Power-law distributions from exponential processes: an explanation for the occurrence of long-tailed distributions in biology and elsewhere

William J. Reed
Department of Mathematics and Statistics,
University of Victoria,
Victoria, British Columbia, Canada V8W 3P4
reed@math.uvic.ca
and
Barry D. Hughes
Department of Mathematics and Statistics,
University of Melbourne,
Victoria 3010, Australia
hughes@ms.unimelb.edu.au

Abstract. A possible explanation for the frequent occurrence of power-law distributions in biology and elsewhere comes from an analysis of the interplay between random time evolution and random observation or killing time. If the system population or its topological parameters grow exponentially with time, and observations on the system correspond to stopping the evolution at an exponentially distributed random time, power-law behaviour in one or both tails of the distribution of observed quantities may result.

We pursue this theme for two specific models. The first model is a randomly killed birth-and-death process, with applications to the numbers of genes per gene family and proteins per protein family, the distribution of taxonomic elements in live taxa, and other areas. The second model is a randomly growing network, with the state of a random node (which thus has a random age) observed. For the growing network, we consider both tree-like networks, appropriate in biological applications, and networks in which closed loops can appear, which model communication networks and networks of human sexual interactions.

1 Introduction Distributions exhibiting power-law (or fractal) behaviour in one or both tails are widespread in biology and elsewhere. Within biology these include various size distributions and the connectivities of nodes in various networks. Size distributions for which power-law tail behaviour has been claimed include the numbers of genes per gene family and proteins per protein family; the number of species per genus; areas burned in wildfires and the length distribution of terrestial animal species. Some biological networks for which the distribution of connectivities of nodes are claimed to exhibit power-law behaviour include protein-protein interactions (proteins being connected if they bind together); food webs and networks of human sexual contacts.

Other examples just outside the realm of biology are the size distribution of family name clades[1] and of language taxa; and the connectivities of words in languages ('word webs'). In the non-biological world some examples are the size distributions of cities, incomes, sand

1991 *Mathematics Subject Classification.* 62P10, 90B15, 92D10, 92D20.
Key words and phrases. power-law distributions, random networks.

[1] In view of this one might expect that the size distribution of haplogroups sharing the same mitochondrial DNA would also exhibit power-law behaviour, given that the mechanism through which mtDNA is transmitted is similar to that through which family names are transmitted.

particles and World Wide Web file sizes along with the connectivities of World Wide Web sites[2].

An obvious question to ask is why these phenomena and many others share similarities in their distributional form: one furthermore in which extremes of size, connectivity etc. readily occur? For the widespread occurrence of the normal distribution in biology and elsewhere there is a simple and elegant explanation, based on de Moivre–Laplace central limit theorem. It seems natural to ask whether there is a similar explanation for the widespread occurrence of distributions with power-law tails. We have provided an answer to this question elsewhere (Reed and Hughes 2002*b*). The purpose of this paper is to provide an outline of the theory and to show how it can be applied to explain the phenomenon of power-law behaviour in some examples from biology.

Put very simply the explanation is that if a process that is growing exponentially in some loose sense is stopped, or observed after an exponentially distributed time, the state of the process at the time of stopping or observation will exhibit power-law behaviour in one or both tails. To illustrate consider the simplest case in which deterministic exponential growth $X(t) = X_0 e^{\mu t}$ is stopped after an exponentially distributed random time. The stopped state is $\bar{X} = X_0 e^{\mu T}$ where T has an exponential distribution, with parameter λ (say), that is, $\Pr\{T > t\} = e^{-\lambda t}$. A simple calculation shows that $\bar{X}$ has a density of the form

$$f_{\bar{X}}(x) = kx^{-\lambda/\mu - 1} \quad \text{for} \quad x > X_0,$$

a Pareto distribution, exhibiting power-law behaviour over its full range. Reed and Hughes (2002*b*) show how this result can be extended to include stochastic processes that are growing exponentially or geometrically *in expectation*. Specifically in continuous time, if either geometric Brownian motion or a homogeneous birth-and-death process are stopped after an exponentially distributed time (*i.e.* with a constant stopping rate); or in discrete time, if either a multiplicative process or a Galton–Watson branching process are stopped after a geometrically distributed number of steps (*i.e.* with a constant stopping probability), then the distribution of the resulting stopped state will exhibit power-law behaviour in one or both tails.

In this article we shall concentrate on models of biological phenomena of the birth-and-death process type, and show how they can be used to explain the observed power-law behaviour in the size distributions of gene and protein families and of genera, and in the distribution of connectivities of nodes in some biological networks.

2 A model for the size distribution of genera and gene and protein families

Consider a genus which begins with one species (or a gene or protein family which begins with one gene or protein) at time $t = 0$. Suppose that in time $(t, t+h)$ there is a probability $\lambda h + o(h)$ that any given species may speciate, giving rise to a new species (or any gene may mutate and create in addition to replicates of itself, a new gene in the family); and a probability $\mu h + o(h)$ that the individual species may go extinct (or a gene alone be selected out of the genome). Suppose further that all speciations and extinctions are independent. Under these assumptions, N_t, the number of species in the genus (genes in the family) in existence at time t follows a homogeneous birth-and death process (see *e.g.* Bailey 1964) for which the probability mass function (p.m.f.)

$$p_n(t) = \Pr\{N_t = n\} \tag{1}$$

[2]Although these last two examples are from outside of biology, their similarity to biological pheneomena is recognized in the name 'internet ecology' used to describe the study of such phenomena.

satisfies the differential-difference equation

$$\frac{d}{dt}p_n(t) = -(\lambda+\mu)np_n(t) + \lambda(n-1)p_{n-1}(t) + \mu(n+1)p_{n+1}(t), \tag{2}$$

with initial condition

$$p_n(0) = 1 \text{ if } n = 1; \quad p_n(0) = 0 \text{ otherwise.}$$

Now let

$$\phi(z;t) = \sum_{n=0}^{\infty} p_n(t)z^n \tag{3}$$

be the generating function for N_t. Multiplying both sides of (2) by z^n and summing over $n = 0, \ldots, \infty$ yields the partial differential equation

$$\phi_t = (\lambda z - \mu)(z-1)\phi_z, \tag{4}$$

with initial condition

$$\phi(z,0) = z. \tag{5}$$

This can readily be solved by the method of characteristics (see *e.g.* Bailey 1964) to yield

$$\phi(z,t) = \begin{cases} \dfrac{\mu(1-z) - (\mu - \lambda z)\exp[-t(\lambda-\mu)]}{\lambda(1-z) - (\mu-\lambda z)\exp[-t(\lambda-\mu)]} & \text{if } \lambda \neq \mu, \\ \\ 1 - (1-z)/[1+\lambda t(1-z)]^{-1} & \text{if } \lambda = \mu. \end{cases} \tag{6}$$

From this the well-known formulas for the p.m.f. of $N(t)$ can be derived. In the case $\lambda \neq \mu$

$$p_0(t) = \frac{\mu - \mu e^{-t(\lambda-\mu)}}{\lambda - \mu e^{-t(\lambda-\mu)}}; \tag{7}$$

$$p_n(t) = \frac{(\lambda-\mu)^2 e^{-t(\lambda-\mu)}}{[\lambda - \mu e^{-t(\lambda-\mu)}]^2}\left\{\frac{\lambda - \lambda e^{-t(\lambda-\mu)}}{\lambda - \mu e^{-t(\lambda-\mu)}}\right\}^{n-1}, \quad n \geq 1. \tag{8}$$

Since different genera (gene families) will have originated at different times, in order to obtain the p.m.f. (or generating function) of the unconditional distribution of family size $\bar{N}$ say, the p.m.f. $\{p_n(t)\}$ or the generating function $\phi(z,t)$ must be integrated with respect to the distribution of t over genera (gene families). It seems reasonable to assume that any genus (or gene family) originated when an individual species mutated to a form so different from others in the genus (gene family) that it could no longer be considered a member of that genus (gene family). Let us suppose that such radical mutations can occur in any existing genus (family) in a time interval of length h with probability $\rho h + o(h)$. This implies that the number of genera (gene families) follows a Yule process (Yule 1924), and that the time in existence of any genus will follow a truncated exponential distribution, the truncation time being the time since the establishment of the first genus in the family (Feigin 1979). Since evolution has been happening for a very long time, this truncation can essentially be ignored, so that the p.m.f. of the distribution of current genus (or gene family) size, $\bar{N}$, is

$$q_n = \Pr(\bar{N} = n) = \int_0^{\infty} p_n(t)\rho e^{-\rho t}dt \quad \text{for } n = 0, 1, \ldots \tag{9}$$

and its generating function is

$$\bar{\phi}(z) = \int_0^\infty \phi(z,t)\rho e^{-\rho t} dt. \tag{10}$$

Neither of the integrals in (9) or (10) have simple closed-form expressions. However by returning to the partial differential equation (36), multiplying throughout by $\rho e^{-\rho t}$ and integrating with respect to t between 0 and ∞, one arrives at the following ordinary differential equation for $\bar{\phi}(z)$:

$$(\lambda z - \mu)(z-1)\frac{d\bar{\phi}}{dz} - \rho\bar{\phi}(z) = -\rho z. \tag{11}$$

While this can be solved in terms of Lerch's phi function (see Reed and Hughes 2002*a*), the form is not particulary useful. However a series solution can be obtained by equating coefficients of powers of z on both sides of (11), yielding the following recursion for the p.m.f. $\{q_n\}$

$$(n-1)\lambda q_{n-1} - [n(\lambda+\mu)+\rho]q_n + (n+1)\mu q_{n+1} = 0, \quad \text{for } n \geq 2 \tag{12}$$

$$-(\lambda+\mu+\rho)q_1 + 2\mu q_2 = -\rho \tag{13}$$

$$-\rho q_0 + \mu q_1 = 0. \tag{14}$$

The p.m.f. $\{\tilde{q}_n\}$ of the size of a non-extinct gene family is obtained as

$$\tilde{q}_n = \frac{q_n}{1-q_0}, \qquad n = 1, 2, \ldots \tag{15}$$

2.1 Power-law behaviour in the genus (gene family) size distribution It is shown in this section that in the case $\lambda > \mu$, the distribution of genus size exhibits power-law behaviour in the upper tail. Consider (from (9) and (8))

$$q_{n+1} = \int_0^\infty \frac{\rho e^{-\rho t}(\lambda-\mu)^2 e^{-t(\lambda-\mu)}}{[\lambda - \mu e^{-t(\lambda-\mu)}]^2} \left\{\frac{\lambda - \lambda e^{-t(\lambda-\mu)}}{\lambda - \mu e^{-t(\lambda-\mu)}}\right\}^n dt. \tag{16}$$

The factor in braces is strictly increasing in t from zero at $t = 0$, approaching unity as $t \to \infty$. Effecting a change of variable from t to τ with

$$t = (\lambda-\mu)^{-1}\log[n(1-\mu/\lambda)/\tau], \tag{17}$$

gives for $n \to \infty$,

$$\left\{\frac{\lambda - \lambda e^{-t(\lambda-\mu)}}{\lambda - \mu e^{-t(\lambda-\mu)}}\right\}^n \sim \left\{1 - \frac{\tau}{n}\right\}^n \to e^{-\tau}, \tag{18}$$

so that in (16)

$$q_{n+1} \sim \frac{\rho}{\lambda}\Big(1 - \frac{\mu}{\lambda}\Big)^{-\rho/(\lambda-\mu)} n^{-1-\rho/(\lambda-\mu)} \int_0^\infty e^{-\tau}\tau^{\rho/(\lambda-\mu)} d\tau. \tag{19}$$

The integral can be evaluated as a gamma function. The important conclusion is that for large n, the p.m.f. exhibits power-law behaviour *i.e.*

$$q_n \sim c_1 n^{-(\rho/(\lambda-\mu)+1)}, \tag{20}$$

where

$$c_1 = \frac{\rho}{\lambda}\Big(1 - \frac{\mu}{\lambda}\Big)^{-\rho/(\lambda-\mu)}\Gamma(1+\rho/(\lambda-\mu)). \tag{21}$$

Thus $\tilde{q}_n$ will also exhibit asymptotic power-law behaviour with exponent $-(\rho/(\lambda-\mu)+1)$.

For $\lambda \leq \mu$, it can be shown (Reed and Hughes 2002*a*) that q_n does not exhibit exact power-law behaviour, but rather behaves in a 'stretched exponential' form.

2.2 A special case: no extinctions If the extinction rate parameter μ is set to zero, the birth-and-death process reduces to the Yule process and the resulting size distribution is the eponymous Yule distribution (*e.g.* Johnson *et al.* 1993, Sec. 6.10.3) with p.m.f.

$$q_n = \Big(\frac{\rho}{\lambda}\Big)\frac{\Gamma(\rho/\lambda+1)\Gamma(n)}{\Gamma(\rho/\lambda+n+1)}, \quad \text{for } n = 1, 2, \ldots. \tag{22}$$

This exhibits power-law behaviour in the upper-tail with exponent $\rho/\lambda + 1$ and indeed as Yule showed, exhibits almost linear behaviour in the log-log plot, over the whole range.

3 Power-law behaviour in evolving biological networks We shall consider various models for the evolution of a network, in which new nodes enter stochastically. The networks considered will be what have been termed 'scale-free' networks (see the reviews of Albert and Barabási 2002 and Dorogovtsev and Mendes 2002). The basic idea of such networks is that nodes that are well-connected are more likely to establish new connections than are nodes that are less well connected. The typical example of such a network is the World-Wide Web where, when new sites are added to the web, they are more likely to have links to sites such as Google, Adobe *etc.* that are already well-connected, than to weakly connected sites. Existing models of scale-free networks (see Albert and Barabási 2002) assume that a new node is added to the network in each period and it connects to a fixed number of existing nodes, with the probability of connection to any given existing node being proportional to that node's current connectivity. We shall consider a model of this type later, but first we consider a simple model, which possibly is more appropriate in biological applications, which gives rise to a scale-free network with a tree structure. This model will then be extended to allow for any network structure.

3.1 A model for an evolving tree-structured network Suppose when there are n nodes in the network, with varying connectivities,[3] any node can connect to an external isolated node, thereby bringing it into the network, with well-connected nodes being more likely to do this than less well-connected nodes. Specifically for a node in the network, with connectivity k_i at time time t, suppose that the probability of it bringing a new node into the network in the infinitesimal time interval $(t, t+h]$ is $\lambda k_i h + o(h)$. Thus the well-connected nodes are much more likely to establish a new connection than the less well-connected nodes. Now denote the number of nodes in the network at time t by $N(t)$ and the number connected to a specific node (call it node $*$) by $K(t)$. Let $p_{k,n}(t) = \Pr\{K(t) = k, N(t) = n\}$. Then

$$\begin{aligned} p_{k,n}(t+h) &= p_{k-1,n-1}(t)\lambda(k-1)h + p_{k,n-1}(t)\lambda h\sum_{i\neq *} k_i \\ &\quad + p_{k,n}(t)\Big(1 - \lambda h\sum_{i=1}^{n} k_i\Big) + o(h). \end{aligned} \tag{23}$$

[3] By the 'connectivity' of a node, we mean the number of other nodes that are directly linked to it. In other contexts, this number is called the 'valence', the 'coordination number', or the 'degree'.

When $N(t) = n$, we know that $\sum_{i=1}^{n} k_i = 2n - 2$, since the system evolves from one in which $\sum_{i=1}^{n} k_i = 2$ when $n = 2$, and the addition of each new node to the network increases the sum over connectivities of all nodes by 2. In the limit $h \to 0$, the recurrence relation (23) yields the differential-difference equation

$$\frac{d}{dt}p_{k,n}(t) = \lambda(k-1)p_{k-1,n-1}(t) + \lambda(2n-4-k)p_{k,n-1}(t) - \lambda(2n-2)p_{k,n}(t). \tag{24}$$

Summing this over n (from 1 to infinity) yields a differential-difference equation for the marginal distribution of $K(t)$:

$$\frac{d}{dt}p_k(t) = \lambda(k-1)p_{k-1}(t) - \lambda k p_k(t). \tag{25}$$

This is the equation of a Yule process (Bailey, 1964, p.85) with parameter λ. It follows from standard results that $K(t)$ has a negative binomial distribution with parameters $k_0 = K(0)$ and $e^{-\lambda t}$. In particular with $k_0 = 1$ this reduces to a geometric distribution with

$$p_k(t) = e^{-\lambda t}(1 - e^{-\lambda t})^{k-1}, \qquad \text{for } k = 1, 2, \ldots \tag{26}$$

But the connectivities of nodes do not follow a geometric distribution, the reason being that they will have been present in the network for different lengths of time. To take this factor into account the distribution (26) must be integrated with respect to the distribution of the time t that nodes have been in existence.

One can determine this by first summing (24) from k from 1 infinity, to yield an evolution equation for the marginal distribution of $N(t)$:

$$\frac{d}{dt}p_n(t) = 2\lambda(n-2)p_{n-1}(t) - 2\lambda(n-1)p_n(t). \tag{27}$$

This is the equation of a non-homogeneous birth process with

$$\Pr\{\text{birth in } (t, t+h] \mid N(t) = n)\} = 2\lambda(n-1). \tag{28}$$

From this it follows that the number of new nodes, $U(t) = N(t) - n_0$, connected in $(0, t]$ is a birth process with immigration and has a negative binomial distribution. Such a process is an 'order statistic' process (see *e.g.* Feigin 1979), which means that the times of births since the start of the process have the same joint distribution as those of the order statistics of a sample of independent identically distributed random variables: in this case of random variables with a truncated exponential distribution, that is, with p.d.f.

$$f(t) = \frac{2\lambda e^{-2\lambda t}}{1 - e^{-2\lambda T}} \tag{29}$$

where T is the elapsed time since the founding of the network (see Feigin 1979, p. 300). This means that the time since the introduction of any existing node (say the specified node $*$) will have this distribution. Assuming the time T elapsed since the founding of the network, is large, to a good approximation, the time since the introduction of node $*$ will have an exponential distribution with p.d.f. $f(t) = 2\lambda e^{-2\lambda t}$ on $(0, \infty)$.

It follows that the distribution of the connectivity of node $*$ at the current time, will be given by the state of a Yule process (governed by (25)) after an exponentially distributed time, that is, the state of an a process growing exponetially in expectation, after an exponentially distributed time. It is not difficult to show (by integrating (26) with respect to the exponential density that the p.m.f. of the connectivity K^* of node $*$ is

$$p_k^* = \frac{2\Gamma(3)\Gamma(k)}{\Gamma(k+3)} = \frac{4}{k(k+1)(k+2)}, \tag{30}$$

a Yule distribution which exhibits power-law behaviour with exponent -3. It is interesting to note that unlike the Yule distribution in Sec. 2.2, this distribution does not depend on any parameters. This is one way in which the network can be said to be scale-free.

A second way of formulating a model for the evolution of a tree-structured network, which gives the same result as above, is closer to the more familiar model of the evolution of scale-free networks (Albert and Barabási, 2002). However rather than assuming a fixed number of new nodes is added to the network in every period, assume instead that new nodes are created in a Yule process at rate 2λ, and that any new node will connect to one of the existing nodes, with probabilities proportional to the current connectivities of these existing nodes. After some steps this yields an ode for $p_k(t)$ the same as (25) and in consequence a geometric distribution for $K(t)$ the same as (26). Since the Yule process is an order statistic process it follows that the time elapsed since the introduction of a specified node $*$ will follow a truncated exponential distribution on $(0, T)$, again with parameter 2λ, and thence to a good approximation that the distribution of the current connectivity of node $*$ will follow the Yule distribution (30).

3.2 The evolution of networks where new internal links can be established We now consider a model for the evolution of more complicated networks, in which any node can establish a new connection with an exterior, isolated node, as before, and in addition can establish a new connection with any existing node in the network. Let $N(t)$ and $K(t)$ denote the number of nodes in the network and the connectivity of a specified node $*$, say at time t, as before. In addition let $M(t)$ denote the total number of links in the network, so that

$$K(t) + \sum_{i \neq *} K(t) = 2M(t). \tag{31}$$

A state of the system in which $K(t) = k, M(t) = m, N(t) = n$ will be denoted by (k, m, n) for brevity. In the infinitesimal interval $(t, t + h]$ the following transitions that add one link to the network can occur:

$$\begin{aligned}
(k-1, m-1, n-1) &\rightarrow (k, m, n) && \text{[new link from } * \text{ to external node]} \\
(k, m-1, n-1) &\rightarrow (k, m, n) && \text{[new link from node other than } * \text{ to external node]} \\
(k-1, m-1, n) &\rightarrow (k, m, n) && \text{[new link from } * \text{ to an existing node]} \\
(k, m-1, n) &\rightarrow (k, m, n) && \text{[new internal link not involving } *\text{]}
\end{aligned}$$

As before, assume that in the infinitesimal interval $(t, t+h]$ any existing node i can establish a new external link with probability $\lambda k_i h + o(h)$, so that the probabilities associated with the first two transitions are $\lambda(k-1)h + o(h)$ and

$$\lambda\Big(\sum_{i \neq *} k_i\Big)h + o(h) = \lambda[2(m-1) - k)]h + o(h),$$

respectively. Assume further that any existing node i can establish a new internal link with probability $\nu k_i h + o(h)$, so that for a transition of the third kind the probability is $\nu(k-1)h + o(h)$, and for a transition of the fourth kind the probability is

$$\nu\Big(\sum_{i \neq *} k_i\Big)h + o(h) = \nu[2(m-1) - k]h + o(h).$$

With these assumptions we deduce the evolution equation

$$\begin{aligned}\frac{d}{dt}p_{k,m,n}(t) &= \lambda(k-1)p_{k-1,m-1,n-1}(t) + \lambda[2(m-1)-k]p_{k,m-1,n-1}(t) \\ &\quad +\nu(k-1)p_{k-1,m-1,n}(t) + \nu[2(m-1)-k]p_{k,m-1,n}(t) \\ &\quad -[\lambda k + \lambda(2m-k) + \nu k + \nu(2m-k)]p_{k,m,n}(t). \end{aligned} \tag{32}$$

Summing this over $n = 1, 2, \ldots$ and $m = 1, 2, \ldots$ yields

$$\frac{d}{dt}p_k(t) = (\lambda+\nu)(k-1)p_{k-1}(t) - (\lambda+\nu)kp_k(t), \tag{33}$$

which is the same as (25) with λ replaced by $\lambda + \nu$. Also summing over $k = 1, 2, \ldots$ yields

$$\frac{d}{dt}p_{m,n}(t) = 2\lambda(m-1)p_{m-1,n-1}(t) + 2\nu(m-1)p_{m-1,n}(t) - 2(\lambda+\nu)mp_{m,n}(t). \tag{34}$$

The solution of (33) with $k_0 = 1$ yields a geometric distribution for $K(t)$ as before (equation (26)), but with parameter $e^{-(\lambda+\nu)t}$ rather than $e^{-\lambda t}$.

To solve (34) define the generating function

$$\phi(s,z,t) = \sum_{m,n=1}^{\infty} p_{m,n}(t)s^m z^n. \tag{35}$$

Multiplying (34) by $s^m z^n$ and summing over m and n yields the following partial differential equation for ϕ:

$$\phi_t(s,z,t) = 2s[\lambda sz + \nu s - (\lambda+\nu)]\phi_s(s,z,t). \tag{36}$$

This equation can be solved (with initial condition $M(0) = m_0, N(0) = n_0$) by the method of characteristics to yield

$$\phi(s,z,t) = z^{n_0}\left[\frac{s(\lambda+\nu)}{(\lambda+\nu)e^{2(\lambda+\nu)t} + s(\lambda z+\nu)(1-e^{2(\lambda+\nu)t})}\right]^{m_0}. \tag{37}$$

It follows that both $N(t)$ and $M(t)$ have negative binomial distributions. In particular $U(t) = N(t) - n_0$, the number of new nodes established in $(0,t]$ has a negative binomial distribution with p.m.f.

$$\Pr\{U(t) = u\} = \binom{m_0+u-1}{u}[p(t)]^{m_0}[q(t)]^u, \qquad u = 0, 1, \ldots \tag{38}$$

where

$$p(t) = \frac{\lambda+\nu}{\lambda e^{2(\lambda+\nu)t} + \nu}, \quad q(t) = 1 - p(t) = \frac{\lambda(e^{2(\lambda+\nu)t}-1)}{\lambda e^{2(\lambda+\nu)t} + \nu}. \tag{39}$$

As before it can be shown that $U(t)$ is an order statistic process, in this case with

$$\mathsf{E}(U(t)) = m_0\frac{\lambda}{\lambda+\nu}\left(e^{2(\lambda+\nu)t} - 1\right). \tag{40}$$

From this using the results of Feigin (1979) it follows that conditional on there being $U(t) = u$ new nodes established, the times elapsed since establishment of these nodes have the same

joint distribution as a random sample of size u from the truncated exponential distribution with p.d.f.

$$f(t) = \frac{2(\lambda+\nu)e^{-2(\lambda+\nu)t}}{1-e^{-2(\lambda+\nu)T}}, \tag{41}$$

where T is the elapsed time since the founding of the network. Thus, assuming T large, we can treat the distribution of the time since establishment of node $*$ as being approximately exponential with parameter $2(\lambda+\nu)$.

As in the previous sub-section, the unconditional distribution of the number of connections at node $*$ can be obtained by integrating the geometric distribution

$$p_k(t) = e^{-(\lambda+\nu)t}(1-e^{-(\lambda+\nu)t})^{k-1}, \quad \text{for } k = 1, 2, \ldots \tag{42}$$

with respect to the exponential density with parameter $2(\lambda+\nu)$, yielding the identical Yule distribution (30) with power-law behaviour with exponent -3. As in previous examples, the power-law distribution arises as the state of a stochastic process (for $K(t)$) growing exponentially in expectation (the Yule process) after an exponentially distributed time.

As in Sec. 3.2 the distribution of conectivities is scale-free. The distributions of the size $N(T)$ and total connectivity $M(T)$ however do depend on model parameters. Specifically if the intial network has $n_0 = 2$ nodes with a single ($m_0 = 1$) link, then $N(T)$ has a (shifted) geometric distribution

$$\Pr(N(T) = n) = p(T)[q(T)]^{n-2}, \quad \text{for } n = 2, 3, \ldots, \tag{43}$$

where $p(T)$ and $q(T)$ are given by (39); and $M(T)$ has a geometric distribution with parameter $e^{-2(\lambda+\nu)T}$, *i.e.* with p.m.f. (42) but with $\lambda+\nu$ replaced by $2(\lambda+\nu)$. Thus

$$\mathsf{E}(N(T)) = 2 + \frac{\lambda}{\lambda+\nu}(e^{2(\lambda+\nu)T}-1) \quad \text{and} \quad \mathsf{E}(M(T)) = e^{2(\lambda+\nu)T}. \tag{44}$$

We have obtained similar results for the distribution of connectivities of nodes in the models of Sec. 3.1 and 3.2. The latter model is more general and seems to provide a reasonable description for the growth of a network of sexual partners. It may also be reasonable for food webs and networks of interacting proteins. If so the exponents in the empirical distributions of connectivities should be close to 3. Estimates cited by Albert and Barabási (2002) are 3.4 for a network of sexual partners; 2.4 for the protein network of the yeast *Saccharomyces cerevisiae;* and 1.05 and 1.13 for two food webs. The first two are not too far from the value of 3 derived in the model. Those for the food webs however are very different, but it should be borne in mind that the food webs are rather small, the largest having 186 nodes, so there is probably considerable sampling error in the estimates.

4 Other examples and conclusions In this article we have used stochastic models to explain why certain distributions in biology exhibit power-law behaviour. The explanations are in a sense special cases of a more general result which predicts power-law tails in distributions which result from stopping an exponentially growing process after an exponentially distributed time. Reed and Hughes (2002*b*) give other examples of distributions with power-law tails which can be explained as the result of this mechanism. These include those of incomes and human settlement sizes, which can be modelled as growing following geometric Brownian motion. If the workforce is growing at a fixed rate, or if new settlements are being created from old in a Yule process, then the time in the workforce of earners, or the time in existence of settlements, should be approximately exponentially distributed, and

the resulting distributions should follow power-laws in both tails, as indeed does occur for empirical distributions of incomes and settlement sizes. Another example is that of internet file sizes. The evolution of file sizes can be modelled as following a discrete multiplicative process, and assuming the random creation of new files from old, it can be shown that the time in existence of files should be approximately geomtrically distributed. This leads to a distribution of file sizes exhibiting power-law behaviour in one or both tails. Large data sets on WWW file sizes indicate power-law behaviour does occur at least in the upper tail.

Another example with potential application to biology is that of the distribution of the size of family name clades. It is assumed that the number of males carrying a family name follows a Galton–Watson branching process, and that new names arise either randomly through immigration, or from the random splitting of existing names (*e.g.* through a spelling change). With these assumptions it can be shown (Reed and Hughes 2003) that the current distribution of the size of name clades should exhibit power-law behaviour in the upper tail.[4] Indeed it seems that empirical distributions of names follow close to a power law over their complete range of support. As pointed out in a footnote in the introduction, the mechanism through which family names are passed on from generation to generation is very similar to that for the transmisssion of mtDNA. Mutations in mtDNA could also be considered similar to changes in spelling of names. This would suggest that the sizes of haplogroups with common mtDNA should have a distribution exhibiting power-law behaviour, at least in the upper tail. We are not familiar with data which could verify or refute this prediction. For now we leave it as a prediction and hope that it will be confirmed and our model validated.

The authors have elsewhere considered the distribution of the number of taxon elements that have ever existed, including both currently live and extinct species (Hughes and Reed 2002) in a model of evolution subject to catastrophic extinction events. Power law distributions are again obtained.

References

[1] Albert, R. and Barabási, A.L. (2002), 'Statistical mechanics of complex networks', *Rev. Mod. Phys.* **74**, 47–97; e-print *arXiv/cond-mat/0106096.*

[2] Bailey, N.T.J. (1964), *The Elements of Stochastic Processes*, John Wiley and Sons, New York.

[3] Dorogovtsev, S.N. and Mendes, J.F.F. (2002), 'Evolution of networks', *Adv. Phys.* **51**, 1079–1187; e-print *arXiv/cond-mat/0106144.*

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

[5] Hughes, B.D. and Reed, W.J. (2002), 'A problem in paleobiology', e-print *arXiv/physics/0211090.*

[6] Johnson, N.L., Kotz, S.I. and Kemp, A.W. (1993), *Univariate Discrete Distributions*, John Wiley and Sons, New York.

[7] Reed, W.J. and Hughes, B.D. (2002*a*), 'On the size distribution of live genera', *J. Theoret. Biol.* **217**, 125–135.

[8] Reed, W.J. and Hughes, B.D. (2002*b*), 'From gene families and genera to incomes and internet file sizes: why power-laws are so common in nature', *Phys. Rev. E* **65** (2002) 067103.

[9] Reed, W.J. and Hughes, B.D. (2003), 'On the distribution of family names', *Physica A* **319**, 579–590.

[10] Yule, G.U. (1924), 'A mathematical theory of evolution, based on the conclusions of Dr. J.C. Willis, F.R.S.', *Phil. Trans. Roy. Soc. Lond., Series B*, **213**, 21–87.

[4] The theory predicts a power-law modulated by a log-periodic factor. The oscillations are of very small amplitude, and would be difficult to detect in data.