Exploration Bias of Complex Networks

Paolo De Los Rios

Institut de Physique Théorique, Université de Lausanne, CH-1015, Lausanne, Switzerland and
INFM UdR Torino Politecnico, Corso Duca degli Abruzzi 24, 10129 Torino, Italy.

Abstract. Recent research on complex networks has had a great development thanks to a great abundance of data. Here we address the problem of whether the methods used to obtain these data can influence the data themselves and, as a consequence, the topology of the observed networks.

Complex networks have recently emerged as the suitable tool to describe different systems such as the Internet, the World Wide Web, food webs, protein interaction networks and social relationships [1]. In particular, the flurry of activity has been supported by the availability of huge amounts of data. As a result there are many different models appearing in the literature trying to recover some of the details of the observed networks. Here we want to discuss the reliability of the available data, given the method used to obtain them.

One of the most interesting features of a large class of the complex networks under study now is their scale-free behavior [2]: each node of the network (representing Internet routers, Web pages, proteins or individuals) is connected to some other k nodes. The number of connections (the "degree" in Graph Thoery language) obeys a power-law distribution, *i.e.* $P(k) \sim k^{-\gamma}$, with $2 < \gamma < 3$ for most networks considered. Such networks are dubbed "scale-free" because the fluctuations of the distribution around the average value $< k >$ are infinite (they do not possess any particular scale). The difference between a scale-free network and a random network (where every link between different nodes is present with a probability p, resulting in a Poisson degree distribution) hints towards some mechanisms that generated the observed network features. One of the most celebrated models that explains the emergence of scale-free networks is the Barabasi-Albert (BA) model [3]. According to the BA model, the two essential ingredients for the formation of scale-free networks are growth and preferential attachment. Growth implies that new nodes are added to the network over time, at a more or less constant rate. Preferential attachment means that a newly added node connects preferentially to nodes that already have a high degree: a new node tries to attach to authoritative nodes, and the degree of a node is an effective representation of its authoritativness. It has been shown that, if the probability to connect to a site is linearly proportional to its degree, then growth and preferential attachment indeed generate scale-free networks [4]. In tis simplest realisation, the BA model produces networks with $\gamma = 3$, although it has been shown that, by tinkering with the details of the attachment propbability (but still keeping it linear with the degree), any exponent $\gamma > 2$ can actually be obtained.

Subsequent to the introduction of the BA model, many new models have been introduced, mostly variations of the basic BA rules, in order to obtain the exact exponents seen for real networks (as an example, for the Internet it has been found $\gamma = 2.2(1)$, and for protein networks it has been claimed that $\gamma = 2.4(2)$). Clearly, shifting from "paradigms" to detailed models implies a higher degree of confidence in the data that one would like to explain. So, it is natural to investigate whether the exploration method could influence what we see. We address this problem with two examples inspired by the Internet and by Protein Interaction Networks (PIN's).

The available Internet data are mostly obtained using a command known as "traceroute": given a starting IP address, it tries to find a path connecting it to a given other address, keeping note all other addresses it went through. By asking it to do it in a systo a certain degree (we verified that scale-free networks stay scale-free, and exponential networks exponential). This result leads us to wonder what is the reliability of present network data, at least for the case where they are obtained via tree-like explorations, and, consequently, to what degree we should try to reproduce the detailed data if we are not still sure of their values.

As a second example, we look at networks where links are discovered one by one, somehow picking them in an uncorrelated way. This is the case of large-scale PIN's, where, in a systematic fashion, interactions between proteins (links and nodes of the networks, respectively) are discovered [5]. There are two main methods, presently, to detect a protein-protein interaction: two-hybrid assays and mass spectrometry. In mass spectrometry protein pairs are separated and then identified according to their mass, through a series of procedures such as immunoprecipitation, elution and centrifugation; the two-hybrid assay method is less "physical". Rather it is based on the ability to manipulate genes: first, a transcription factor (a protein that allows the expression of a specific gene) is broken in two parts. The two parts are then glued to the two proteins whose interaction is under investigation. If the two proteins do not interact, the two parts of the transcription factor do not come close together, and the corresponding gene is not expressed; otherwise, if the two proteins interact, the transcription factor is reconstituted and the expression of the gene can be observed, revealing the protein interaction. PIN's obtained either way show again fat-tail behavior, clearly different from random networks, and it has been claimed that actually they also are scale free with exponents $\gamma = 2.3(2)$.

Looking carefully at the methods used to detect interactions, in both cases a large binding constant is needed between the two proteins: for mass spectrometry in order to resist the various processes of separations, in two-hybrid assays so that the two transcription factor fragments stay together long enough to allow gene expression. The strength of a protein-protein interaction is given, at least partially, by their solubility degree: two unsoluble proteins stick to each other much more easily than two soluble ones. Solubility can be measured as a free energy gain Δf to be dissolved ($\Delta f > 0$ for unsoluble proteins), and the free energy of interaction between proteins i and j is therefore $\Delta f_{ij} = \Delta f_i + \Delta f_j$. Asking for a large enough binding constant is tantamount to asking that $\Delta f_{ij} > \Delta f_c$.

Now, let us assume that the "true" PIN is a random network of N proteins, where every possible protein pair is linked with probability p (p in general quite small, the smaller the larger the network: it accounts for specificity of interactions). Yet, we can detect only those interactions for which $\Delta f_{ij} > \Delta f_c$: if we assume that the solubilities Δf's are randomly distributed over the nodes, according, say, to an exponential distribution, $p(\Delta_f) = \exp(-\Delta f)$, we have that the typical number of links of a node of free energy Δf is

$$k(\Delta f) = pN \int_{\Delta f_c - \Delta f}^{\infty} e^{-y} dy = pNe^{-\Delta f_c + \Delta f} \tag{1}$$

that provides us with an expression for $P(k)$ by $P(k)dk = P(\Delta f)d(\Delta f)$, which results in

$$P(k) \simeq \frac{1}{k^2} \tag{2}$$

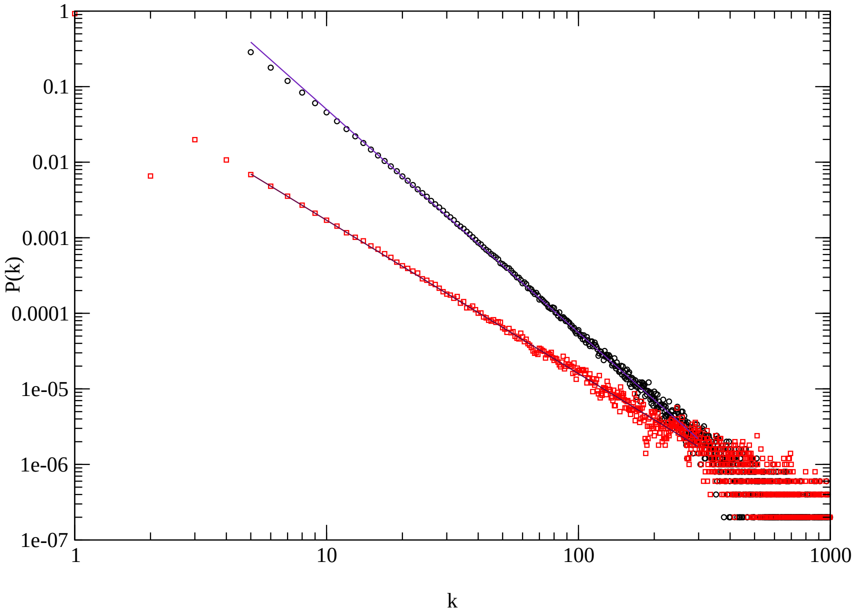

FIGURE 1. Tree-like exploration of a BA network (circles, exponent $\gamma = 3$), where 10% of the links starting from every node are explored (squares, exponent $\gamma = 2.1(1)$).

The result of a simulation is shown in Fig.2 [6].

We see therefore that, although the underlying network was a simple random Poisson one, the observed one shows striking scale-free features, that are a clear artifact of the exploration method (although of course it says something of the physics behind the network).

In conclusion we have analysed the effects of the exploration method on the tolopgy of the network: indeed we have found that different methods can distort the topological properties of the network, and introduce "systematic" errors in the statistics. The extent to which this actually occurs in reality could be uncovered only by exploring the same networks with different methods, and then comparing them.

The author thanks G. Caldarelli, A. Capocci, M.A. Muñoz and T. Petermann for their fruitful collaboration on these topics. This work has been supported by the FET Open Project IST-2001-33555, by the OFES-Bern (grant No. 02.0234) and by the MCYT (grants No. BFM2001-2841 and HI2001-0173).

REFERENCES

1. R. Albert and A.-L. Barabási, *Rev. Mod. Phys.* **74**, 47 (2002).
2. R. Albert, H. Jeong and A.-L. Barabási, *Nature* **401**, 130 (1999).
3. A.-L. Barabási and R. Albert, *Science* **286**, 509 (1999).
4. L. Krapivsky and S. Redner, *Phys. Rev. Lett.* **63**, 066123 (2001).
5. H. Jeong, S.P. Mason, A.-L. Barabási and Z.N. Oltvai, *Nature* **411**, 41 (2001).
6. G. Caldarelli, A. Capocci, P. De Los Rios and M.A. Muñoz, *ArXiv:cond-mat/0207366*, to appear on Phys. Rev. Lett.

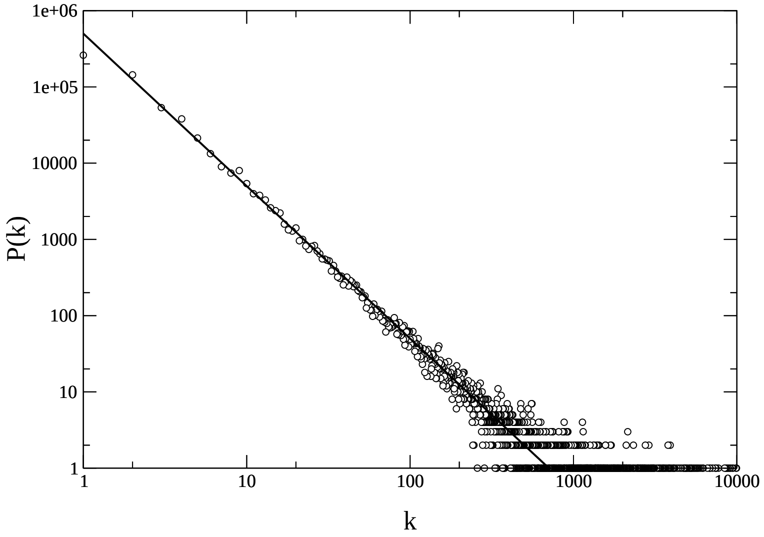

FIGURE 2. Degree distribution for a network of 10000 nodes, with $p = 0.1$, and $\Delta f_c = 10$, average over 1000 realizations. The straight line is k^{-2}.