Immunization Strategies for Networks with Scale-Free Topology

STEFAN MOCANU, SEBASTIAN TARALUNGA
POLITEHNICA University of Bucharest, ROMANIA
Splaiul Independentei 313, Faculty of Control & Computers

Abstract: - The paper propose a model for complex networks with scale-free degree distribution in which the fraction of sites having k connections follows a power law: $P(k)=k^{-\lambda}$. By studying the percolation in such a scale-free model it was established that in the regime $2<\lambda<3$ the networks are resilient to random breakdown. By computing the percolation critical exponents and the resilience to attacks of the scale free networks it was possible to show a similitude between the methods to fight against worms in computer networks and the procedures for immunization against the spread of diseases in social environments. Finally SFN based strategies for stopping an epidemic spread are discussed.

Key-Words: random graphs, scale free networks, computer virus, small world, connectivity degree, power laws

1 Introduction

In a classic paper of Bellovin [1] it is described the Domain-Name-Server (DNS) cache corruption spreading as a *natural computer virus* proliferating on the Internet. Computers on the Internet rely upon DNS servers to translate Internet protocol addresses into computer names and vice-versa. On their turn, DNS servers communicate with their DNS peers to share and update this information. The updating is periodic in time and in the meanwhile, translation tables are “cached” and eventually transmitted to the other DNS peers. If any portion of this cache is corrupted, the DNS server will provide incorrect addresses not only to requesting computers but to DNS peers as well, propagating the error. At the same time, any DNS server can get “cured” by an updating with an error-free DNS peer. The same kind of processes can occur with routing tables exchanged by routers. This propagation of errors occurring on routers and servers that are physically linked is a typical example of epidemic process, in which the corruption (virus) is transmitted from infected to healthy individuals. From a more familiar point of view, however, computer viruses are usually referred to as little programs that can reproduce themselves by infecting other programs [2]. The basic mechanism of infection is as follows: when the virus is active inside the computer, it is able to copy itself, by different ways, into the code of other, clean, programs. When the newly infected program is run into another computer, the code of the virus is executed first, becoming active and being able to infect other programs. Apart from reproducing themselves, computer viruses perform threatening tasks that range from flashing innocuous messages on the screen to seriously corrupt data stored in the computer. These deleterious effects render most computer viruses as dangerous as their biological homonyms, and explain the interest, both commercial and scientific, arisen around their study.

Computer viruses can be classified into three main classes, or *strains*. The first strain includes *file viruses* that infect application programs. A second and more harming family contains the *boot-sector viruses* that infect the boot sector of floppy disks and hard drives, a portion of the disk containing a small program in charge of loading the operating system of the computer. A third and nowadays prevailing strain is formed by the *macro viruses*. These viruses are independent of the platform’s hardware and infect data files, such as documents produced with spreadsheets or word processors. They are coded using the *macro* instructions that are appended in the document, instructions used to perform a set of automatic actions, such as formatting the documents or typing long sequences of characters. In addition, with the ever more efficient deployment of antivirus software, more harmful viruses combining together the properties of the main strains have been developed. Noticeably, however, the nowadays dominant and most aggressive type of cyber organisms is represented by the *worms* family. Worms are actually viruses infecting the computer with mechanisms similar to usual viruses and making a particularly effective use of the e-mail for infecting new computers. In fact, by using the instructions of some commercial mail software applications, worms are capable of sending themselves to all the e-addresses found in the address-book of the person receiving the infected mail. This possibility renders worms the most effective viruses, especially in terms of the velocity at which they can propagate starting from a single infection.

The spreading of computer viruses has been studied for long years, in close analogy with the models developed for the study of the transmission of biological diseases. In this biological framework, the key point is the description of the epidemic process in terms of *individuals* and their *interactions*. In this simplified formalism, individuals can only exist in a discrete set of states, such as susceptible (or healthy), infected (and ready to spread the disease), immune, dead (or removed), etc. On the other hand, the interactions among individuals are schematized in the structure of the contacts along which the epidemics can propagate. Within this formalism, the system can be described as a *network* or graph, in which the nodes represent the individuals and the links are the connections along which the epidemics propagates. Standard epidemiological models usually consider *homogeneous* networks, which are those that have a connectivity distribution peaked at an average connectivity $\langle k \rangle$, and decaying exponentially fast for k<<$\langle k \rangle$ and k>>$\langle k \rangle$. A typical example of deterministic homogeneous network is the standard hyper-cubic lattice, while among the random homogeneous network we can count the Erdos-Renyi model and the Watts-Strogatz model. On the other hand, as we shall see in the following, computer viruses and worms spread in environments characterized by scale-free connectivity. This will lead to the failure of the standard epidemic picture and will naturally introduce the scale-free connectivity as an essential ingredient for the understanding the spread of a computer virus.

2 Graph models for networks

In the 1960s, Paul Erdos and Alfred Renyi initiated the study of random graphs [3]. Random graph theory is, in fact, not the study of graphs, but the study of an ensemble of graphs (or, as mathematicians prefer to call it, a *probability space* of graphs). The ensemble is a class consisting of many different graphs, where each graph has a probability attached to it. A property studied is said to exist with probability *P* if the total probability of all the graphs in the ensemble of having that property is *P*. Two well-studied graph ensembles are $G_{N,M}$ - the ensemble of all graphs having *N* vertices and *M* edges, and $G_{N,p}$ - consisting of graphs with *N* vertices, where each possible edge is realized with probability *p*. These families are known to be similar if $M = \binom{N}{2} p$, so long as *p* is not too close to 0 or 1 and are referred to as ER graphs.

An important attribute of a graph is the average degree, *i.e.*, the average number of edges connected to each node. We shall denote the degree of the *i* th node by k_i and the average degree by $\langle k \rangle$. N -vertex graphs with $\langle k \rangle = \mathrm{O}(N^0)$ are called sparse graphs. In what follows, we concern ourselves exclusively with sparse graphs.

An interesting characteristic of the ensemble $G_{N,p}$ is that many of its properties have a related threshold function, $p_t(N)$, such that if $p<p_t$ the property exists with probability **0** , in the "thermodynamic limit" of N→∞, and with probability **1** if $p>p_t$. This phenomenon is similar to the physical notion of a phase transition. An example of such a property is the existence of a giant component, *i.e.*, a set of connected nodes, in the sense that a path exists between any two of them, whose size is proportional to *N*. Erdos and Renyi showed that for ER graphs such a component exists if $\langle k \rangle > 1$. If $\langle k \rangle < 1$ only small components exist, and the size of the largest component is proportional to ln *N*. Exactly at the threshold, $\langle k \rangle = 1$, a component of size proportional to $N^{2/3}$ emerges. This phenomenon was described by Erdos as the "double jump". Another property is the average path length distance between any two sites, which in almost every graph of the ensemble is of order ln *N*.

Recently, several studies of real world networks have indicated that the ER model fails to reproduce many of their observed properties. One of the simplest properties of a network that can be measured directly is the degree distribution, or the fraction *P(k)* of nodes having *k* connections (degree *k*). A well-known result for ER networks is that the degree distribution is Poissonian, $P(k)=e^{-z}z^k/k!$, where $z=\langle k \rangle$ is the average degree. Direct measurements of the degree distribution for networks of the Internet ([4], [5]), WWW [6], metabolic networks [7], network traffic control [8] and many more, show that the Poisson law does not apply. Most often these nets exhibit a scale-free degree distribution: $P(k)=k^{-\lambda}$, $k=m,..,K$ where $c\approx(\lambda-1)m^{(\lambda-1)}$ is a normalization factor, and *m* and *K* are the lower and upper cut-offs for the connectivity of a node, respectively. The divergence of moments higher than [λ-1] (as *K*→∞ *when N*→∞) is responsible for many of the special properties attributed to scale-free networks.

All real-life networks are finite (and all their moments are finite), so the actual value of the cut-off K plays an important role. It may be approximated by noting that the total probability of nodes with $k>K$ is of order $1/N$ [9] that yields the result $K \sim mN^{1/(\lambda-1)}$. The degree distribution does not characterize the graph or ensemble in full. There are other quantities, such as the degree-degree correlation (between connected sites), the spatial correlations, etc. Several models have been presented for the evolution of scale-free networks, each of which may lead to a different ensemble. The first suggestion was the *preferential attachment* model by Barabasi and Albert, which came to be known as the "Barabasi-Albert (BA)" model. Several variants have been suggested to this model. One of them known as the "Molloy-Reed construction" [10], which ignores the evolution and assumes only the degree distribution and no correlations between nodes, will be considered in the following. Thus, the site reached by following a link is independent of the origin.

3 Distances and bounds in SFN

In most random network models the structure is locally tree-like (since most loops occur only for $n(l) \sim N$), and, since the number of sites grows as $n(l) \sim (k-1)^l$, they are also infinite-dimensional. As a consequence, the diameter of such graphs (*i.e.*, the minimal path between the most distant nodes) scales like $D \sim \ln N$. This small diameter is to be contrasted with that of finite-dimensional lattices, where $D \sim N^{1/d_l}$. Watts and Strogatz [11] have suggested a model which retains the local high clustering of lattices while reducing the diameter to $D \sim \ln N$. This so called *small world network* is achieved by replacing a fraction z of the links in a regular lattice with random links, to random distant neighbors.

We now aim to show that scale-free networks with degree exponent $2<\lambda<3$ has a diameter $D \sim \ln \ln N$, smaller even than that of ER and small world networks. If the network is fragmented, we will only be interested in the diameter of the largest cluster (assuming there is one). In this study we consider the diameter of a Molloy-Reed scale-free network definite as the *average* distance between any two sites on the graph. Actually, it easier still to focus on the radius of a graph, $L \equiv \langle l \rangle$ as the average distance of all sites from the site of highest degree in the network (if there is more than one, we pick one arbitrarily). The diameter of the graph, D, is restricted to $L \leq D \leq 2L$ and thus scales like L .

Cohen, *et al.*, show that the radius of any scale-free graph with $\lambda>2$ has a rigorous lower bound that scales as ln ln N [12]. It is easy to convince oneself that the smallest diameter of a graph, of a given degree distribution, is achieved by the following construction: Start with the highest degree site, then connect to each successive layer the extant sites of highest degree, until the layer is full. By construction, loops will occur only in the last layer.

To bound the radius L of the graph, we will assume that the low degree sites are connected randomly to the giant cluster. We pick a site of degree $1<<k^*<<(\ln\ \ln\ N)^{1/(\lambda=1)}$. If $l_1 \approx \ln\ \ln\ N/\ln(\lambda\text{-}2)$ then $K_{l_1} < k^*$, so, with probability 1 all sites of degree $k \geq k^*$ lie within l_1 layers from the site we picked. On the other hand, if we start uncovering the graph from any site - provided it belongs to the giant component – then within a distance l_2 from this site there are at least l_2 bonds. Since $l = l_1+l_2$, all sites are at a distance of order ln ln N from the highest degree site, and $L=$ ln ln N is a rigorous lower bound for the diameter of scale-free networks with $\lambda>2$.

In a similar way one can demonstrate that the scaling of $D \sim \ln \ln N$ is actually realized in the general case of *random* scale-free graphs with $2<\lambda<3$. For $\lambda>3$ and $N>>1$, k is independent of N , and the radius of the network is $L \sim \ln N$ [13]. The lower bound is obtained from the highest degree site for $\lambda=3$, with $K = m\sqrt{N}$. Then, assuming $\ln \ln N>>1$, the upper bound results $L \sim \ln N/\ln \ln N$. This result has been obtained rigorously for the maximum distance in the BA model where $\lambda=3$, for $m \geq 2$. For $m = 1$, the graphs in the BA model turn into trees, and the behavior of $D \sim \ln N$ is obtained. It should be noted that for $m = 1$ the giant component in the random model contains only a fraction of the sites (while for $m \geq 2$ it contains all sites - at least to leading order). This might explain why exact trees and BA trees are different from Molloy-Reed random graphs.

4 Nodes percolation

The term of site (node) percolation, usually defined on lattices, signify that the sites (nodes) are present (or *occupied*) with probability q, or equivalently, removed (blocked) with probability $p=1-q$. The (infinite) network undergoes a sharp phase transition at a critical threshold q_c from a connected, or percolating phase, where a spanning cluster runs across the entire size of the system, for $q < q_c$, to a fragmented phase, where only finite clusters exist, for $q > q_c$. The percolation transition is continuous (second order), and near the transition point many properties behave as power laws. The percolation threshold is $q_c=1/\langle k\rangle$ for ER graphs.

The problem of percolation on scale-free networks has important practical applications, especially concerning the resilience of the Internet in the face of random breakdown of servers as well as under intentional attack, and to immunization strategies against the spread of contagious epidemics in population and computer networks.

For a graph having degree distribution $P(k)$ to have a spanning cluster, a site j which is reached by following a link (from site i on) the giant cluster must have at least one other link, on average, to allow the cluster to exist. For this to happen the average degree of site j must be at least 2 (one incoming and one outgoing link), given that site i is connected to j. A spanning cluster exists for graphs with $k>2$, while graphs with $k<2$ contain only small clusters whose size β is negligible compared to the entire network. Neglecting the loops is justified below the transition, since the probability for a bond to form a loop in a β-node cluster is proportional to $(s/N)^2$ (*i.e.*, proportional to the probability of choosing two sites in that cluster). An estimate of the fraction of loops P_{loop} in the network yields to S/N where S is the size of the biggest cluster.

The above reasoning can be applied to the problem of percolation in a generalized random network. If we randomly remove a fraction p of the sites (along with the emanating links), the degree distribution of the remaining sites will change. For instance, sites with initial degree k_0 will have, after the random removal of nodes, a different number of connections k, depending on the number of removed neighbors. This yield the critical threshold for percolation - or with the same meaning, the critical epidemic threshold $q_c=1-p_c=1/(R_0-1)$ where the ratio $R_0 \equiv \langle k_0^2\rangle/\langle k_0\rangle$ is calculated using the original distribution, before the random removal of sites. This formula is applicable to random graphs of arbitrary degree distribution. For example, applying the above criterion to the Poisson distribution of ER graphs yields $R=2$, which reduces to the known result $\langle k\rangle=1$.

In the limit of $K>>m$, we have

$$R_0 \rightarrow \left|\frac{2-\lambda}{3-\lambda}\right| \times \begin{cases} m, \lambda>3 \\ m^{\lambda-2}K^{3-\lambda}, 2<\lambda<3 \\ K, 1<\lambda<2 \end{cases}$$

We see that for $\lambda>3$ the ratio R_0 is finite and there is a percolation transition. In *finite* systems a transition is always observed, though for $\lambda<3$ the transition threshold is exceedingly high. For the case of the Internet ($\lambda\approx 5/2$) we have $R_0\approx K^{1/2}\approx N^{1/3}$. Considering the enormous size of the Internet, ($N>10^6$), one needs to destroy over 99% of the nodes before the spanning cluster collapses. For $\lambda>4$ calculation of R shows that it is lower than 2 even before the breakdown occurs.

Because the ultimate goal is to compute the size of the giant component as well as the critical exponents associated with the percolation transition in scale-free networks, a model of interest for study, suggested in [14], is that of intentional attack on the most highly connected nodes of the network. In this model an attacker (*e.g.*, computer hackers trying to cause damage to the network, or doctors trying to disrupt a contagious epidemic) succeeds in knocking off a fraction p of the most highly connected sites in the network. As might be expected, such a strategy is far more effective than *random* dilution. Let now consider analytically the consequence of such an attack, or sabotage, on scale-free networks. Recall that the upper cut-off K , before the attack, may be estimated from $\sum_{k=K}^{\infty} P(k) = 1/N$. Similarly, the new cut-off K', after the attack, follows from $\sum_{k=K'}^{\infty} P(k) - (1/N) = p$.

We estimate the impact of the attack on the distribution of the remaining sites as follows. The removal of a fraction p of the sites with the highest degree results in a random removal of links from the remaining sites - links that had connected the removed sites with the remaining sites. The probability p' for a link to lead to a deleted site equals the ratio of the number of links belonging to deleted sites to the total number of links:

$$p' = \sum_{k=K'}^{K} kP(k)/\langle k_0\rangle$$

where $\langle k_0\rangle$ is the initial average degree.

With the above results we can compute the effect of intentional attack. Essentially, the network after attack is equivalent to a scale-free network with cut-off K', that has undergone random removal of a fraction p' of its sites. The critical fraction is rather sensitive to the lower degree cut-off *m*. For larger *m* the networks are more robust, though they still undergo a transition at a finite p_c. These results can lead for a definition of the critical exponents expressed by

$$\beta = \begin{cases} 1/(3-\lambda), 2<\lambda<3 \\ 1/(\lambda-3), 3<\lambda<4 \\ 1, \lambda>4 \end{cases}$$

In other words, the transition in $2<\lambda<3$ is a mirror image of the transition in $3<\lambda<4$. An important difference is that $q_c=0$ is not λ-dependent in $2<\lambda<3$. Also notice that the critical exponents are not model-dependent but depend only on λ.

5 Efficient immunization strategies

It was shown that epidemic processes in SFN do not possess, in the limit of an infinitely large network, an epidemic threshold below which diseases cannot set into an endemic state. SFN are in this sense very prone to the spreading and persistence of infections. In view of this weakness, it becomes a major task to find optimal immunization strategies oriented to minimize the risk of epidemic outbreaks.

The simplest immunization procedure one can consider consists in the random introduction of immune individuals in the population, in order to get a uniform immunization density. In this case, for a fixed spreading rate *R*, the relevant control parameter is the density of immune nodes present in the network, the immunity *g*. The presence of a uniform immunity will have the effect of reducing the spreading rate *R* by a factor *1-g*, *i.e.* the probability of finding and infecting a susceptible and no immune node will be *R(1-g)*. For homogeneous networks we can easily see that, for a constant *R*, the stationary prevalence R_g is 0 if $g>g_c$ and respectively $(g_c-g)/(1-g)$ if $g \leq g_c$, where g_c is the critical immunization value above which the density of infected individuals in the stationary state is null and depends on *R* as g_c =1- R_c/R. Thus, for a uniform immunization level larger than g_c, the network is completely protected and no large epidemic outbreaks are possible. On the contrary, uniform immunization strategies on SFN are totally ineffective. The presence of uniform immunization is able to locally depress the infection's prevalence for any value of *R* , but it does so too slowly, and it is impossible to find any critical fraction of immunized individuals that ensures the infection eradication

However, we can take advantage of the heterogeneity of SFN, by devising an immunization strategy that takes into account the inherent hierarchy in the network nodes. SFN posses a noticeable resilience to *random* connection failures, which implies that the network can resist a high level of damage (disconnected links), without loosing its global connectivity properties; *i.e.* the possibility to find a connected path between almost any two nodes in the system. At the same time, SFN are strongly affected by *selective* damage; if a few of the most connected nodes are removed, the network suffers a dramatic reduction of its ability to carry information. Applying this argument to the case of epidemic spreading, one can devise a *targeted* immunization scheme [15] in which the most highly connected nodes are progressively make immune. While this strategy is the simplest solution to the optimal immunization problem in heterogeneous populations, its efficiency is comparable to the uniform strategies in homogeneous networks with finite connectivity variance. In SFN, on the contrary, it produces an arresting increase of the network tolerance to infections at the price of a tiny fraction of immune individuals. In order to make an approximate calculation of the immunization threshold in the case of a random SF network, let consider the situation in which a fraction *g* of the individuals with the highest connectivity have been successfully immunized. This corresponds, in the limit of a large network, to the introduction an upper cut-off k_g which is obviously an implicit function of the immunization *g*, such that all nodes with connectivity $k>k_g$ are immune. The introduction of immune nodes implies at the same time the elimination of all the links emanating from them, which translates into a probability *p(g)* of deleting any link in the network. This elimination of links yields the approximate solution for the immunization threshold in the case of targeted immunization as $g_c \approx e^{-2/mR}$. This clearly indicates that the targeted immunization program is extremely convenient in SFN where the critical immunization is exponentially small in a wide range of spreading rates *R* .

The main disadvantage in using targeted immunization is that it requires global knowledge of the topology of the network in question, which is sometimes difficult. A most effective strategy seems to be a selective one, based on the immunization of a small fraction of *random acquaintances* of randomly selected individuals, that prevents epidemics without requiring global knowledge of the network. In a previous section referring to the percolation on broad distribution networks it was shown that a large fraction f_c of the nodes need to be removed (immunized) before the integrity of the network is compromised. This is particularly true for scale-free networks with 2<λ<3 where the percolation threshold $f_c \rightarrow 1$, and the network remains connected (contagious) even after immunization of most of its nodes. To implement such a strategy one chooses a random fraction *p* of the population (of size *N*) and asks each individual to point at an acquaintance to whom they are in contact. The acquaintances, rather than the individuals themselves, are the ones immunized. The fraction f_c needed to be immunized in order to stop the epidemic can be computed analytically.

6 Conclusions

The main goal of this paper has been to study the effect of the special nature of scale-free distribution on the properties of random network models. Some general methods have been presented for the study of generalized random networks. Those include methods for the study of the layer structure of the graph, the percolation threshold and the critical exponents. The paper analyses also an epidemiological framework obtained in population networks characterized by a scale-free connectivity pattern. SFN are very weak in face of infections, and pts susceptibility to epidemic spreading is reflected also in an intrinsic difficulty in protecting them with uniform immunization policies. But targeted or selective immunization procedures achieve the desired lowering of epidemic outbreaks and prevalence. The special properties of scale-free networks, in conjunction with the general method presented for the study of scale-free and other networks, might prove useful for applications such as the design of more robust networks, the improvement of routing algorithms and the prevention of an epidemic broadcast of computer or human viruses. In addition, simple rules defining the temporal patterns of the networks, such as the frequency of forming new connections, the actual time that a connection exists, or different types of connections, should be included trough future work in the aim to obtain better epidemic modelling.

References:

[1] Bellovin, S. M. (1993) Packets found on an Internet. *Comput. Commun. Rev.* **23**, pp.26–31

[2] Harley, C. D., R. Slade, D. Harley, E. H. Spafford and U E. Gattiker (2001) *Viruses Revealed*, McGraw-Hill, New York

[3] P. Erdos and A. Renyi, On the evolution of random graphs. *Publications of the Mathematical Institute of the Hungarian Academy of Sciences* **5**, 17–61 (1960).

[4] Barabasi, A.-L. and R. Albert (1999). Emergence of scaling in random networks. *Science* **286**, pp. 509–512

[5] Faloutsos, M., P. Faloutsos and C. Faloutsos (1999). On power-law relationships of the Internet topology. *Computer Communications Review,* **29**, pp.251–262

[6] Broder, A. R,. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins and J. Wiener (2000). Graph structure in the web. *Computer Networks,* **33**, pp. 309–320

[7] Jeong, H., B. Tombor, R. Albert, Z. N. Oltvai and A.-L. Barabasi (2000). The large-scale organization of metabolic networks *Nature*, **407**, p. 651

[8] Dobrescu, R., M. Dobrescu and St. Mocanu (2004). Using Self Similarity To Model Network Traffic, WSEAS Transactions on Computers, Issue 6, **3**, pp. 1752-1757

[9] Dorogovtsev S. N. and J. F. F. Mendes (2001). Natural Scale of Scale-free Networks. *Phys. Rev. E* **63**, p. 62101

[10] Molloy, M. and B. Reed (1998). The size of the giant component of a random graph with a given degree sequence. *Combin. Probab. Comput.* **7**, pp. 295–305

[11] Watts D. J. and S. H. Strogatz (1998). Collective dynamics of ‘small-world’ networks. *Nature,* **393**, pp.440–442

[12] Cohen, R., K. Erez, D. ben-Avraham and S. Havlin (2000). Resilience of the Internet to Random Breakdown. *Phys. Rev. Lett.* **85**, pp. 4626–4628

[13] Newman, M. E. J.,S. H. Strogatz and D. J. Watts, (2001) Random graphs with arbitrary degree distributions and their applications. *Phys. Rev. E,* **64**, 026118

[14] Albert, R., H. Jeong and A.-L. Barabasi (2000). Attack and error tolerance of complex networks. *Nature,* **406**, pp. 378–382

[15] Pastor-Satorras R. and A. Vespignani (2001). Epidemic spreading in scale-free networks. *Phys.Rev. Lett.* **86**, pp. 3200–3203.