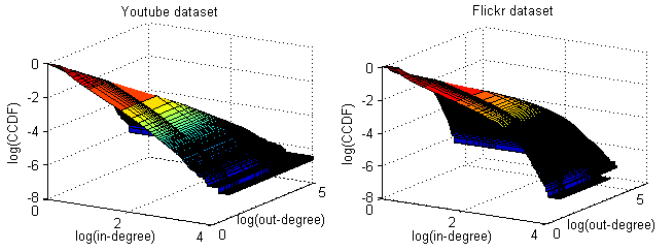


Figure 1: 2D power law data in social networks

2D Model with a Shared Poisson Counter

The following 2D model is a PCSDE formulation of the bivariate Pareto distribution of the second kind in [13],

$$dX_i = X_i dt + (1 - X_i)(dN_0 + dN_i) \tag{1}$$

where N_0, N_1, and N_2 are independent Poisson counters with rate λ_0, λ_1, and λ_2, respectively. The marginal distribution of this model is $f_{X_i}(x_i) =$

$(\lambda_0+\lambda_i)x_i^{-(\lambda_0+\lambda_i+1)}$, $x_i \geq \epsilon_i$; and the CCDF is $\bar{F}_{X_i}(x_i) = x_i^{-(\lambda_0+\lambda_i)}$, $x_i \geq \epsilon_i$. The joint CCDF of this model is $\bar{F}_{X_1,X_2}(x,x) = x^{-\lambda_+}$, where $\lambda_+ = \lambda_0+\lambda_1+\lambda_2$.

The asymptotic behavior of this model can be studied by computing the 'asymptotic dependence coefficient', which is defined as $\lim_{x\to\infty} P(X_2 > x|X_1 > x)$. In this model, we have

$$P(X_2 > x|X_1 > x) = \frac{x^{-\lambda_+}}{x^{-(\lambda_0+\lambda_1)}} = x^{-\lambda_2} \xrightarrow{x\to\infty} 0. \tag{2}$$

This model is useful in generating correlated 2D power law data; however, the two variables in this model are asymptotic independent. In the next section, we will pursue modulations to the original sharing Poisson counter model to produce nonzero asymptotic dependence coefficient.

3 Modulated 2D PCSDE Model of Type I

The basic model is asymptotic independent due to the existence of two independent Poisson Counters. In the first modified model, we consider shutting down the two independent Poisson Counters during some period of the growing processes.

3.1 2D Models with Markov On-off Modulation

Define a Markov on-off process Y_t and our modified 2D PCSDE model is as follows:

$$\begin{aligned} dY &= (1-Y)dM_1 - dM_2, \\ dX_i &= X_i dt + (1-X_i)((1-Y)dN_0 + Y dN_i), \end{aligned} \tag{3}$$

where M_i has rate μ_i, N_i has rate λ_i. In this model, the shared Poisson counter N_0 is effective in "off" period, and the two independent Poisson counters is effective in "on" period.

Let

$$\xi_\pm^{(1)} = \frac{\lambda_0+\lambda_1+\mu_1+\mu_2 \pm \sqrt{(\lambda_1-\lambda_0+\mu_2-\mu_1)^2+4\mu_1\mu_2}}{2} > 0,$$

$$\xi_\pm = \frac{\lambda_0+\lambda_1+\lambda_2+\mu_1+\mu_2 \pm \sqrt{(\lambda_1+\lambda_2-\lambda_0+\mu_2-\mu_1)^2+4\mu_1\mu_2}}{2} > 0.$$

It is easy to check that $\xi_- - \xi_-^{(1)} > 0$, which makes

$$P(X_2 > x|X_1 > x) = \Theta(x^{-(\xi_- - \xi_-^{(1)})}) \xrightarrow{x\to\infty} 0.$$

This model is still asymptotic independent. However, consider a special case when the arrival rates of Poisson counters N_0, N_1 and N_2 are the same, which gives $\lambda_1 = \lambda_2 = \lambda_0 \triangleq \lambda$. If $\mu_1, \mu_2 \ll \lambda$, we have $P(X_2 > x | X_1 > x) \sim \frac{\mu_2}{\mu_1+\mu_2} x^{-\xi_- - \xi_-}$ and $\xi_- - \xi_- < \mu_1$. If μ_1 is small enough, the model can produce non-zero dependence coefficient over a few decades under this special condition (as shown in Figure 2).

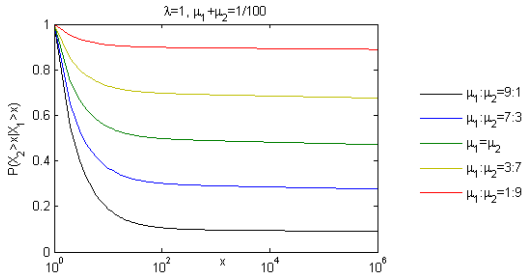


Figure 2: Theoretical dependence coefficient for the modulated model with Markov on-off process in a special case

3.2 Modulated Model with 'Manually Resetting'

The above model separates the effective time of the independent and shared Poisson Counters. However, when we are talking about tail behavior, which corresponding to a very long growing process, the probability for the independent Poisson counters never being effective goes to 0. So we manually revert the growing process to initial value whenever the Markov on-off process changes its state. The new model is as follows:

$$dX_i(t) = X_i dt + (1 - X_i)((1 - Y)(dN_0 + dM_1) + Y(dN_i + dM_2)) \tag{4}$$

Let $\lambda_1 = \lambda_2 \triangleq \lambda$, the marginal and joint CCDFs are as follows:

$$\bar{F}_{X_i}(x) = x^{-(\lambda+\mu_2)} \frac{\mu_1}{\mu_1 + \mu_2} + x^{-(\lambda_0+\mu_1)} \frac{\mu_2}{\mu_1 + \mu_2}, \tag{5}$$

and

$$\bar{F}_X(x, x) = x^{-(2\lambda+\mu_2)} \frac{\mu_1}{\mu_1 + \mu_2} + x^{-(\lambda_0+\mu_1)} \frac{\mu_2}{\mu_1 + \mu_2}. \tag{6}$$

Then let $\Delta\mu = \mu_1 - \mu_2$, we have the asymptotic dependence coefficient of this model:

$$P(X_2 > x | X_1 > x) \xrightarrow{x\to\infty} \begin{cases} 1 & \lambda > \lambda_0 + \Delta\mu \\ \frac{\mu_2}{\mu_1+\mu_2} & \lambda = \lambda_0 + \Delta\mu \\ 0 & \lambda < \lambda_0 + \Delta\mu. \end{cases} \tag{7}$$

This model successfully generates non-zero asymptotic dependence coefficient under special conditions. It seems that the conditional probability depends on the parameters in a discontinuous way. When a parameter is perturbed, the conditional probability goes to 1 or 0 as the value x goes to infinity.

We use this model to fit the joint CCDF of Youtube and Flickr datasets mentioned in Section 2. The "rmse (root mean square error)" for "Youtube" data fitting is $8.2335e-006$ and for "Flickr data" is $2.4703e-005$. We compare the CCDFs of the original datasets and our model fitting results with heat map, as shown in Figure 3.2. Using the fitting parameters we suggest the asymptotic dependence coefficients to the two datasets. We suggest in-degree and out-degree in "Youtube" dataset go towards asymptotic dependence; while the in-degree and out-degree in "Flickr" datasets are asymptotically independent.

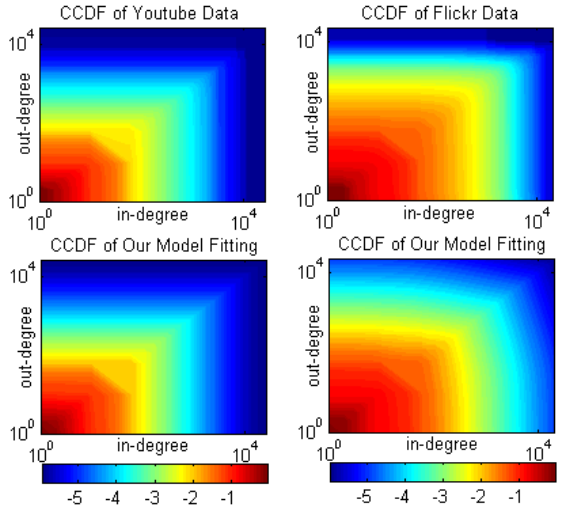


Figure 3: 2D power law data fitting results

4 2D Modified Model of Type II

Modulated model with Markov on-off process and manually resetting in Section 3.2 successfully generate nonzero asymptotic dependence coefficient. However, the case when fractional asymptotic dependence coefficient appears is unstable. We are interested in inventing a more simple and natural model, which could generate fractional asymptotic dependence coefficients in an elegant way.

In this section, we proposed a new model where the exponential growth process of the two variables X_1 and X_2 are not independent any more. The model is as follows:

$$\begin{aligned} dX_1 &= (X_1 + \beta X_2)dt + (1 - X_1)dN_1; \\ dX_2 &= (\beta X_1 + X_2)dt + (1 - X_2)dN_2. \end{aligned} \tag{8}$$

We omit the computation here and only give the results. Let the Poisson rates $\lambda_1 = \lambda_2 \triangleq \lambda$. For the marginal tail, we prove that

$$P(X > x) \sim Cx^{-\alpha},\ x \to \infty. \tag{9}$$

The marginal tail exponent α can be computed by solving the equation $EA^\alpha = 1$, and

$$EA^\alpha = \frac{1}{2}I_1 + I_2^2\frac{1}{4 - 2I_1}, \tag{10}$$

where $I_1 = \frac{\lambda 2^{-\alpha}}{\beta}\int_0^1 z^{\frac{2\lambda-\alpha(1+\beta)}{2\beta}-1}(1+z)^\alpha dz$ and $I_2 = \frac{\lambda 2^{-\alpha}}{\beta}B\left(\frac{2\lambda-\alpha(1+\beta)}{2\beta}, \alpha+1\right)$. Since an analytic solution to I_1 is not available, we use MATLAB to compute the integration numerically and solve α numerically with different β values.

The conditional probability can be computed with Breiman's Theorem [14]. Let $T \sim \exp(2\lambda)$ and given $T = t$, $u \sim U(0, t)$, we get the asymptotic dependence coefficient for this model,

$$\mathbb{P}(X_2 > x | X_1 > x) \xrightarrow{x\to\infty} \frac{2EV^\alpha}{EV^\alpha + EW^\alpha}, \tag{11}$$

where

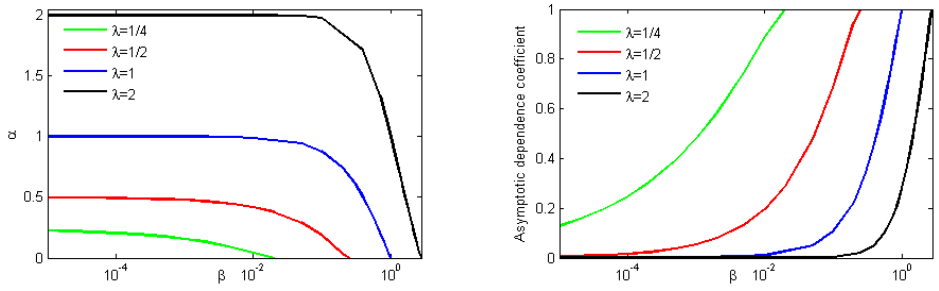
$$V = \frac{e^{u(1+\beta)} - e^{u(1-\beta)}}{2}, W = \frac{e^{u(1+\beta)} + e^{u(1-\beta)}}{2}. \tag{12}$$

EV^α and EW^α can be computed numerically using MATLAB. Since $0 < V < W$ with $\beta > 0$, the asymptotic dependence coefficient of this model is between 0 and 1.

Let $\lambda = 1/4$, $\lambda = 1/2$, $\lambda = 1$ and $\lambda = 2$, the results are plotted in Figure 4. As shown in the Figure, with β increasing, α decreases until it reaches 0, which means that the tail becomes heavier. On the other hand, the asymptotic dependence coefficient increases with the increasing of β and it reaches 1 when the marginal tail exponent α reaches 0.

5 Discussions

In this paper, we develop two types of 2D PCSDE models to generate bivariate power law distribution with tail dependence. The model of Type I



(a) α value as a function of β

(b) Asymptotic dependence coefficient as a function of β

Figure 4: Numerical results of α and asymptotic dependence coefficient as a function of β with different λ values($\lambda = 1/4, 1/2, 1$, and 2))

makes modulation on the Poisson Counter part. We aware that the problem of the basic 2D shared Poisson counter model in Section 2 is the existence of independent Poisson counters. Shared Poisson counter means the two variables start their growth at the same time, which means a common cause lying behind. In social networks, the reason could be that the user is very active. When the empirical data indicates dependence in the regions of very large values we would be able to infer that there is a common cause for the vary large values observed. On the other hand, sometimes an ID in a social network may have a very high in-degree but quite ordinary out-degree, which contributes to the independence at the tail part. Obviously, there is no common cause in this case. For example, this ID may have a very high in-degree because of he/she being famous; however being famous does not ensure high out-degree.

Our modulated model in Section 3 could be seen as a mixture of the two cases. Our study suggests that the conditional probability of one variable given the other tends to be either 0 or 1 asymptotically, and the case where such conditional probability is fractional is not stable. This phenomenon could be explained by the fact that a user only belongs one of the two groups, which is realized by the "Manually resetting" procedure in the model in Section 3.2. As the value increasing, the users in one group may become dominant.

To sum up, although our model of Type I is not that concise and could not produce stable fractional asymptotic dependence coefficient, it does connect to the evolvement of real social networks. We have the theoretical joint distribution of this model, so we can fit real world data and predict the asymptotic behavior based on the data observed.

The model of Type II keeps the two independent Poisson counters but makes modulation on the differential equation part. This model is more complicated to analyze. In this paper we get some conclusions about the tail behavior. This model successfully generate fractional asymptotic dependence coefficient. The coefficient can be adjusted by the parameters in the model.

Since we could not get the precise joint distribution of this model, we did not use this model to fit any empirical data. The principle of this model is to make the increment of one variable also depend on the current value of another one. This inspire us to invent new complex network generative models. For example, a new node with an outgoing link selects the target node with a probability proportional to not only the node's current in-degree, but also the node's current out-degree.

Our on-going and further work includes: (1) exploring datasets which can be explained by our model of type II potentially; (2) generalizing our 2D models to multivariate models.

References

[1] A. Clauset, C.R. Shalizi and M.E.J. Newman, *Power-Law Distributions in Empirical Data*, SIAM REVIEW, 51(4), 661-703, 2009.

[2] D.J. Schwab, I. Nemenman and P. Mehta, *Zipfs Law and Criticality in Multivariate Data without Fine-Tuning*, Physical Review Letters, PRL 113, 068102, 2014.

[3] A. Barabási and R. Albert, *Emergence of Scaling in Random Networks*, Science, 286, 509-512, 1999.

[4] P.L. Krapivsky, G.J. Rodgers and S. Redner, *Degree Distributions of Growing Networks*, Physical Review Letters 86(23),5401-5404, 2001.

[5] P.L. Krapivsky and S. Redner, *A statistical physics perspective on Web growth*, Computer Networks, 39, 261-276, 2002.

[6] B. Bollobás, C. Borgs, J. Chayes and O. Riordan *Directed Scale-free Graphs*, Proceedings of SODA'03, 132-139, 2003.

[7] W.J. Reed, *The Pareto, Zipf and other power laws*, Economics Letters, 74, 15-19, 2001.

[8] W.J. Reed and B.D. Huges, *From gene families and genera to incomes and internet file sizes: Why power laws are so common in nature*, Physical Review E 66,067103,2002.

[9] W.J. Reed and B.D. Huges, *Power-law distributions from exponential processes: an explanation for the occurrence of long-tailed distributions in biology and elsewhere*, Scientiae Mathematicae Japonicae, 58(2), 473-483, 2003.

[10] M. Mitzenmacher, *A Brief History of Generative Models for Power Law and Lognormal Distributions*, Internet Mathematics, 1(2), 226-251, 2004.

[11] B. Jiang, R.W. Brockett, W. Gong and D. Towsley *Stochastic Differential Eqations for Power Law Behavior*, 51st IEEE Conference on Decision and Control, 6696-6701, Dec 10-13, 2012.

[12] J. Kunegis, *KONECT: the Koblenz Network Collection*,WWW 2013 Companion, May 13-17, 2013, Rio de Janeiro, Brazil.

[13] A.V. Asimit, E. Furman and R. Vernic, *On a Multivariate Pareto Distribution*, Insurance: Mathematics and Economics, 46(2), 308-316, 2010.

[14] D. Denisov and B. Zwart, *On a theorem of Breiman and a class of random difference equations*, Journal of Applied Probability, 44(4), 1031-1046, 2005.