A BRIEF INTRODUCTION TO SCALE-FREE NETWORKS.

William J. Reed*
Department of Mathematics and Statistics
University of Victoria
PO Box 3045
Victoria B.C.
Canada V8W 3P4.
(**e-mail:** reed@math.uvic.ca)

May 18, 2004

*Research supported by NSERC grant OGP 7252.

Abstract

This article provides a brief introduction to scale-free networks (SFNs). The notion of SFN is defined and some examples given. Properties frequently exhibited by SFNs are discussed. The importance of the phenomenon of preferential attachment in generating SFNs is illustrated with two examples for the spread of a persistent disease. The models are similar in that they both yield a total infected population (1) which is geometrically distributed, and growing exponentially in expectation; and (2) in which the average distance from the original source of infection grows in a similar way over time. However one model, which has preferential attachment (infection), yields an SFN, while the other which has homogeneous infectivity does not. The possible application of the theory of SFNs to resource management is briefly discussed.

Keywords: power-law; preferential attachment; small world network; network resilience; persistent disease model.

1 Introduction.

As the Internet and the World-Wide Web (WWW) have grown and become ever more influential and pervasive in daily live, so has the interest in studying the properties of complex and evolving networks of which they are exemplars. While the Internet and WWW are ultimately the product of human choice and technological ingeneuity they are of such complexity that, in many ways, it is convenient to think of them as organic entities evolving according to their own rules. This viewpoint is reflected in the term *internet ecology* which has been coined to describe such an approach to understanding their properties.

It has been discovered that other evolving networks share many properties with the WWW and Internet. These include networks of collaboration among scientists and movie actors; citation networks; telephone-call networks; networks of human sexual contacts; ecological and food webs and networks of metabolic reactions and of protein interactions. In this paper no attempt will be made to describe all of these networks. Rather the reader is referred to three excellent and comprehensive review articles (Albert and Barabási, 2002; Dorogovtsev and Mendes, 2002 and Newman, 2003). Likewise this article will not attempt to cover all of the ground in these survey articles. Rather the aim is to provide a gentle introduction to the subject, in the hope that some of the ideas may prove useful to modellers in the fields of resource management and ecology. To illustrate some of the ideas concerning SFNs discussed in Sec. 2, two very simple models for the spread of a disease are developed in Sec. 3 and contrasted.

2 What are scale-free networks?

Mathematically a network is simply a graph (directed or otherwise) comprising *nodes* (or vertices) some of which are joined by *links* (or edges). The *degree* of a node is simply the number of links attached to that node. For a directed graph each node has an *in-degree* and an *out-degree.* The *degree distribution* of the network is simply the frequency distribution of the degrees of all nodes in the network. The network is called *scale-free* if its degree distribution exhibits power-law behaviour, at least in its upper tail *i.e.* if

$$p_k \sim k^{-\gamma} \qquad \text{as } k \to \infty$$

All of the three survey articles cited in the Introduction have logarithmic plots of the degree distributions of networks such as the WWW, Internet routers, collaboration networks, citation networks *etc.* to illustrate that they possess the scale-free property. The term *scale-free* is used because the degree distribution looks essentially the same when looked at on any scale – or technically that

$$\frac{\sum_{k\geq a} p_k}{\sum_{k\geq b} p_k}$$

depends only on the ratio a/b and not on the individual scales of a and b.

Scale-free networks (SFNs) often exhibit the following properties:

Small-world property. This is the property that the length of the shortest path between any two nodes is small compared to the size of the network. A well-known example of this is the "six degrees of separation" concept of the social psychologist S. Milgram (1967) who observed that there was a path of typical length about 6 linking any two people in the U.K. More technically the idea can be expressed as the property that the diameter of the network (maximum path length between any two nodes) is bounded by a polynomial in $\log n$ where n is the number of nodes in the network (see *e.g.* Baldi *et al.*, 2003)

Clustering. This is the property that the network comprises many clusters or cliques. Within in each cluster there is a high density of links, but between clusters the density of links is much lower.

Network resilience. If the removal of one or more nodes results in a considerable increase in the distance between nodes (or indeed causes some nodes to become disconnected) then the efficiency of the network (at least as a communications network) will be diminished. The resilience of the network is its vulnerability to removal of nodes. Albert *et al.* (2000) claim that the Internet and WWW are highly resilient to random removal of nodes, but are highly vulnerable to deliberate attack on the nodes of highest degree.

Albert and Barabási (2002) give two conditions which they identify as being necessary for a network to have the scale-free property, *viz.*

(i) **growth** – the network grows over time;

(ii) **preferential attachment** – nodes with a high degree are more likely to create links to new nodes than nodes with a low degree.

One can see that the WWW satisfies both of these conditions. New nodes are being added constantly and at the same time well-connected nodes (*e.g.* Google, Adobe *etc.*) are more likely to be linked to new nodes than more obscure, low-degree nodes (*e.g.* the Web page of the author).

In the following Section we present two network models for the spread of a persistent disease, which illustrate the importance of preferential attachment. While similar in some respects (growth in size, mean distance from original source of infection) the models differ in the important respect that one exhibits the scale-free property, while the other does not. One provides a crude model for the spread of a sexually transmitted disease, while the other could provide a crude model for the spread of, for example, oral herpes.

3 Simple models for the spread of an infectious disease.

The spread of an infectious disease within a community can be represented by a growing tree network, with infected individuals represented by nodes, and with any two nodes A and B considered connected if A infected B or *vice versa*. Suppose that one individual introduced the infection into the community at time $t = 0$ and that in the infinitesimal time increment $(t, t + h)$ any infected individual i can infect a new individual with probability $\lambda_i(t)h + o(h)$ and that all new infections are independent. It will be assumed that the disease is persistent, so that once infected an individual remains infectious.

For some diseases it is reasonable to assume that individuals are homogeneous with respect to their infectivity, and thus that $\lambda_i(t) \equiv \lambda$ (a constant). However when considering a sexually-transmitted disease, it is important to recognize the fact that individuals vary greatly in their promiscuity, so that the above homogeneity assumption is not appropriate. The two cases are considered in turn in the following sub-sections.

3.1 A homogeneous model.

Let the number of individuals infected (nodes) at time t be denoted by $N(t)$ and assume $N(0) = 1$. Let $K(t)$ denote the number of links to a specified node (call it node *), so that $K(t) - 1$ is the number of individuals infected by * by time t. Let

$$p_{k,n}(t) = \mathrm{P}(K(t) = k, N(t) = n)$$

with $p_{k,n}(t) = 0$ for $n < 0$ or $k < 0$. Then conditioning on what happens in $(t, t + h)$ one obtains

$$p_{k,n}(t + h) = p_{k-1,n-1}(t)\lambda h + p_{k,n-1}(t)\lambda h(n - 2) + p_{k,n}(t)[1 - \lambda hn] + o(h)$$

The first term on the right-hand side corresponds to a new node infected by * in $(t, t + h)$, while the second term corresponds to a new node infected by a node other than *. The third term corresponds to no new infections in $(t, t + h)$. Subtracting $p_{k,n}(t)$ from both sides, dividing by h and passing to the limit as $h \to 0$ yields the following Kolmogorov forwards equation

$$\frac{d}{dt}p_{k,n} = \lambda p_{k-1,n-1} + \lambda(n - 2)p_{k,n-1} - \lambda np_{k,n}. \tag{1}$$

Summing over, in turn n and k, yields the following differential equations for $p_n(t) = \mathrm{P}(N(t) = n)$ and $p_k(t) = \mathrm{P}(K(t) = k)$:

$$\frac{d}{dt}p_k = \lambda p_{k-1} - \lambda p_k \tag{2}$$

and

$$\frac{d}{dt}p_n = \lambda(n-1)p_{n-1} - \lambda n p_n. \tag{3}$$

The first of these equations is easily recognized as the Kolmogorov equation of a *Poisson process* and the second as the Kolmogorov equation of a *Yule process* (homogeneous pure birth process – see *e.g.* Karlin and Taylor, 1975). From standard results it follows that

$$p_n(t) = e^{-\lambda t}(1 - e^{-\lambda t})^{n-1} \tag{4}$$

i.e. that $N(t)$ follows a *geometric distribution* with parameter $e^{-\lambda t}$ and

$$\mathrm{E}(N(t)) = e^{\lambda t} \qquad \mathrm{var}(N(t)) = e^{\lambda t}(1 - e^{\lambda t}). \tag{5}$$

Also one can show

$$\mathrm{E}\left(\frac{1}{N(t)}\right) = \frac{\lambda t e^{-\lambda t}}{1 - e^{-\lambda t}} \tag{6}$$

Now let t^* denote the time of infection of node $*$, so that $K(t^*) = 1$. It then follows from standard results on the Poisson process that for $t \geq t^*$

$$p_k(t) = e^{-\lambda(t-t^*)}\frac{[\lambda(t-t^*)]^{k-1}}{(k-1)!} \tag{7}$$

i.e. that $K(t) - 1$ follows a Poisson distribution with parameter $\lambda(t - t^*)$.

Both the Poisson process and the Yule process exhibit what is known as the *order statistic* property (Feigin, 1979). For such processes it can be shown that the ordered event times have the same distribution as that of the order statistics of independent identically distributed random variables (iid rvs). For example for the Poisson process the event times in $[0, t]$ have the same distribution as iid rvs uniformly distributed on $[0, t]$.

From the corresponding result for the Yule process it can be shown that the times since infection of nodes in existence at time t (other than the primal node from which the infection originated) are independent random variables following a truncated exponential distribution with probability density function (pdf)

$$f(\tau) = \frac{\lambda e^{-\lambda\tau}}{1 - e^{-\lambda t}}, \qquad 0 \leq \tau \leq t.$$

This gives the distribution of the time in existence $t - t^*$ of all nodes in the network (save the primal node). To include the primal node, this distribution must be mixed with an atomic distribution at $\tau = t$ (with mixing weights $\mathrm{E}(1/N(t))$ and $1 - \mathrm{E}(1/N(t))$. By integrating (7) with respect to this mixed distribution, and using (6) one obtains the probability mass function (pmf) of the degree distribution of a randomly chosen node from the network at time t as:

$$p_k(t) = \frac{e^{-2\lambda t}(\lambda t)^{k+1}}{k!(1 - e^{-\lambda t})} + \frac{\lambda(1 - (1 + \lambda t)e^{-\lambda t})}{(k-1)!(1 - e^{-\lambda t})^2} \int_0^t (\lambda\tau)^{k-1} e^{-2\lambda\tau} d\tau. \qquad (8)$$

The integral can be expressed in terms of an incomplete gamma function. Now let $t \to \infty$ so that $\mathrm{E}(1/N(t)) \to 0$ to get

$$p_k \to 2^{-k} \qquad \text{for } k = 1, 2, \ldots \qquad (9)$$

a geometric distribution with parameter 1/2. Thus we can conclude that at a suitably long time after the introduction of the infection, the degree distribution over the network will follow a geoemetric form, with parameter 1/2. Note that this distribution is *not* scale free ($\sum_{k\geq a} 2^{-k}/\sum_{k\geq b} 2^{-k} = 2^{b-a}$ which depends on the scale of a and b not just their ratio).

Other properties of this model can be established *e.g.* it can be shown that the distribution of the *ring number* R_t (the distance of a node to the primal node) has a distribution over the network

$$\mathrm{P}(R_t = r) = \frac{(\lambda t)^{r-1}}{(e^{\lambda t}-1)(r+1)!} \quad r = 0, 1, \ldots$$

i.e. $R_t + 1$ has a zero-truncated Poisson distribution (Chan *et al.*, 2003). Also $\mathrm{E}(R(t)) = \lambda t(1-e^{-\lambda t})^{-1} \sim \lambda t$ as $t \to \infty$.

In summary, for this homogeneous model (i) the size of the infected population grows exponentially in expectation (and is geometrically distributed); (ii) after a suitable time the degree distribution follows close to a geometric form (not scale-free); and (iii) the ring number distribution is related to a Poisson distribution.

We turn now to a non-homogeneous model for the spread of an STD.

3.2 A non-homogeneous model for the spread of an STD.

A characteristic of networks of sexual partners is the extreme variability in degree (numbers of partners) over the network. Indeed, Liljeros *et al.*, (2001) inferred from data in a Swedish survey, that such networks are scale-free. In light of this the homogeneity assumption, used in the previous sub-section, appears innappropriate for describing the spread of an STD.

In its place suppose that the infectivity rate $\lambda_i(t)$ for individual i is of the form

$$\lambda_i(t) = \rho K_i(t)$$

where $K_i(t)$ is the degree of node i. Thus individuals who have already infected many people are more likely to infect a new person than those individuals who have to date infected few or no others. This assumption of preferential attachment captures in a coarse way the idea that promiscuous individuals will be more important than their more restrained fellows in spreading the STD.

Under this assumption the Kolmogorov equation analogous to (1) is

$$\frac{d}{dt}p_{k,n} = \rho(k-1)p_{k-1,n-1} + \rho(2n-4-k)p_{k,n-1} - \rho(2n-2)p_{k,n}. \qquad (10)$$

Summing out respectively n and k leads to (analogous to equations (2) and (3))

$$\frac{d}{dt}p_k = \rho(k-1)p_{k-1} - \rho k p_k \qquad (11)$$

and

$$\frac{d}{dt}p_n = 2\rho(n-2)p_{n-1} - 2\rho(n-1)p_n. \qquad (12)$$

The first of these equations is the Kolmogorov equation for a Yule process and so it follows that the distribution of the degree of node * at time t (*i.e.* $t - t^*$ time units after node * was infected) is (for $t \geq t^*$)

$$p_k(t) = e^{-\rho(t-t^*)}(1 - e^{-\rho(t-t^*)})^{k-1}, \qquad \text{for } k = 1, 2, \ldots \qquad (13)$$

To solve (12) one needs to specify an initial condition - so assume $N(0) = 2$ (although infection may begin with one individual with degree one, this individual will eventually infect another – with probability $1 - e^{-\rho t}$ by time t). The solution to (12) with this initial condition is

$$p_n(t) = e^{-2\rho t}(1 - e^{-2\rho t})^{n-2}, \qquad \text{for } n = 2, 3, \ldots \tag{14}$$

i.e. $N(t) - 1$ is geometrically distributed with parameter $e^{-2\rho t}$ with

$$\mathrm{E}(N(t)) = 1 + e^{2\rho t} \qquad \mathrm{var}(N(t)) = e^{2\rho t}(1 - e^{2\rho t}). \tag{15}$$

Equation (12) is the Kolmogorov equation for a non-homogeneous birth process with instananeous birth rate

$$\mathrm{P}(\text{ birth in } (t, t+h]|N(t) = n) = 2\rho(n-1).$$

From this it follows that the number of new infections $U(t) = N(t) - 2$ by time t is an order statistic process, so that the times of these new infections are iid rvs with support on $(0, t]$. In fact as before, the distribution of the times since infection, $t - t^*$, of these new infections follow the trunctaed exponential distribution

$$f(\tau) = \frac{2\rho e^{-2\rho\tau}}{1 - e^{-2\rho t}}, \qquad 0 \le \tau \le t.$$

As for the homogeneous model in Sec 3.1, the distribution of the degree of a randomly selected node is found by integrating the pmf (13) with respect to this density. Doing this and letting $t \to \infty$ leads to

$$\lim_{t\to\infty} p_k(t) \;=\; \int_0^\infty 2\rho e^{-2\rho\tau} e^{-\rho\tau}(1 - e^{-\rho\tau})^{k-1} d\tau$$

$$\begin{aligned} &= \frac{2\Gamma(3)\Gamma(k)}{\Gamma(k+3)} \\ &= \frac{4}{k(k+1)(k+2)} \quad \sim \quad 4k^{-3}. \end{aligned} \tag{16}$$

Thus for this model at a suitably long time after the introduction of the infection, the degree distribution over the network of infected individuals exhibits a power-law behaviour in the upper tail and thus is scale-free, with exponent -3. It is worth noting that Liljeros *et al.*, (2001) estimated the exponent for the Swedish network of lifetime sexual partners as -3.1 ± 0.3 for females and -2.6 ± 0.3 for males.

Chan *et al.* (2003) obtain the following mean-field approximation to the distribution of the ring number

$$P(R(t) = r) = \frac{e^{\lambda t}(\lambda t)^{r-1}}{(e^{2\lambda t}+1)(r-1)!} \quad r = 1, 2, \dots \tag{17}$$

with $P(R(t) = 0) = 1/(e^{2\lambda t} + 1)$. This distribution is related to the Poisson distribution and has mean value $(\rho t + 1)(1 + e^{-2\rho t})^{-1} \sim \rho t + 1$ (as $t \to \infty$).

Comparing the results for this model with those of the homogeneous infection model it can be seen that they behave similarly in terms of the growth of the network (both grow exponentially in expectation); and likewise in terms of ring number (both grow linearly in expectation in t for large t). They differ however in the degree distributions over the network. The homogeneous model produces a geometric degree distribution (non-scale free); while the non-homogeneous (STD) model produces a degree distribution which is scale-free. It is the phenomenon of preferential attachment with gives rise to this scale-free behaviour.

A more complex two-sex model for the spread of an STD has been developed (see Reed, 2004). The network in this case is a bipartite graph, with males able to infect females with infectivity rate $\lambda K_i(t)$; and females able to infect males with infectivity rate $\rho L_j(t)$, where $K_i(t)$ and $L_j(t)$ are numbers, respectively, of females infected by male i and of males infected by female j by time t. For this model as $t \to \infty$, the degree distribution for males is scale-free with exponent $-(2 + \rho/\lambda)$; and that for females scale-free with exponent $-(2 + \lambda/\rho)$. Typically males are more promiscuous than females, so one would expect $\lambda > \rho$. This implies that the power-law exponent is smaller in absolute value for males than females, corresponding to a longer tail in the degree distribution for males. This is in agreement with the results of Liljeros (2001) *et al.* mentioned above.

4 Conclusions.

Scale-free and other complex evolving networks have received much attention in recent years. Whether the results of this research is important for resource management remains to be seen. One possible application may be in terms of the resilience of an ecological network to harvesting certain species. There is some evidence that food webs exhibit small-world and clustering properties (Williams *et al.*, 2002; Camacho *et al.*, 2002) and it has also been claimed by Montoya and Solé (2002) that the food webs they have studied are consistent with a scale-free degree distribution with a small exponent (≈ 1.1) *i.e* with a very long tail. If indeed food webs are scale-free, they may be vulnerable to the harvesting of certain well-connected nodes (keystone species). Knowledge of the resilience of such ecological networks may provide important information to resource managers.

With respect to disease control and prevention, knowledge of the nature of the network of connections among susceptible individuals may prove to be of great value. A disease will spread much faster on a SFN than on a more homogeneous network, and treating or isolating individuals with a high degree of connectivity may prove to be the most effective way of controlling the disease. There has been debate recently as to whether networks of sexual partners, if indeed scale-free, have a power-law exponent greater or less than 3 (Jones and Handcock, 2003). If less than three Jones and Handcock claim that no epidemic threshold would exist for STDs, in which case the infection could persist indefinitely regardless of interventions such as vaccination or barrier contraceptives *etc.*

References.

Albert, R. and A.-L. Barabási, [2002]. Statistical mechanics of complex networks. *Rev. Modern Phys.*, **74**, 47–97.

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

Baldi, P, P. Frasconi and P Smyth [2003]. *Modeling the Internet and the Web.* John Wiley & Sons, Chichester, U.K.

Camacho, J., R. Guimer and L.A.N. Amaral, [2002]. Robust patterns in food web structure. *Phys. Rev. Letters*, **88**, 228102.

Chan, D. Y. C., B. D. Hughes, A. S. Leong and W. J. Reed, [2002]. Stochastically evolving networks. *Phys. Rev. E*, **68**, 066124.

Dorogovtsev, S. N. and J. F. F. Mendes, [2002]. Evolution of networks. *Adv. in Phys.*, **51**, 1079–1187.

Feigin, P. D. [1979], On the characterization of point processes with the order statistic property. *J. Appl. Prob.* **16**, 297–304.

Jones, H. J. and M. S. Handcock, [2003]. Sexual contacts and epidemic thresholds. *Nature*, **423**, 605–606.

Karlin, S. and H. M. Taylor [1975]. *A First Course in Stochastic Processes, Second Edition.* Academic Press, New York.

Liljeros, F., C. R. Edling, L. A. N. Amaral, H. E. Stanley and Y. Åberg [2001]. The web of human sexual contacts. *Nature*, **411**, 907-908.

Milgram, S. [1967]. The small world problem. *Physch. Today*, **2**, 60-67.

Montoya, J. M. and R. V. Solé. [2002] Small world patterns in food webs. *J. Theor. Biol.*, **214**, 405-412.

Newman, M. E. J. [2003]. The structure and function of complex networks. *SIAM Rev.* **45**, 167-256.

Reed, W. J. [2004]. A scale-free network model for the spread of sexually transmitted diseases. Submitted to *Math. Biosciences.*

Williams, R. J., E. L. Berlow, J. A. Dunne, A.-L Barabasi and N. D. Martinez, [2002]. Two degrees of separation in complex food webs. *Proc. Natl. Acad. Sci. U. S. A.* **99**, 12913-12916