

Brief Announcement: Revisiting the Power-law Degree Distribution for Social Graph Analysis

Alessandra Sala, Haitao Zheng, and
Ben Y. Zhao
UC Santa Barbata, USA
{alessandra, htzheng, ravenben}
@cs.ucsb.edu

Sabrina Gaito, and Gian Paolo Rossi
Universita degli Studi di Milano, Italy
gaito@dsi.unimi, rossi@dico.unimi.it

ABSTRACT

The study of complex networks led to the belief that the connectivity of network nodes generally follows a Power-law distribution. In this work, we show that modeling large-scale online social networks using a Power-law distribution produces significant fitting errors. We propose the use of a more accurate node degree distribution model based on the Pareto-Lognormal distribution. Using large datasets gathered from Facebook, we show that the Power-law curve produces a significant over-estimation of the number of high degree nodes, leading researchers to erroneous designs for a number of social applications and systems, including shortest-path prediction, community detection, and influence maximization. We provide a formal proof of the error reduction using the Pareto-Lognormal distribution, which we envision will have strong implications on the correctness of social systems and applications.

Categories and Subject Descriptors

I.6.4 [Simulation and Modeling]: Model Validation and Analysis;
G.3 [Probability and Statistics]: Distribution Functions

General Terms

Analysis, Measurement

Keywords

Social Networks, Power-law, Pareto-Lognormal

1. INTRODUCTION

With billions of users, online social networks (OSNs) and social applications have changed the way users interact with the Internet. In recent years, the research community has seen a rise in large-scale measurements of deployed social networks [?, ?, ?]. Analysis of the resulting online social graphs is critical to our understanding of the mechanisms underlying the formation and evolution of complex networks, and serves a particularly important role in guiding the design of social network applications.

Scientists have long accepted the Power-law model for complex networks, and have applied it to online and offline social networks. Current measurement studies on online social networks [?, ?, ?] use the Power-law coefficient as a standard graph metric, ignoring the fitting errors and the implications of those errors.

Our work shows that the Power-law model largely over-predicts the number of “high degree nodes” and this over-prediction has

a significant impact on a set of algorithms and protocols whose designs rely on estimates of high degree nodes generated by the Power-law model. Because they leverage the population of high degree nodes, analysis and simulations of these algorithms must be re-evaluated to remove the error introduced by the Power-law model. Examples include: distributed resource replication strategies to minimize routing delay and perform load balancing [?], epidemic dissemination strategies to maximize information spread [?], landmark selection strategies to accurately predict shortest paths in graphs [?, ?], and community detection to improve social recommendation systems.

Our work, in its full version [?], has two contributions. First, we propose an alternative distribution to the Power-law to better capture and predict node degree distributions in online social networks. We provide formal proof of how much the modified distribution improves upon the poor fit of the Power-law model. Second, we evaluate the real implications of using our more accurate model for algorithmic problems, and quantify its impact on several important network measures.

2. MIXTURE OF PROBABILITIES BEHIND THE DEGREE DISTRIBUTION OF OSN

Exploring the degree distribution is the first step towards a deep understanding of online social networks. As mentioned above, we recognize that the degree distribution in these networks do not match the presumed Power-law distribution. By investigating the family of distributions that fuses Pareto with Lognormal distributions we identify a new model that does not have the Power-Law limitations.

The model we propose as an alternative to Power-law is a *limit form* of the Double Pareto-Lognormal (DPLN) distribution [?]. The DPLN is a complex model with four parameters: two Pareto exponents α and β which identify the slope of the upper and lower tails of the distribution and μ and σ which describe the Lognormal parameters which connect the two Pareto tails. The DPLN originates other two distributions: its left side Pareto-Lognormal (PLN) and its right side Lognormal-Pareto (LNP). Both these distributions are expressed by three parameters: β is the Pareto exponent, and μ and σ characterize the Lognormal component.

We propose, here, to use the PLN distribution, whose cumulative distribution function is:

$$F(X) = \Phi\left(\frac{\log x - \mu}{\sigma}\right) + x^\beta e^{(-\beta\mu + \frac{\beta^2\sigma^2}{2})} \Phi^c\left(\frac{\log x - \mu + \beta\sigma^2}{\sigma}\right)$$

with $E[X] = \mu - \frac{1}{\beta}$ and $VAR[X] = \sigma^2 + \frac{1}{\beta^2}$; the Pareto component describes the low values and the Lognormal one dominates the high values.

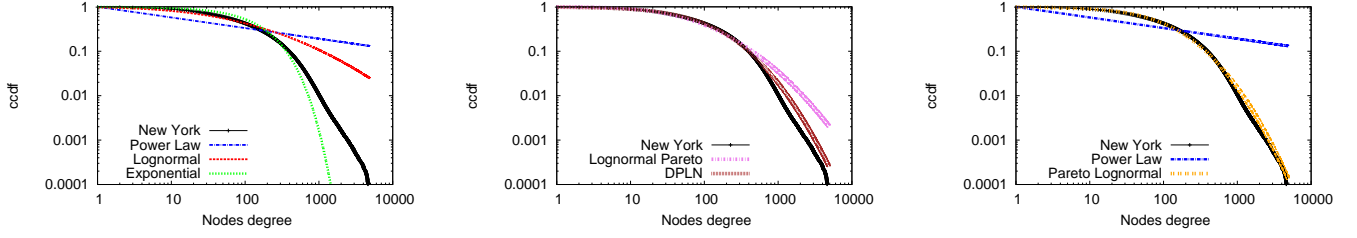


Figure 1: Fitting the degree distribution of New York Facebook network; first, with Elementary distributions; second with LNP and DPLN, and third with Power-law and PLN.

In the following section, we provide statistical evidences of the improved fitting obtained with the Pareto-Lognormal distribution model compared to the Power-law distribution and other distributions. We leverage some of the popular statistical methods to evaluate the *Goodness-of-Fit Tests* of the abovementioned distributions. We compute, for each distribution, the minimum log-likelihood, the residual sum of squares (RSS) in order to estimate the discrepancy between the data and the estimate model, and finally, the Akaike Information Criterion (AIC) that includes a penalty that is an increasing function of the number of the estimated parameters.

Experimental Results Among the social graphs that we have analyzed (*i.e.* more than 11 Facebook social graphs from [?]), we present, here, a Facebook regional network, New York which has 855K nodes and 66 million edges. In Figure ?? we report the fitted distributions on the New York network. In the first plot, we show the complementary cumulative distribution function for the Power-law, Lognormal and Exponential models, while the second plot shows the DPLN and its right limit form (*i.e.*, LNP). Finally, the third plot of Figure ?? shows how PLN distribution more accurately models the real data compared to the Power-law.

We compute the likelihood of each distribution model considered together with the RSS and AIC statistical methods to formally compare goodness of the fit of those distributions. We found, as reported in our full version [?], that the Pareto-Lognormal model outperforms the others almost on any statistical and graphical test.

Theoretical Results In this section, we aim to reveal that fundamental flaws in evaluating network measures if a poor model is used to describe those networks.

We intend to quantify the γ % of high degree nodes in the network, in particular we study the minimum degree ξ_γ , so that all nodes with degree higher than ξ_γ are the ones that account for the γ % of the high degree nodes. Therefore, ξ_γ is the γ -th quantile of the complementary cumulative distribution.

Assuming a Power-law model, the degree ξ_γ is $(\frac{1}{\gamma})^{\frac{1}{\alpha}}$. In order to quantify the same quantile for the PLN distribution we first studied its limit behavior, which appears to be a Lognormal distribution whose parameters are estimated from the Pareto-Lognormal model. We use the asymptotic expansion of the complementary Gauss error function to get the tight upper bound as reported in Lemma ??.

LEMMA 1. *A tight upper bound of the γ -th quantile of the PLN distribution is $\xi_\gamma = e^{\mu + \sqrt{-\log(\frac{\sqrt{2\pi}\gamma}{\sigma})}2\sigma^2}$.*

In order to compute the difference in the order of magnitude between the Power-law quantile ξ_γ and the PLN one we approximate the PLN ξ_γ as e^μ , since $\sqrt{-\log(\frac{\sqrt{2\pi}\gamma}{\sigma})}2\sigma^2$ is negligible compared to μ . Since μ and $\frac{1}{\alpha}$ are approximated by the mean value, ν , of the sample logarithms the following theorem applies:

THEOREM 1. *The Power-law overestimation, computed as the ratio of the ξ_γ values, is $\approx ((\frac{1}{\sigma\gamma})^\nu)$.*

The ξ_γ degree is expressed as an exponential function respectively by the PLN and Power-law, although with *different base*. This difference in the New York network, for example, produces, on the 10% of the highest degree nodes, a minimum super-node degree estimation of ≈ 1099 with the PLN and of ≈ 15000 with the Power-law. The same quantity computed on the sample is 395, so Power-law reveals an intolerable error compare to the PLN one.

The results derived by the quantile estimation allow us to bound few other related networks quantities which confirm the large overestimation derived by the erroneous use of Power-law distributions. In particular, we bounded [?] the number of incident edges to high degree nodes and the total number of high degree nodes, defined as the nodes with degree higher than a given minimum value. For example, in New York network, fixed the minimum degree of a super-node equal to 500, the Power-law model predicts 193840 super-nodes while the PLN estimates 53394 ones, which is a “tight” approximation of the 55103 from the real data.

3. CONCLUSION

Current solutions to challenging problems on complex networks often rely on assumptions about properties of the network. In particular we refer to those algorithms and protocols which rely on the population of high degree nodes or their connectivity. Today, the overestimation of these network properties is so large, that we need to revisit application-level results of numerous algorithms that relied on an erroneous Power-law assumption.

4. REFERENCES

- [1] AHN, Y.-Y., ET AL. Analysis of topological characteristics of huge online social networking services. In *Proc of WWW* (2007).
- [2] CHEN, W., SOMMER, C., TENG, S.-H., AND WAN, Y. Compact routing in power-law graphs. Tech. rep., 2009.
- [3] CHEN, W., WANG, Y., AND YANG, S. Efficient influence maximization in social networks. In *Proc. of KDD* (2009).
- [4] KUMAR, R., NOVAK, J., AND TOMKINS, A. Structure and evolution of online social networks. In *Proc. of KDD* (2006), pp. 611–617.
- [5] MISLOVE, A., KOPPULA, H. S., GUMMADI, K. P., DRUSCHEL, P., AND BHATTACHARJEE, B. Growth of the flickr social network. In *Proc. of WOSN* (Seattle, WA, August 2008).
- [6] MISLOVE, A., MARCON, M., GUMMADI, K. P., DRUSCHEL, P., AND BHATTACHARJEE, B. Measurement and analysis of online social networks. In *Proc. of Internet Measurement Conference* (Oct 2007).
- [7] POTAMIAS, M., BONCHI, F., CASTILLO, C., AND GIONIS, A. Fast shortest path distance estimation in large networks. In *Proc. of CIKM* (2009).
- [8] PUTTASWAMY, K. P. N., SALA, A., AND ZHAO, B. Y. Searching for rare objects using index replication. In *Proc. of INFOCOM* (Phoenix, AZ, April 2008).
- [9] REED, W. J., AND JORGENSEN, M. The double pareto-lognormal distribution - a new parametric model for size distributions, 2003.
- [10] SALA, A., GAITO, S., ROSSI, G. P., ZHENG, H., AND ZHAO, B. Manuscript (2010).
- [11] WILSON, C., ET AL. User interactions in social networks and their implications. In *Proc. of EuroSys* (April 2009).