On a Group Formation Model Leading to the Logarithmic Series Distribution

by

BU-1139-M

August 1991

S.P. Blythe

Dept. Statistics & Modelling Science

Strathclyde University

Glasgow G1 1XH

Scotland

and

K. Dietz

Department of Medical Biometry

Eberhard-Karls-University

W-7400 Tübingen

Germany

ABSTRACT

A deterministic model is presented, motivated by the growth and decline of groups of sexually linked individuals in a closed, finite, population. Group sizes change through the loss and gain of individuals. It is shown that the steady-state relative frequency distribution of group sizes follows a truncated power series distribution, which tends rapidly to the logarithmic series distribution as the population size becomes large. By contrast, most group size models lead to a truncated Poisson or negative binomial distribution.

1. INTRODUCTION

The logarithmic series distribution, described originally by Fisher *et al.* (1943), has been found to be the limiting distribution of many processes (*e.g.: loc. cit.*, Kendall 1948, Nelson and David 1967. See also Johnson and Kotz 1972). A common theme seems to be the existence of groupings of individuals within a population; then the distribution can for example arise through sampling effects.

In this paper we consider a group-size model, which arose from consideration of sexual networks in human populations: sexually active individuals may be considered to belong to social groupings or networks (Klovdähl 1985) where every individual in each group has a connection (often *via* other members) to all of the others. Within a population, there will be very many such groups, distinct and unconnected at any given time, with individuals continually leaving and joining groups, so producing new groups, dissolving old ones, and linking and separating pairs of groups. The motivation is that models for the spread of sexually transmitted diseases in a population composed of such groups of concurrent sexual linkages (Ward and May 1991, Blythe 1991), may be quite different from those where partner change rates per unit time are used (*e.g.* Anderson *et al.* 1987), or those based on pair formation/dissolution (*e.g.* Dietz and Hadeler 1988, Blythe *et al.* 1991). For a review of this literature, see Castillo-Chavez (1989).

There is a considerable literature on group-size distributions (see Okubo 1986 for a review), including the physical clustering of animals and humans. Models in the class with which we are concerned consist of a set of differential equations for the number of groups of all feasible sizes, with (generally quadratic) terms encapsulating the various processes of group growth and decline. Such models are strictly deterministic descriptions, though this is not always recognized by the authors; see Goodman (1964)

and Okubo (1986) for examples and a discussion of this point. The steady-state frequency distribution of group sizes for these deterministic models is typically a truncated Poisson distribution (*op. cit.*).

A significant feature of the sexual grouping problem is that individuals are not usually aware of the size of the groupings which they join, as these exist in some form of "social space." As a consequence, the sizes of the groups do not (necessarily) directly effect the probabilities of joining them: instead, only the fraction of the population in groups of a given size is important.

2. MODEL FORMULATION

The model is an extremely simplified description of the processes of group initiation, growth, and dissolution. We consider a closed population, and construct a deterministic description of the processes controlling the frequency distribution of group sizes. We make the strong simplifying assumption that individuals must leave a group before joining another one; this precludes the linking and unlinking of groups. The model will encapsulate two distinct but inter-related processes. First, there is the flow of individuals: this occurs *from* groups of all sizes, *to* the pool of singles, as individuals leave groups, and *from* the singles pool *to* groups of all sizes, as singles take partners. The second process is the change in the numbers of groups of all sizes: as individuals flow into and out of a given group, the size of that group inevitably changes. Hence, the flow of individuals causes the numbers of groups of different sizes to vary, in a manner we will now specify.

Let $N > 2$ be the total number of individuals in the population, and let g_k ($k=1,2,\ldots$) be the number

of groups of size k. The group-size for single individuals is, of course, one. We describe the process of individuals leaving a group using a parameter η. It is helpful to note that the *total* number of individuals who are in any k-group is $n_k = k g_k$, so that the total loss rate from k-groups is ηn_k individuals per unit time, *i.e.* if this were the only process involving the k-group individuals, their residence time in a k-group would be exponentially distributed with parameter η. All of these individuals leaving the k-groups enter the singles pool, so that g_1 increases, g_k decreases, and g_{k-1} increases, as k-groups become $(k-1)$-groups. Groups of size 2 are a special case: separation occurs in these groups at a rate 2η per group, so that the rate of production of new singles is $4\eta g_2$: a single leaving a pair produces *two* singles (a similar calculation is required in counting the number of daughter cells in binary fission models of unicellular populations, *e.g.* Nisbet and Gurney 1982). When a single leaves a group of size $k+1$ ($k > 2$), g_{k+1} decreases by one, and g_k increases by one, so we need terms of the form $\eta(k+1)g_{k+1} - \eta k g_k$ in each size-k equation, describing *increases* in the number of k-groups due to dissolution of $(k+1)$-groups, and *decreases* in the number of k-groups due to their own dissolution.

We may view the accretion process as one where individuals leave the singles pool, and form partnerships with another individual, who may either be single, or already part of a group of size two or more. A fraction ρ per unit time of singles undergo this process. We need to be able to assign the groups to which individuals leaving the singles pool go. Unlike physical clustering models, for example, the group sizes are not readily detectable by individuals; hence, we do not expect that the rate of accretion will depend simply on k. Instead, we will assume that individuals mix at random; this means that the rate at which a group of a given size $k > 2$ acquires another member is proportional to the fraction of all other people in size-k groups, $k g_k/(N-1)$. This fraction is then also the fraction of new partnerships among singles (ρg_1) which occur with k-group individuals. the number of groups of size k

will increase due to partnerships being added to $(k-1)$-groups, and will decrease due to partnershps being added to k-groups. Hence each $k > 2$ equation will have terms of the form

$$\rho g_1 \frac{(k-1) g_{k-1}}{N-1} - \rho g_1 \frac{k g_k}{N-1}$$

The net effect of joining existing $(k > 1)$ groups on the singles population will be that losses occur at a rate proportional to the fraction of the total population in non-single groups, $(N-g_1) \big/ (N-1)$, while the loss due to pairs of singles forming *new* 2-groups will be proportional to the fraction of other singles in the population, $(g_1-1) \big/ (N-1)$.

We can now write down the deterministic equations for the processes of group formation, accretion, and dissolution, for all group sizes:

$$\begin{cases} \dot{g}_1 = 4\eta g_2 + \eta \sum_{i=3}^{\infty} i\, g_i - \rho g_1 \frac{N-g_1}{N-1} - 2\rho g_1 \frac{g_1-1}{N-1} \\[4mm] \dot{g}_2 = \rho g_1 \frac{g_1-1}{N-1} + 3\eta g_3 - 2\eta g_2 - \rho g_1 \frac{2 g_1}{N-1} \\[4mm] \dot{g}_k = \rho g_1 \frac{(k-1) g_{k-1}}{N-1} + \eta(k+1) g_{k+1} - \eta k g_k - \rho g_1 \frac{k g_k}{N-1} \quad , \; k \geq 3, \end{cases} \qquad (1)$$

for appropriate initial conditions $g_k(0) = \psi_k \geq 0$. Here ($\dot{}$) denotes differentiation with respect to time t. The value of k in (1) is permitted to run to infinity only for convenience; in practice we must truncate $\{g_k\}$ at some finite value, which will be at most N. Provided the tail of $\{g_k\}$ is sufficiently

small for $k > N$, we may conveniently disregard this truncation; in what follows, we take particular account of the effects of this act of convenience.

Eqs (1) sspecify a modified member of the very general class of quadratic models described by Okubo (1986). Because the final term in the $\dot{g}_1$ equation has g_1-1 as a numerator, reflecting the reduction by one of single partners available *to* a single. In Okubo's notation, $a(\ell,k;t)$ is the rate of amalgamation for groups of sizes k and ℓ in, and $b(k,\ell;t)$ is the rate of splitting of k-groups into groups of size ℓ and $k-\ell$, so that (1) may be characterized by

$$a(\ell,k) = \begin{cases} 2\rho \dfrac{1}{g_1} \dfrac{g_1-1}{N-1} & , \ell = 1, k = 1 \\[2mm] \rho \dfrac{\ell g_\ell}{N-1} & , \ell \geq 2, k = 1 \\[2mm] 0 & , \ell \geq 1, k \neq 1 \end{cases}$$

and

$$b(k,\ell) = \begin{cases} 4\eta & , \ell = 1, k = 2 \\[2mm] \ell\eta & , \ell = 1, k \geq 3 \\[2mm] 0 & , \ell \geq 1, k \geq 1 \end{cases} .$$

Eqs (1) are similar in form to many stochastic models, based upon continuous-time Markov processes, for birth and death processes (*c.f.* Nisbet and Gurney 1982), epidemic models (Bailey 1963), and models of macroparasitic infection (*e.g.* Anderson and May 1985, Kretzschmar 1989*a,b*). In the group-size examples discussed by Okubo (1986), the equilibrium distribution of group sizes follows a

truncated Poisson, negative binomial, or geometric form. This is also true for the models of Bartholomew (1982), James (1953), and Shorrocks (1962). As we show below, this is not the case with (1). We note in passing that (1), with an arbitrary truncation point, is a rather numerically unstable system, requiring adaptive time-stepping with generally very small steps if numerical error is to be avoided.

3. ASYMPTOTIC GROUP-SIZE DISTRIBUTION

We introduce the following notation: $G \equiv \sum_{k=1}^{\infty} g_k$ is the total number of groups, so $f_k \equiv g_k/G$ is the fraction of groups of size k, clearly subject to $1 = \sum_{k=1}^{\infty} f_k$. Then the average group size is given directly by

$$\mu \equiv \sum_{k=1}^{\infty} k f_k = \frac{N}{G}.$$

Note that because the population is closed, N is a constant, and it may readily be shown that

$$\dot{G} = \frac{d}{dt} \sum_{k=1}^{\infty} g_1 = \eta N - (\eta + \rho) g_1,$$

i.e.

$$\frac{\dot{G}}{G} = \eta \mu - (\eta + \rho) f_1,$$

and hence asymptotically, we have $f_1 = \eta \mu / (\eta + \rho)$.

We consider the asymptotic $(t \to \infty)$ behavior of (1), with $\{f_k^*\}$ denoting the steady-state values of the relative distribution of group sizes. Let the probability generating function of $\{f_k^*\}$, and its first derivative, be defined in the usual way as

$$\phi(X) \equiv \sum_{k=1}^{\infty} f_k^* X^k, \quad \text{and} \quad \phi'(X) \equiv \sum_{k=1}^{\infty} k\, f_k^* X^{k-1}, \quad X \in [0,1],$$

respectively. In particular, note that we require $\phi'(1) = \mu = N/G$, $\phi(1) = 1$, and $\phi(0) = 0$.

At the steady-state of (1), with time $t \to \infty$, and the $\{g_k\}$ going to their asymptotic values, the system reduces to a set of algebraic equations which we can rearrange to give the following equation for $\phi'(X)$:

$$\phi'(X) = \frac{\eta f_1^* + \left[\dfrac{\rho f_1^*(N-2)}{N-1} - \eta \dfrac{N}{G}\right]X + \dfrac{\rho f_1^*}{N-1}X^2}{(1-X)\left(\eta - \dfrac{\rho f_1^* G}{N-1}X\right)}, \quad X \in [0,1].$$

$$= \frac{N}{G}\, \frac{s + \left[\dfrac{(1-s)(N-2)}{(N-1)} - 1\right]X + \dfrac{(1-s)}{(N-1)}X^2}{(1-X)\left(1 - \dfrac{(1-s)N}{(N-1)}X\right)}, \tag{2}$$

using $f_1^* = sN/G$, $s \equiv \eta/(\eta+\rho)$, implied by the asymptotic result $g_1^* = \eta N/(\eta+\rho)$. Inspection of the terms in (2) reveals that, provided $sN > 1$, the numerator is positive for $X \in [0,1)$, and the second term in the denominator is positive for $X \in [0,1]$. Further, routine application of l'Hôpital's rule gives

$$\lim_{X \to 1} \left\{\phi'(X)\right\} = \frac{N}{G},$$

again provided $sN > 1$. Hence,

$$\phi(X) = \int_0^X \phi'(u)\,du$$

$$= \frac{1}{G}\left\{X - \frac{(Ns-1)(N-1)}{N(1-s)}\ln\left[1 - \frac{N(1-s)}{(N-1)}X\right]\right\}.$$

As $\phi(1) = 1$, this implies

$$G = 1 - \frac{(Ns-1)(N-1)}{N(1-s)}\ln\left[\frac{Ns-1}{N-1}\right]$$

(where $sN > 1$ is again required). Thus the p.g.f. is of the form $\phi(X) = \lambda(\theta X)/\lambda(\theta)$ (with λ such that $\lambda(\theta) = \theta G$, and $\theta \equiv N(1\text{-}s)/(N\text{-}1)$), so $\{f_k^*\}$ belongs to the class of power series distributions (Johnson and Kotz 1972).

The p.g.f. of the tail of the distribution $(X > T)$ is (Bailey 1962)

$$\omega(T) = \frac{1 - \phi(T)}{1 - T}$$

$$= \frac{1}{G}\left[1 + \frac{(Ns-1)(N-1)}{N(1-s)(1-T)}\ln\left[\frac{1 - \frac{N(1-s)}{(N-1)}T}{1 - \frac{N(1-s)}{(N-1)}}\right]\right],$$

and in the series expansion

$$\omega(T) = 1 + \alpha_1 T + \alpha_2 T^2 + \cdots$$

it may readily be demonstrated that the weight in $k > K$ is

$$\alpha_K \equiv \frac{\frac{(Ns-1)(N-1)}{N(1-s)}\left\{\ln\left[\frac{Ns-1}{N-1}\right]+\beta_K\right\}}{\frac{(Ns-1)(N-1)}{N(1-s)}\ln\left[\frac{Ns-1}{N-1}\right]-1},$$

(3)

where

$$\beta_K \equiv \sum_{j=1}^{K}\frac{1}{j}\left(\frac{N}{N-1}(1-s)\right)^j.$$

The Nth term, α_N, gives the region which should be truncated by the finiteness of N (and hence is forbidden for group sizes unless $g_k = 0$ is explicitly enforced for $k > N$ in Eqs (1)). As $N \to \infty$ the sum β_N tends to the limiting value of

$$-\ln\left[\frac{Ns-1}{N-1}\right] > 0 \quad \text{for } s \in (0,1),$$

so that $\alpha_N \to 0$, as required. For finite N, we may readily calculate the value $\nu(\epsilon)$ such that for $N > \nu(\epsilon)$, we have $\alpha_\nu < \epsilon$ (ϵ some suitably small number). Table I shows $\nu(\epsilon)$ for $\epsilon = 10^{-5}$ and 10^{-6}, for various values of $s \in (0,1)$. *Unless s is very small, there is a negligible error from not truncating explicitly above $k = N$, for any reasonably sized model population.* (Note: the $\nu(\epsilon)$ were calculated exactly, using Mathematica$^{\text{TM}}$, but accurate approximations to the sum β_N may be found in Johnson and Kotz (1972)).

We may now exploit the relationship $f^*_{k+1} = \alpha_k - \alpha_{k+1}$, $k > 1$, to obtain the p.d.f. of the group-size distribution $\{f^*_k\}$ from Eq (3); this is in the form of a power series distribution, truncated at zero:

$$f^*_k = \frac{a_k \, \theta^k}{V(\theta)} \, , \quad k = 1,2,3,\ldots$$

(4)

with parameter

$$\theta = \left(\frac{N}{N-1}\right)(1-s),$$

and where

$$a_k = \begin{cases} \dfrac{Ns}{Ns-1} & , \quad k=1 \\[2ex] \dfrac{1}{k} & , \quad k>1 \end{cases}$$

and

$$V(\theta) = \frac{\theta}{(Ns-1)} - \ln[1-\theta].$$

V and a_1 are finite and positive provided $Ns > 1$. Note that $f_1^* = sN/G$ is maintained in Eq (4). The mean and variance are respectively $\mu = N/G$ and

$$\sigma^2 = \frac{N}{G}\left[\left(\frac{N+s-2}{Ns-1}\right)-\left(\frac{N}{G}\right)\right].$$

As $N \to \infty$, Eq (4) reduces to the p.d.f. for the logarithmic series distribution,

$$f_k^* = \frac{1}{-\ln[s]}\frac{(1-s)^k}{k}, \quad k = 1,2,3\ldots$$

with mean and variance

$$\mu = \frac{(1-s)}{-s\,\ln[s]},$$

and

$$\sigma^2 = \mu\left(\frac{1}{s}-\mu\right) = \frac{(1-s)}{-s^2\ln[s]}\left[1+\frac{(1-s)}{\ln[s]}\right],$$

respectively. Thus in an infinite population, the frequency distribution of group sizes follows the logar-

ithmic series distribution. Both the mean and the variance tend towards infinity as $s \to 0$, so the requirement for large $\nu(\epsilon)$ when s is small (the finite N cases of Table I) is at once explained. Likewise, as $s \to 1$, $\mu \to 1$ and the variance tends to zero, so $\nu(\epsilon)$ becomes very small.

4. DISCUSSION

The model (1) is a somewhat crude approximation to the processes which govern the growth and decline of sexual networks (which are essentially invisible to those participating). Age-structure, two sexes, inter-group preference and heterogeneity, and variable population size, are some of the notable features of real networks which are not included. Nonetheless, the model produces a testable hypothetical structure for group-size distributions, which can provide the basis for further model development when informed by appropriate data.

The requirement that $f_1^* = sN/G$ is suggestive, stating that the number of singles in a population, g_1, is simply proportional to the total population size N, and that the constant of proportionality is the key parameter s. This in particular should be testable. There are clearly problems in estimating N (all the contacts of individuals), but capture-recapture techniques have recently been employed successfully to estimate precisely this quantity (Rubin *et. al* 1991) for a college population who have significant sexual mixing with "outsiders." Of all the group size data requirements, the fraction of singles is certainly one of the least difficult to estimate, as the number and nature of contacts need not be addressed (it is however true that sampling is biased in favor of those who spend longer times between partnerships). We can therefore speculate that s might be estimable directly from data on singles and

population sizes, hence giving an estimate of the mean group size in a population, $\mu = N/G$. As there are many studies on pair-durations (*e.g.* marriage durations), this would seem an additional useful data source. (We may also note that $f_1^* \propto \mu$ arises in a variety of different group-size models, when there is a requirement for conservation of the number of groups (Goodman 1964, Cohen 1971, Okubo 1986)).

Furthermore, because (1) in principle provides trajectories of all the group-size variations with time, it may be possible to greatly improve the processes of estimation and model testing, if any time-series data on network change in a population can be obtained.

There are of course many *caveats*. The formation and separation parameters ρ and η are unlikely to be the same for pairs, for example, as for other groups, as "pairs" covers a variety of social forms. The model assumption that groups cannot coalesce may be a severe restriction: in reality there may well be link individuals who are members of multiple groups that would otherwise not be connected (*e.g.* a college student with home-town and college contacts, or individuals who are prostitutes' clients as well as having dating-based sexual contacts). These problems again constitute testable hypotheses, and may be addressed by more detailed modeling, informed by available data and survey results.

The appearance of the logarithmic series distribution in the steady-state relative frequency distribution of group-sizes is a novel feature of the model, as truncated Poisson or negative binomial models are more commonly found in group-size problems. This is probably ascribable to the fact that individuals joining a group do so at a rate proportional to the number in that group, rather than to the proportion of groups of that size. An intriguing possibility for future work is the inclusion in models of measures of certainty in the group-size of prospective partners. A population model where the logarithmic distribution appears may be found in Kendall (1948), for the case of an unstructured population

-15-

founded by a single parthenogenetic ancestor; here the distribution is of population size, rather than group size.

ACKNOWLEDGEMENTS

REFERENCES

Anderson, R.M. and May, R.M. (1985) Helminth infections of humans: mathematical models, population dynamics and control. *Adv. Parasit.* 24: 1-101.

Anderson, R.M., Medley, G.F., May, R.M. and Johnson, A.M. (1987) A preliminary study of the transmission dynamics of the human immunodeficiency virus (HIV), the causative agent of AIDS. *IMA J. Math. Appl. Med. &. Biol.* 3, 229-263

Bailey, N.T.J. (1962, Rep. 1990) "The Elements of Stochastic Processes; with applications in the natural sciences." *Wiley Interscience*, New York

Bartholomew, D.J. (1982, 3rd Ed.) "Stochastic Models for Social Processes." *J. Wiley*, New York

Blythe, S.P. (1991) Multiple concurrent sexual partners, and sexually transmitted disease dynamics. (Submitted to *Theor. Pop. Biol.*)

Blythe, S.P., Castillo-Chavez, C., Palmer, J. and Cheng, M. (1991) Towards a unified theory of mixing and pair formation. *Math. Biosci.* (to appear)

Castillo-Chavez, C. (1989) Review of recent models of HIV/AIDS transmission. *pp.* 253-262 *in* S.A. Levin, T.G. Hallam, and L.J. Gross (Eds), "Applied Mathematical Ecology", *Biomathematics* 18, Springer-Verlag, Berlin *etc.*

Cohen, J.E. (1971) "Causal Groups of Monkeys and Men: Stochastic Models of Elemental Social Systems." *Harvard U. P.*, Cambridge, Mass.

Dietz, K. and Hadeler, K.P. (1988) Epidemiological models for sexually transmitted diseases. *J. Math. Biol.* **26**, 1-25.

Fisher, R.A., Corbet, A.S., and Williams, C.B. (1943) The relation between the number of species and the number of individuals in a random sample of an animal population. *J. anim. Ecol.* **12**, 42-58

Goodman, L.A. (1964) Mathematical methods for the study of systems of groups. *Amer. J. Sociol.* **70**, 170-192

James, J. (1953) The distribution of free-forming small group size. *Amer. Sociol. Rev.* **18**, 569-570

Johnson, N.L. and Kotz, S. (1972) "Discrete Distributions" *Wiley Interscience*, New York

Kendall, D.G. (1948) On some modes of population growth leading to R.A. Fisher's logarithmic series distribution. *Biometrika* **35**, 6-15

Klovdähl, A.S. (1985) Social networks and the spread of infectious diseases: the AIDS example. *Soc. Sci. Med.* **21**, 1203-1216

Kretzschmar, M. (1989a) A renewal equation with birth-death process as a model for parasitic infections. *J. Math. Biol.* **27**: 191-221.

Kretzschmar, M. (1989b) Persistent solutions in a model for parasitic infections. *J. Math. Biol.* 27: 549-573.

Nelson, W.C. and David, H.A. (1967) The logarithmic distribution: a review. *Virginia J.Science* 18, 95-102

Nisbet, R.M. and Gurney, W.S.C. (1982) "Modelling Fluctuating Populations," *Wiley Interscience*, New York

Okubo, A. (1986) Dynamical aspects of animal grouping: swarms, school, flocks, and herds. *Adv. Biophysics* 22, 1-94

Rubin, D., Umbach, D. Shyu, S.-F., and Castillo-Chavez, C. (1991) Using capture-recapture methodology to estimate the size of the population at risk for AIDS and other sexually transmitted diseases (Biometrics Unit Technical Report BU-1116-M).

Shorrocks, A.F. (1975) On stochastic models of size distributions. *Rev. Econ. Stud.* 42, 631-641

Ward, C.H. and May, R.M. (1991) The influence of concurrent partnerships on the dynamics of HIV/AIDS. (Submitted to *Math. Biosci.*)

Table I

'Population size $N = \nu(\epsilon)$ such that for group-size $k > \nu(\epsilon)$ the residual tail weight α_ν is less than ϵ

	$\nu(\epsilon)$	
s	$\epsilon = 10^{-5}$	$\epsilon = 10^{-6}$
0.01	873	1081
0.1	89	109
0.25	34	41
0.5	15	18
0.75	8	10
0.9	5	6
0.99	5	3

Legend

For $N \geq \nu(\epsilon)$, and the given value of parameter $s = \eta/(\eta+\rho)$, the error in using Eqs (1) without explicitly forcing $g_k = 0$ for $k > N$, is less than ϵ.