Caracterização da distribuição de carga em redes complexas submetidas a um tráfego uniforme

Elias Bareinboim[1], **Valmir C. Barbosa**[1] *

[1]Programa de Engenharia de Sistemas e Computação, COPPE,
Universidade Federal do Rio de Janeiro,
Caixa Postal 68511, 21941-972 Rio de Janeiro - RJ, Brazil

{eliasb,valmir}@cos.ufrj.br

Abstract. *The load of a node in a network is the total traffic going through it. We express nodal load in terms of the more elementary notion of a node's descents in breadth-first-search trees, and study both the descent and nodal-load distributions in the case of scale-free networks. Our main result is that the load distribution, even though it can be disguised as a power-law through subtle (but inappropriate) binning of the raw data, is in fact a succession of sharply delineated probability peaks, each of which can be clearly interpreted as a function of the underlying BFS descents. This find is in stark contrast with previously held belief, based on which a power law of exponent -2.2 was conjectured to be valid regardless of the exponent of the power-law distribution of node degrees.*

Resumo. *A carga de um nó em uma rede é o tráfego total passando atavés dele. Nós expressamos a carga do nó em termos de uma noção mais elementar de descendência em árvores de busca em largura, e estudamos as distribuições da descendência e da carga do nó no caso de redes scale-free. Nosso resultado principal é que a distribuição da carga, mesmo podendo ser obtida como uma power-law através da aplicação sútil (mas inapropriada) de binning sobre os dados brutos, é na verdade uma sucessão de picos de probabilidades bem delineados, cada um podendo ser interpretado como uma função da descendência BFS subjacente. Tal descoberta encontra-se em oposição direta a crença previamente estabelecida, baseada na qual uma power-law com expoente -2.2 foi conjecturada como valida independente do expoente da distribuição da power-law para os graus dos vértices.*

1. Introduction

In a scale-free network, node connectivities (or degrees) are distributed according to a power law, that is, the probability that a randomly chosen node has degree k is

*Esta dissertação foi orientada pelo professor Valmir C. Barbosa e apresentada em Set/2007 à COPPE/UFRJ para a obtenção do grau de M. Sc.. Participaram da banca examinadora os professores Celina M. H. de Figueiredo (COPPE/UFRJ), Raul Donangelo (IF/UFRJ) e Virgílio A. F. Almeida (DCC/UFMG). Além deste artigo, publicamos uma versão mais detalhada sobre este trabalho na revista Physical Review E [Bareinboim and Barbosa(2008)], relevante periódico da área. Gostaríamos de agradecer ao suporte financeiro parcial da SBC, CNPq, CAPES, e FAPERJ BBP. A versão completa da dissertação encontra-se em meu site (http://www.cos.ufrj.br/∼eliasb/mestrado/).

proportional to $k^{-\tau}$ for some $\tau > 1$. Scale-free networks are therefore strictly diverse from networks of the classic Erdős-Rényi type [Erdős and Rényi(1959)], in which node degrees are Poisson-distributed. The importance of scale-free networks in various natural, social, and technological settings (the latter encompassing now ubiquitous structures such as the Internet and the WWW) has motivated considerable research along several fronts during the last decade. For the main results that have been attained the reader is referred to the chapters in [Bornholdt and Schuster(2003), Newman et al.(2006)Newman, Barabási, and Watts].

Most of these research efforts have concentrated on either extracting a scale-free network structure out of data on some particular domain, or the creation of mechanisms of network evolution to function as generative models of such networks. As a consequence, it seems fair to state that so far the greatest thrust has been directed toward what may be called the "syntactic" aspects of scale-free networks, as opposed to their "semantic" (or "functional") aspects, these being related to the higher processes, either natural or artificial, that depend on the underlying networks as a substrate. In the case of computer networks, for example, this issue is illustrated by the networks' topological properties, on the one hand, and their utilization (for end-to-end communication protocols, data storage and retrieval, etc.), on the other.

We see, then, that even as we move from the merely topological aspects of a network toward its higher-level, functional aspects, there remain entities that make up a node's set of local characteristics (e.g., node congestion) which ultimately can be understood as originating higher up at more abstract levels (e.g., the protocols that steer information this way or that as it moves through the network). Clearly, understanding such entities seems to be one of the fundamental keys to better design decisions at the upper levels. And even though the setting of computer networks provides good examples here, other examples can also be considered such as networks of road or street maps, or any other where some kind of end-to-end flows intersect one other.

In this work we study the load of a node in a scale-free network. This property, also known as "betweenness centrality," was analyzed in [Goh et al.(2001)Goh, Kahng, and Kim] and gives, for the node in question, the total communication demand on that node when all node pairs sustain a uniform, bidirectional message traffic between them on shortest paths. Clearly, the load of a node is one of the aforementioned entities, bridging the various levels of abstraction at which the network may be analyzed. The study in [Goh et al.(2001)Goh, Kahng, and Kim] is essentially based on simulations and ends with the conjecture that nodal load is distributed as a power law whose exponent is invariant with respect to τ in the range $(2, 3]$. We follow a different approach, providing both a semi-analytical treatment and results from computational simulations. As we discuss in the sequel, we have found that nodal-load distribution in the scale-free case is richly detailed in a way that can be understood by resorting to appropriate graph-theoretic concepts, such as breadth-first-search (BFS) trees and descents. This contrasts sharply with the purported nature of such a distribution as a power law, and also with the conjecture of a universal exponent [1].

[1]The universality of the nodal-load distribution has also been contested in the correspondence of [Barthélemy(2003), Goh et al.(2003)Goh, Ghim, Kahng, and Kim], but the discussion seems to have lacked a satisfactory conclusion.

2. Descents and nodal load

We conduct our study entirely on undirected random graphs whose degrees are distributed as a power law. Also, in order to avoid any spurious effects resulting from the existence of node pairs joined by no path at all, we concentrate exclusively on each graph's giant connected component (GCC), which for $\tau < 3.47$ is guaranteed to exist [Stauffer and Barbosa(2007)]. For the sake of the analysis in this section, we then assume that G is a connected undirected graph. We let n be the number of nodes in G.

Shortest paths in G are intimately connected with the graph's so-called BFS trees [Cormen et al.(2001)Cormen, Leiserson, Rivest, and Stein]. For each node r of G, a BFS tree of G rooted at r spans all of G's nodes and results from the process of visiting all nodes, beginning at r. Of course, depending on the order of addition of a node's neighbors to the queue, multiple BFS trees may exist for the same root r, and consequently multiple shortest paths from r to each of the other nodes.

Let t_r be the number of distinct BFS trees rooted at r and $T_r^1, \ldots, T_r^{t_r}$ the trees themselves. If T_r^t is one of these trees, then we define the descent of node i in T_r^t, denoted by $d_r^t(i)$, as the number of nodes in the sub-tree of T_r^t rooted at i. This definition is also valid for $i = r$ and includes i in its own descent [thus $d_r^t(i) = n$ if $i = r$ and $d_r^t(i) = 1$ if i is a leaf in T_r^t]. We see that, by definition, $d_r^t(i)$ is the number of shortest paths on T_r^t that lead from r to some other node through node i.

A node's descents are then related to its load [2]. Assuming, as we do henceforth, that the notion of load includes traffic from the node in question to itself, then one possibility for expressing the load of node i in terms of its descents might seem to be to write it as $\sum_{r=1}^{n} \sum_{t=1}^{t_r} d_r^t(i)$. Notice, however, that this would make each pair of nodes weight in the load of node i in proportion to the number of shortest paths between them going through i, which is not acceptable: the definition of load refers to uniform traffic between all node pairs, meaning that the traffic between pairs interconnected by multiple shortest paths is distributed among those paths.

In order to avoid this distortion and still be able to do some mathematical analysis, we consider node i's average descent in trees $T_r^1, \ldots, T_r^{t_r}$, denoted by $d_r(i)$, and substitute it for $\sum_{t=1}^{t_r} d_r^t(i)$ in the previous expression. Since $d_r(i) = \sum_{t=1}^{t_r} d_r^t(i)/t_r$, this corresponds to assuming that each of the multiple shortest paths between a node pair carries the same fraction of the total traffic between the two nodes. If $\ell(i)$ is the load of node i, the approximation we use is then

$$\ell(i) = \sum_{r=1}^{n} d_r(i). \tag{1}$$

As we move to the setting of the GCC of a random graph whose degrees are power-law distributed, even a relation as simple as the one in Eq. (1) on the corresponding random variables is of little help, since a node's descents in the various BFS trees are not independent of one another. For this reason, in the remainder of this section we limit ourselves to pursuing the relatively simpler goal of analyzing the descent distribution of a randomly chosen node in a randomly chosen BFS tree.

[2] This relation has also been pointed out elsewhere in a manner similar to the one we develop in this work [Newman(2004), Newman and Girvan(2004)].

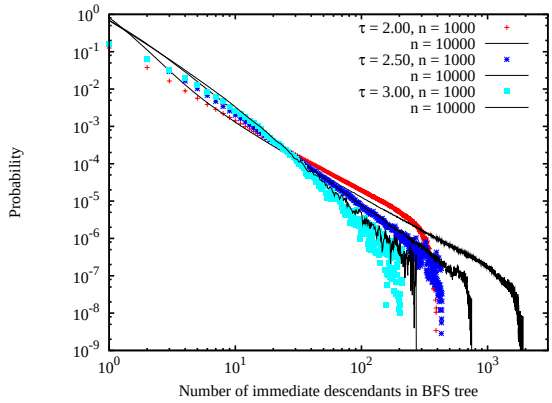

Figure 1. (Color online) Distributions of immediate BFS descents.

If i and r are such a node and the root of such a tree, respectively, and if i has c_i immediate descendants on the tree, then clearly

$$d_r^t(i) = \begin{cases} 1, & \text{if } c_i = 0; \\ 1 + \sum_{j=1}^{c_i} d_r^t(j), & \text{if } c_i > 0, \end{cases} \tag{2}$$

where T_r^t is assumed to be the tree in question. In the case of formally infinite n, it is possible to model descents via the branching process whose branching probabilities are given by the distribution of immediate descendants on the tree. If such a distribution is Poisson, for example, then descents can be found to be distributed according to the Borel distribution [Aldous(1998)]. Other examples include a generalization of the Poisson case, yielding a generalization of the Borel distribution [Barbosa et al.(2003)Barbosa, Donangelo, and Souza]. The branching probabilities of interest to us, however, are of difficult analytical determination, and for this reason, unlike the Poisson case or its aforementioned generalization, there is little hope of determining the descent distribution as a closed-form expression. Even so, some analytical characterization remains within reach. After careful (but tedious) calculation we obtain a form of this expression[Bareinboim and Barbosa(2008)].

3. Computational results and discussion

In respect to our computational methodology, we use $n = 1\,000$ in all our simulations. The reason for such a relatively modest value of n is that, for statistical significance, sufficiently man repetitions are needed for each of the three sources of randomness. These are the number of graphs for each value of τ (we use $10\,000$), the number of roots for each graph (we use all nodes in the graph's GCC, whose number we denote simply by n_{GCC} even though it depends on the graph), and the number of BFS trees for each root (we use 50). For each value of τ, the two distributions of interest (viz. the descent distribution and the nodal-load distribution) can be obtained by computing descents and accumulating them as needed to yield the nodal loads as in Eq. (1).

First of all, to compound the descent distribution, we use the distribution of a node's immediate descendants on BFS trees, which to our knowledge cannot be determined analytically with satisfactory correctness or accuracy [3]. What we do is to resort

[3]One noteworthy attempt is recorded in [Achlioptas et al.(2005)Achlioptas, Clauset, Kempe, and Moore], where the authors ingeniously model the process of BFS-tree construction in continuous time and derive

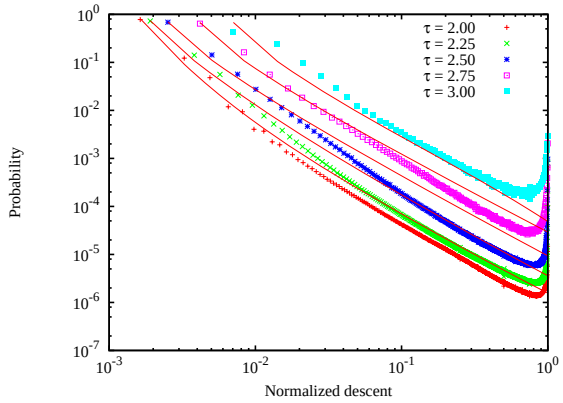

Figure 2. (Color online) Descent distributions. Solid lines give the analytical predictions. Abscissae are normalized to n_{GCC} and binned.

to simulation data to fill in for this distribution, but even this has to be approached carefully, for reasons that are apparent in Fig. 1. In this figure, the distribution of immediate BFS descents within the GCC is shown for three values of τ and two values of n. For fixed τ, the distribution seems to be approximately the same (except for variations due to finite-size effects) for both $n = 1\,000$ and $n = 10\,000$. So, although all our simulations are carried out for the smaller of these values of n, we use simulation data relative to the larger one, since the effects of finite n only become manifest for significantly higher degrees.

Our computational results are summarized in Figs. 2 and 3 for five values of τ in the interval $[2, 3]$. Fig. 2 gives the descent distributions and also their analytical predictions as given by its equation. Since no descent value is larger than the GCC size (n_{GCC}) for the graph in question, all data are shown normalized to the appropriate n_{GCC}: simulation data are normalized to the corresponding GCC sizes occurring during the simulation, and analytical data to the mean GCC size for the τ value at hand.

Notice that all simulated probabilities accumulate significantly at the largest possible normalized descent. While this is clearly due to the finiteness of n, for $\tau \leq 2.75$ it also indicates that, had we been able to afford substantially larger values of n, we could expect this accumulated probability to spread through values of normalized descent one to two orders of magnitude below the maximum and make the simulation data agree with the analytical predictions ever more closely from below. As we discussed in [Bareinboim and Barbosa(2008)], this is in good agreement with the limitations we expect in equation to have for relatively small values of n. As for the remaining value of τ ($\tau = 3$), recall that in this case the effect of relatively small n is considerably severer, since n_{GCC} has a very low mean and is also very widely spread. So, while we may still expect good agreement between simulation and analytical data as n grows, this seems to be reasonable only for values of n even larger than for the previous τ values.

the required probabilities from this model. However, their analysis assumes that degrees in the graph are at least 3 (which we find unreasonable) and, furthermore, seems to involve a probability that is ill defined (may be valued beyond 1). All of this can in principle be fixed in accordance with our work [Bareinboim(2007)], but currently requires BFS queues to be modeled in a way that we think is not possible.

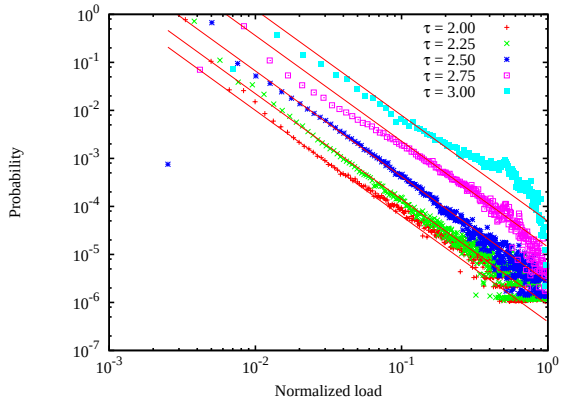

Figure 3. (Color online) Load distributions. Solid lines give power laws of exponent -2.2. Abscissae are normalized to n_{GCC}^2 and binned.

All the simulation data in Fig. 3 are also normalized, but now to n_{GCC}^2, since the greatest load value a node may have grows quadratically with the number of nodes [4]. These data are plotted against power laws of exponent -2.2, which is the exponent that in [Goh et al.(2001)Goh, Kahng, and Kim] is conjectured to be universal with respect to τ for large n. And in fact the agreement of these power laws with the simulation data seems good for $\tau \leq 2.5$, as in these cases GCC sizes have a relatively high mean and low spread. However, unlike the case of the descent distributions, normalizing and binning the raw simulation data for the load distributions has the deleterious effect of masking important information that is present in the raw data and allows nodal-load distributions to be interpreted in terms of the underlying descents.

Filtering the data so that normalization is no longer needed yields figure 4, for example, where the raw simulation data are shown for $\tau = 2$ but restricted to the 95 graphs having $n_{\mathrm{GCC}} = 904$, where 904 is the observed mean GCC size. What we see in this figure is a succession of sharply defined probability peaks. The first peak occurs for a load value of $1\,807$, the second one for $3\,611$, the third for $5\,413$, and so on. If we examine these numbers in the light of Eq. (1), which expresses a node's load in terms of n_{GCC} average descents, one for each possible root, then they can be explained as follows:

- The first peak's location can be decomposed as $1\,807 = 904 \times 1 + 1 \times 903$, and therefore accounts for those nodes whose average descent is 904 for exactly one root (this happens for every node and corresponds to the trees rooted at it) and 1 for all the remaining 903 roots (in whose trees they are leaves). This, clearly, is true of all degree-1 nodes. Note also that the roots in whose trees the nodes in question have average descent 1 constitute the near totality of the roots.
- The location of the second peak can be similarly decomposed, for example as $3\,611 = 904 \times 1 + 903 \times 1 + 2 \times 902$, referring to those nodes whose average descent is 904 when they are root, is 903 for one other root, and 2 for the remaining 902 roots. There may exist degree-2 nodes that conform to this arrangement of average descents, but this is no longer necessary. Also, now it is the roots in whose trees the nodes in question have average descent 2 that constitute the overwhelming majority of the roots.

[4] Consider the case of a star graph and the load of the center node.

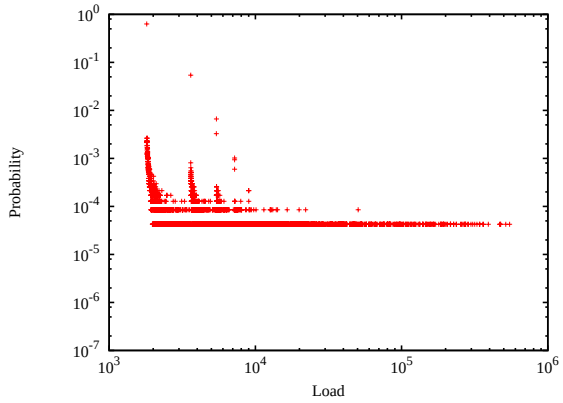

Figure 4. (Color online) Load distribution for $\tau = 2$ and $n_{\mathrm{GCC}} = 904$.

- For the third peak, we can write $5\,413 = 904 \times 1 + 903 \times 2 + 3 \times 901$, now referring to nodes that have average descent 904 when they are root, 903 for two other roots, and 3 for the remaining 901 roots. Once again it is possible, though not necessary, for degree-3 nodes to exist conforming to this arrangement. Continuing the trend established by the previous two cases, the roots in whose trees the nodes in question have average descent 3 are by far the most numerous.

This same pattern of "diophantine" decomposition can be applied to the subsequent peaks and, although the correspondence to node degrees beyond 1 is not guaranteed, we see that peak locations tend to become chiefly determined by the average descents which, from our previous analyses, we know are the most frequently occurring: 1, then 2, then 3, etc.

Naturally, similar filtering can be applied to different values of n_{GCC} with similar results. As for larger values of τ, we remark that the same type of behavior can also be observed, provided τ is sufficiently small for GCC sizes to be relatively large and concentrated around the mean.

4. Concluding remarks

We have considered the load of nodes in scale-free networks and have studied its distribution from the perspective of expressing a node's load in terms of the node's descents in all BFS (or shortest-distance) trees in the graph. We have characterized the descent distribution semi-analytically by resorting to a generating-function formalism and to simulated data on the distribution of immediate BFS descents. We then studied the distribution of nodal load, but through computer simulations only (analytical work in this case would require independence assumptions that we found to be too strong).

Our results have allowed us to revisit the results of [Goh et al.(2001)Goh, Kahng, and Kim] on the load distribution, particularly the conjecture that such a distribution is a power law whose exponent does not depend on τ (i.e., is independent of the underlying graph's degree distribution in the scale-free case). The purported universal exponent of the load distribution is -2.2, and indeed we have been able to confirm that such an exponent seems satisfactorily accurate for large networks after data have been conveniently normalized and binned.

Looking at the raw data, however, reveals that the load distribution is richly structured in a way that can be understood precisely by resorting to the characterization of

nodal load in terms of descents in BFS trees. In our view, this discovery indicates that nodal load is not power-law-distributed and that the conjecture of a universal exponent makes, after all, little sense. Of course, the origin of the previously accepted conclusion and conjecture seems to have been the mishandling of data by inappropriate binning. This, along with other pitfalls of a similar nature, is often the source of inaccurate data interpretation [Clauset et al.(2007)Clauset, Shalizi, and Newman].

We note, finally, that studying quantities like descents in trees and nodal load is well aligned with what we think should be the predominating direction in complex-network investigations. The overwhelming majority of network studies so far have concentrated primarily on structural notions of a predominantly local nature (e.g., node-degree distributions). Descents and loads, on the other hand, are examples of structural notions of a more global nature and, for this very reason, their study constitutes an important step toward complex-network research that emphasizes the networks' functional, rather than structural, properties.

References

E. Bareinboim and V. C. Barbosa, Phys. Rev. E **77**, 046111 (2008).

P. Erdős and A. Rényi, Publ. Math. **6**, 290 (1959).

S. Bornholdt and H. G. Schuster, eds., *Handbook of Graphs and Networks* (Wiley-VCH, Weinheim, Germany, 2003).

M. Newman, A.-L. Barabási, and D. J. Watts, eds., *The Structure and Dynamics of Networks* (Princeton University Press, Princeton, NJ, 2006).

K.-I. Goh, B. Kahng, and D. Kim, Phys. Rev. Lett. **87**, 278701 (2001).

M. Barthélemy, Phys. Rev. Lett. **91**, 189803 (2003).

K.-I. Goh, C.-M. Ghim, B. Kahng, and D. Kim, Phys. Rev. Lett. **91**, 189804 (2003).

A. O. Stauffer and V. C. Barbosa, IEEE ACM T. Network. **15**, 425 (2007).

T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein, *Introduction to Algorithms* (The MIT Press, Cambridge, MA, 2001), 2nd ed.

M. E. J. Newman, in *Complex Networks*, edited by E. Ben-Naim, H. Frauenfelder, and Z. Toroczkai (Springer, Berlin, Germany, 2004), pp. 337–370.

M. E. J. Newman and M. Girvan, Phys. Rev. E **69**, 026113 (2004).

D. Aldous, in *Microsurveys in Discrete Probability*, edited by D. Aldous and J. Propp (American Mathematical Society, Providence, RI, 1998), pp. 1–20.

V. C. Barbosa, R. Donangelo, and S. R. Souza, Phys. A **321**, 381 (2003).

D. Achlioptas, A. Clauset, D. Kempe, and C. Moore, in *Proceedings of the Thirty-Seventh Annual ACM Symposium on Theory of Computing* (ACM Press, New York, NY, 2005), pp. 694–703.

E. Bareinboim, Master's thesis, Systems Engineering and Computer Science Program, Federal University of Rio de Janeiro, Rio de Janeiro, Brazil (2007), in Portuguese.

A. Clauset, C. R. Shalizi, and M. E. J. Newman, *Power-law distributions in empirical data* (2007), URL `http://arxiv.org/abs/0706.1062`.