Features of complex networks in a free-software operating system

2012 J. Phys.: Conf. Ser. 365 012058

(http://iopscience.iop.org/1742-6596/365/1/012058)

Journal of Physics: Conference Series **365** (2012) 012058 doi:10.1088/1742-6596/365/1/012058

Features of complex networks in a free-software operating system

Rajiv Nair[1], **G Nagarjuna**[2], **Arnab K Ray**[3]

[1] Tata Institute of Social Sciences, V N Purav Marg, Deonar, Mumbai 400088, India
[2] Homi Bhabha Centre for Science Education, Tata Institute of Fundamental Research, V N Purav Marg, Mankhurd, Mumbai 400088, India
[3] Department of Physics, Jaypee University of Engineering and Technology, A B Road, Raghogarh, Guna 473226, Madhya Pradesh, India

E-mail: rajiv@tiss.edu, nagarjun@gnowledge.org, arnab.kumar@juet.ac.in

Abstract. We propose a mathematical model to fit the degree distribution of directed dependency networks in free and open-source software. In this complex system, the intermediate scales of both the in-directed and out-directed dependency networks follow a power-law trend (specifically Zipf's law). Deviations from this feature are found both for the highly linked nodes, and the poorly linked nodes. This is due to finite-size effects in the networks, and the parameters needed to model finite-size behaviour make a quantitative distinction between the in-directed and out-directed networks. We also provide a model to describe the dynamic evolution of the network, and account for its saturation in the long-time limit.

1. Introduction

The last decade has seen the emergence of complex networks [1, 2, 3, 4, 5] across diverse domains like (to name a few) the World Wide Web, the Internet, social, ecological and biological systems, income and wealth distributions, trade and business networks, highway networks, linguistic structures with implications for their syntax and semantics, electronic circuits and the architecture of computer software. A recurring theme in many of these systems is a power-law (scale-free) distribution, a property that is significant for the inherent robustness of the network and its evolutionary aspects. And so it is that power-law features have been found [6] in the dependency networks of free and open-source software (*FOSS*) as well, which have further been shown actually to follow Zipf's law [7]. However, it has also been realised that simple power-law properties do not suffice to provide a complete global model for the networks in a *FOSS*. For any system with a finite size, the power-law trend is not manifested indefinitely, and deviations from this trend appear for both the profusely linked and the sparsely linked nodes in the network. Modelling the finite-size effects (equivalently the saturation properties) in the complex network underlying a *FOSS*, and to study how these effects are related to the dynamic properties of the network are the principal objectives of our work [8].

2. Modelling the degree distributions in a *FOSS* network

When it comes to installing a software package from the *Debian GNU/Linux* repository, many other packages — the "dependencies" — are also called for as prerequisites. This leads to a dependency-based network among all the packages, and each of these packages may be treated as a node in a network of dependency relationships. Each dependency relationship connecting any two packages (nodes) is treated

as a link (an edge), and every link establishes a relation between a prior package and a posterior package, whereby the functions defined in the prior package are invoked in the posterior package. As a result, the whole operating system emerges as a functional network, whose fundamental character is determined by the direction of the links.

Networks in a *FOSS* appear in two types: one comprising only the incoming links among the nodes (the in-degree distribution) and the other comprising only the outgoing links among the nodes (the out-degree distribution). To develop any mathematical model that purports to give an accurate global prescription for the degree distribution in both of these networks, it will be necessary first to count the actual number of software packages, ϕ, which are connected by a particular number of links, x, in either kind of network. This gives an unnormalised frequency distribution plot of $\phi \equiv \phi(x)$ versus x. To give continuum description for the power-law feature as well as any finite-size effect in this kind of a frequency distribution, we forward a nonlinear logistic-type equation going as $(x+\lambda)\,\phi'(x) = \alpha\phi\,(1-\eta\phi^{\mu})$, in which $\phi'(x)$ is a first derivative with respect to x, with α being a power-law exponent, μ being a nonlinear saturation exponent, η being a "tuning" parameter for nonlinearity and λ being another parameter that will be instrumental in setting a limiting scale for the poorly connected nodes. The parameter μ crucially controls the advent of scale-free features in the network, while the parameters, η and λ, quantify the finite-size properties in the network on its extremal scales. These two parameters also make a quantitative distinction between the in-directed and out-directed networks.

The integral solution of the foregoing differential equation is obtained as (for $\mu \neq 0$)

$$\phi(x) = \left[\eta + \left(\frac{x+\lambda}{c}\right)^{-\mu\alpha}\right]^{-1/\mu}, \tag{1}$$

in which c is an integration constant (as far as this static model is concerned). It is quite obvious that when $\eta = \lambda = 0$ (with the former condition implying the absence of nonlinearity), there will be a global power-law distribution for the data, going as $\phi(x) = (x/c)^{\alpha}$, regardless of any non-zero value of μ. The situation becomes quite different, however, when both η and λ have non-zero values. In this situation, the network will exhibit a saturation behaviour on extreme scales of x (both low and high). For the high values of x, this can be easily appreciated from the proposed differential equation itself, wherefrom the limiting value of ϕ is obtained as $\phi = \eta^{-1/\mu}$.

The mathematical model implied by Eq. (1) is based on data from two stable *Debian FOSS* releases[1], *Etch* (*Debian GNU/Linux 4.0*) and *Lenny* (*Debian GNU/Linux 5.0*), and it has been tested subsequently on the latest stable release *Squeeze* (*Debian GNU/Linux 6.0*). As mentioned earlier, each release of *Debian FOSS* has two kinds of dependency networks — the in-degree and the out-degree. In fitting the degree distribution of all these networks, the parameter values $\alpha = -2$ and $\mu = -1$, emerge as a universal feature in *all* the cases. While the obvious implication of the former value is that Zipf's law governs the intermediate scales of the dependency networks, the latter value imposes a scale-free character on the network. Carrying out a power series expansion on Eq. (1), it is not difficult to see that a self-contained and natural truncation of the series can only be achieved when $\mu = -1$, i.e. when a single power law operates on all scales of x.

With the values of $\alpha = -2$ and $\mu = -1$, the finite-size properties in the network can be expressed from Eq. (1) as $\phi(x) = \eta + c^2\,(x+\lambda)^{-2}$. The implications of this result are noteworthy. One of these is that the parameter, η, defines a finite lower bound to the discrete count of nodes, regardless of any arbitrarily high value of x, i.e. $\phi \longrightarrow \eta$ as $x \longrightarrow \infty$. The other point worth noting is that when the model is fitted with the data, η makes a quantitative distinction between the in-degree and the out-degree distributions. Going back to the data taken from the *Etch* release, the values of η for the in-degree and the out-degree distributions, respectively, are -8 and 1. For the *Lenny* release, the corresponding values are -15 and 1. The role of η in determining the finite-size properties for high values of x, and in

[1] `http://www.debian.org/releases`

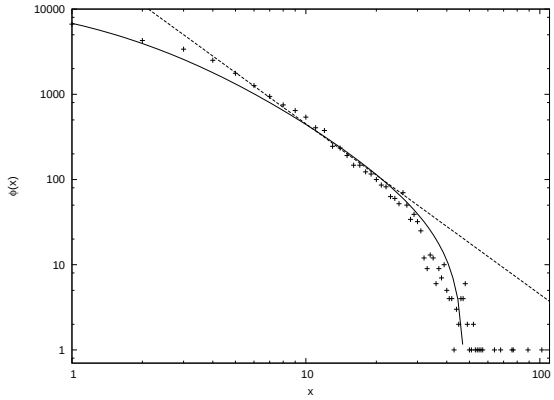

Figure 1. For the network of incoming links in the *Squeeze* release, the degree distribution shows a good fit in the intermediate region with a power-law exponent, $\alpha = -2$ (as indicated by the dotted straight line), which validates Zipf's law. However, for large values of x, there is a saturation behaviour towards a limiting scale that is fitted by the parameter value, $\eta = -28$. On the other hand, when x is small, the fit is good for $\lambda = 2.2$. For this plot, $c \simeq 265$.

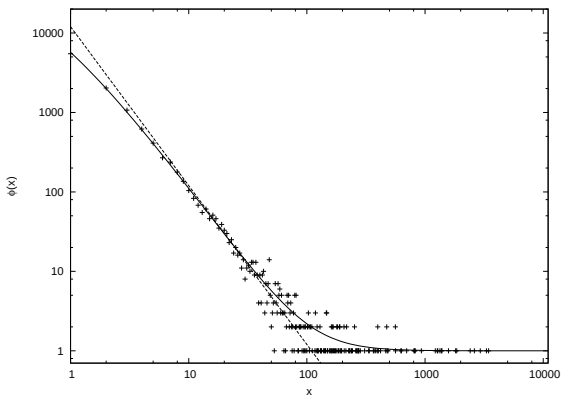

Figure 2. For the network of outgoing links in the *Squeeze* release, the degree distribution of intermediate nodes are again fitted by $\alpha = -2$ (as the dotted straight line shows). However, the saturation behaviour for large values of x is different from that of the network of incoming links. ϕ converges to a limit given by $\eta = 1$. When x is small, the convergence is attained for $\lambda = 0.45$. Thus the sign of η distinguishes the dependency networks. For this plot, $c \simeq 110$.

making a distinction between the two degree distributions is very much in evidence in Figs. 1 & 2, which show the in-degree and the out-degree distributions for *Squeeze*, the latest stable release of *Debian*. It is also interesting to note that for the out-degree distribution, the value of η remains unchanged at 1 with increasing values of x, across all generations of *Debian* releases. Making use of this value, and by requiring the two terms on the right hand side of Eq. (1) to be in rough equipartition with each other, we obtain a scale, $x_{\rm sat} \sim c$, at which the degree distribution starts saturating to a finite limit.

Considering the other extreme of finite-size effects at small values of x, a noticeable deviation from the power-law solution is also seen. This is especially true for the in-degree distribution in Fig. 1. In the limit of small degree distributions for both the in-directed and out-directed networks, where η ceases to have a quantitative significance, and where $x \sim 1$ (which, in the discrete count of links, is the lowest value that x can assume practically), we can find an upper bound to the number of the very sparsely linked nodes, given as $\phi_{\rm ub} \simeq c^2 \, (1+\lambda)^{-2}$, with the full range of ϕ, therefore, going as $\eta \leq \phi \lesssim \phi_{\rm ub}$.

3. Dynamics and long-time saturation in the network

The *FOSS* network is a dynamically evolving network, undergoing continuous additions (even deletions) and modifications across several generations of *Debian* releases. A realistic model should account for this evolutionary aspect of the network distribution, and at the same time accommodate the finite-size features of the network.

Going by the fact that for the out-degree distributions, the value of η remains fixed at unity, one can reason that the richly connected nodes in this case form the irreducible nucleus of the *FOSS* network. These nodes are by far the most influential in the network and it will be useful to study their dynamic properties. From the perspective of a continuum model, the frequency distribution of the nodes in the network of outgoing links can be seen as a field, $\phi(x,t)$, evolving continuously through time, t, with the saturation in the number of nodes for high values of x, emerging of its own accord from the dynamics. In keeping with this need, an ansatz with a general power-law feature inherent in it, may be framed as $\phi(x,t) = c^{-\alpha} \, (x+\lambda)^{\alpha} + \varphi(x,t)$ in which $\varphi \longrightarrow \eta$, as $t \longrightarrow \infty$. Under this requirement, we describe

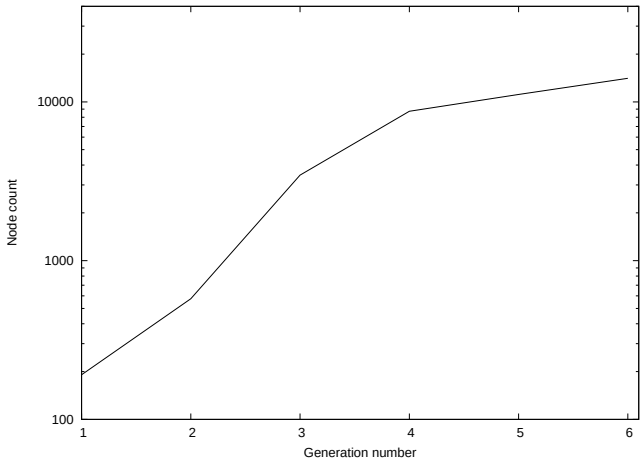

Figure 3. The curve plots the growth of the total number of nodes in the network of outgoing links. Saturation in growth can be seen for the later releases of *Debian*. The generation number plotted along the horizontal axis, effectively marks the time through which the network has evolved. Data for this plot have been taken from all the six releases of *Debian*.

the temporal evolution of the network by a model equation, going as,

$$\tau \frac{\partial \phi}{\partial t} = \frac{\partial \phi}{\partial x} - \frac{\alpha}{c^{\alpha}} \left(x + \lambda\right)^{\alpha - 1} \, , \tag{2}$$

in which τ is a parameter to scale time. The solution of Eq. (2) is obtained by the method of characteristics [9] as $\phi(x,t) = \eta + c^{-2}\left(x + \lambda\right)^{\alpha} - c^{-2}\left[x + \lambda + (t/\tau)\right]^{\alpha}$, and this, under the condition that $\alpha = -2$, will converge to the expected static distribution for $t \longrightarrow \infty$. The long-time saturation of the total number of nodes in the out-degree distribution can then be evaluated as

$$N_{\rm out}(t) = \int_{1}^{x_{\rm m}} \phi(x,t)\, {\rm d}x \, \simeq \, \eta x_{\rm m} + \frac{c^2}{1+\lambda} - c^2 \left(1 + \lambda + \frac{t}{\tau}\right)^{-1} \, , \tag{3}$$

in which $x_{\rm m} (\gg 1)$ is the maximum number of links that a node has in the network. Going by the data taken from all the six releases of *Debian*, the first three of which are *Buzz* (*Debian GNU/Linux 1.1*), *Hamm* (*Debian GNU/Linux 2.0*) and *Woody* (*Debian GNU/Linux 3.0*), we see in Fig. 3 that the total number of nodes in the out-degree distribution does indeed saturate to a finite end in accordance with what one can expect from Eq. (3), when $t \longrightarrow \infty$.

Acknowledgments

The authors thank J. K. Bhattacharjee, C. Gershenson, P. Majumdar, S. Sinha, S. Spaeth, R. Stinchcombe, V. M. Yakovenko and S. Zacchiroli for their helpful comments.

References

[1] Strogatz S H 2001 *Nature* **410** 268
[2] Albert R and Barabási A L 2002 *Rev. Mod. Phys.* **74** 47
[3] Newman M, Barabási A L and Watts D J 2006 *The Structure and Dynamics of Networks* (Princeton and Oxford: Princeton University Press)
[4] Barrat A, Barthélemy M and Vespigniani A 2008 *Dynamical Processes on Complex Networks* (Cambridge: Cambridge University Press)
[5] Newman M E J 2011 Complex Systems: A Survey *Preprint* arXiv:1112.1440
[6] LaBelle N and Wallingford E 2004 Inter-Package Dependency Networks in Open-Source Software *Preprint* cs/0411096
[7] Maillart T, Sornette D, Spaeth S and von Krogh G 2008 *Phys. Rev. Lett.* **101** 218701
[8] Nair R, Nagarjuna G and Ray A K 2009 Semantic Structure and Finite-Size Saturation in Scale-Free Dependency Networks of Free Software *Preprint* arXiv:0901.4904
[9] Debnath L 1997 *Nonlinear Partial Differential Equations for Scientists and Engineers* (Boston: Birkhäuser)