IJCSS International Journal of Complex Systems in Science

Int. J. Complex Systems in Science
vol. **1(2)** (2011), pp. 129–133

Fitting the Degree Distribution of Real-World Networks

Faustino Prieto[1,†] and José María Sarabia[1]

[1] *Department of Economics, University of Cantabria*

Abstract. In this paper, the degree distribution of real-world networks is studied. Many authors have found out that those networks appear to follow a power law behaviour, however other functional forms, such us Lognormal, Gamma or Weibull have been proposed for that. Here, six different probabilistic distributions are fitted by maximum likelihood to the entire range: two particular cases of Pareto Positive Stable (PPS), Lognormal, Gamma, Weibull and Generalized Pareto distributions. These models are fitted to several real-world network datasets in a socioeconomics context. The different models are compared using the Akaike information criterion (AIC), the Bayesian Information Criterion (BIC) and the Kolmogorov-Smirnov (KS) statistic. Finally, the model validation is done by using log-log rank-size plots.

Keywords: Power law, PPS distribution, Degree distribution.
MSC 2000: 60E05, 60H35, 90B15, 91B44, 93A30.

† **Corresponding author:** faustino.prieto@unican.es
Received: October 14, 2011
Published: October 24, 2011

1. Introduction

The study of the degree distribution of paper citation network of scientific publications is an important issue in networks science in general, and in information and informetrics science in particular [1]. In this field, an opened question is which distribution is the most adequate to describe the degree distribution in the entire domain. Among all known probability distributions, the Pareto distribution (power law) is the most popular due to its simplicity and properties. However, in many cases, Pareto distribution is only valid in the upper tail. For that reason, other functional forms have been proposed. In this paper, we look for the best descriptive model. We restrict our research to models with two parameters. We fit and compare six models: two particular cases of the Pareto Positive Stable (PPS), the Generalized Pareto, the Weibull, the Gamma and the Log-Normal distributions.

2. Data and Methodology

Our dataset is composed of information about the total number of cites that papers, published in high-impact journals and international proceedings (2000-2011), have received from other papers. We have included papers from the following Web of Science categories: Computer Science Information Systems; Computer Science Interdisciplinary Applications; Economics; Education Educational Research; Nanoscience Nanotechnology; Telecommunications. We have tested the adequacy of the power law to the observed data by employing two methods: a graphical method based on the rank-size plot and an analitical method based on the Hill estimator [2] and the Kolmogorov-Smirnov statistic [3]. Then, we have compared the six models enumerated before by using the Akaike information criterion AIC [4] and we have validated graphically the model chosen by using log-log rank-size plots.

3. Results

Figure 1 and Table 1 show that the observed data do not follow a power law behaviour in the entire domain and even, there is no a power law in the upper tail of the Economics category. As alternative models, we have considered two particular cases of the PPS [5] distribution, PPSa and PPSb, defined in terms of cumulative distribution function (cdf) as follows:

$F(x;\lambda,\nu) = \Pr(X \leq x) = 1 - \exp\left\{-\lambda[\log(x+1)]^{\nu}\right\},\ x \geq 0, \quad (PPSa)$

$F(x;\lambda,\nu) = \Pr(X \leq x) = 1 - \exp\left\{-\lambda[1+\log(x)]^{\nu}\right\},\ x \geq 0, \quad (PPSb)$

Table 2 shows that PPSb model presents the lowest value of the AIC statistic. Figure 2 shows that PPSb gives a reasonably description of the observed data.

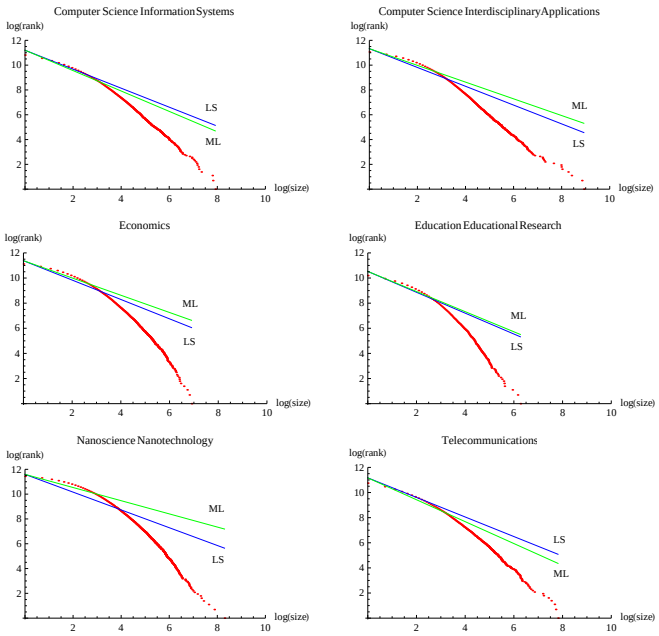

Figure 1: Complementary of the cdf of Pareto distribution (solid lines), estimated by least-square linear regression (LS) and maximum likelihood (ML), and the observed data, on log-log scale.

Table 1: *pvalue* test [6] of the power law with the dataset considered.

Category	x_{min}	α	p_{value}	Papers (upper tail)	Papers (cites> 0)
Computer Science Inform. Syst.	64	1.690	0.878	1206	73352
Computer Science Interd. App.	32	1.623	0.142	7503	83195
Economics	36	1.758	0.063	4076	87675
Education Educat. Research	27	2.043	0.544	1477	37311
Nanoscience Nanotechnology	146	2.126	0.381	1071	108608
Telecommunications	25	1.455	0.147	4119	70858

Table 2: *AIC* values obtained from the fitting by maximum likelihood.

Category	PPSb	PPSa	Log-Normal	Weibull	Gamma	Generalized Pareto
Computer Sc. Inf. Syst.	397141	411425	410458	439036	451320	421081
Computer Sc. Int. App.	503517	513637	512702	538862	550917	522801
Economics	522523	533320	532098	557662	565635	543515
Education Ed.. Research	200382	205878	204784	215520	217228	211067
Nanoscience Nanotechn.	761282	768327	767793	792638	803156	777115
Telecommunications	370050	385063	383894	413244	425415	394898

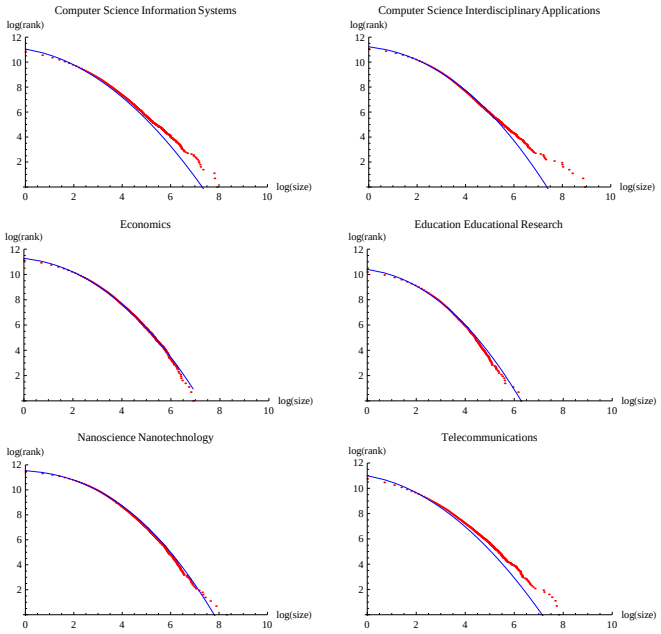

Figure 2: Complementary of the cdf of the PPSb distribution (solid lines) and the observed data (six Web of Science categories, 2000-2011) on log-log scale.

4. Conclusions

PPSb distribution is the best descriptive model for the degree distribution of paper citation network, in comparison with other alternative models proposed: Pareto, Generalized Pareto, Weibull, Gamma and LogNormal distributions.

Acknowledgements

The authors thank to Ministerio de Ciencia e Innovación (project ECO2010-15455) for partial support of this work.

References

[1] S. Onel, A. Zeid and S. Kamarthi, *Scientometrics* **89**, 119-138 (2011).

[2] B.M. Hill, *Annals of Statistics* **3**, 1163-1174 (1975).

[3] W.H. Press, S.A. Teukolsky, W.T. Vetterling and B.P. Flannery, *Cambridge University Press* **2nd ed.**, Cambridge, UK (1992).

[4] H. Akaike, *IEEE Transactions on Automatic Control* **19**, 716-723 (1974).

[5] J.M. Sarabia and F. Prieto, *Physica A: Statistical Mechanics and its Applications* **388(19)**, 4179-4191 (2009).

[6] A. Clauset, C.R. Shalizi and M.E.J. Newman, *SIAM review* **51(4)**, 661-703 (2009).