Available online at www.sciencedirect.com

SCIENCE @ DIRECT®

Theoretical Population Biology 65 (2004) 413–422

Theoretical Population Biology

http://www.elsevier.com/locate/ytpbi

Likelihood-based inference for stochastic models of sexual network formation☆

Mark S. Handcock[a,]* and James Holland Jones[b]

[a] *Center for Statistics and the Social Sciences, University of Washington, Box 354320, Seattle, WA 98195-4320, USA*
[b] *Department of Anthropological Sciences, Stanford University, Stanford, CA 94305-2117, USA*

Received 28 May 2003

Abstract

Sexually-transmitted diseases (STDs) constitute a major public health concern. Mathematical models for the transmission dynamics of STDs indicate that heterogeneity in sexual activity level allow them to persist even when the typical behavior of the population would not support endemicity. This insight focuses attention on the distribution of sexual activity level in a population. In this paper, we develop several stochastic process models for the formation of sexual partnership networks. Using likelihood-based model selection procedures, we assess the fit of the different models to three large distributions of sexual partner counts: (1) Rakai, Uganda, (2) Sweden, and (3) the USA. Five of the six single-sex networks were fit best by the negative binomial model. The American women's network was best fit by a power-law model, the Yule. For most networks, several competing models fit approximately equally well. These results suggest three conclusions: (1) no single unitary process clearly underlies the formation of these sexual networks, (2) behavioral heterogeneity plays an essential role in network structure, (3) substantial model uncertainty exists for sexual network degree distributions. Behavioral research focused on the mechanisms of partnership formation will play an essential role in specifying the best model for empirical degree distributions. We discuss the limitations of inferences from such data, and the utility of degree-based epidemiological models more generally.

Keywords: Stochastic models; Sexual networks; Sexually-transmitted diseases; HIV/AIDS; Multi-model inference

1. Introduction

Sexually-transmitted diseases (STD), including HIV/AIDS, constitute a major global public health concern. UNAIDS estimates that there were 40 million adults and children living with HIV/AIDS in 2001. In addition to being a major humanitarian calamity, the AIDS pandemic represents a substantial barrier to economic development in many resource-poor settings throughout the world.

STDs other than HIV have been labeled a "hidden epidemic" by the National Institute of Medicine (1997), who estimate that the annual economic cost of STDs other than HIV/AIDS in the United States alone to be \$16.4 billion. While the prevalence of some of the traditional bacterial STDs such as Gonorrhea (*Neisseria gonorrhoeae*) and Syphilis (*Treponema pallidum*) have undergone steady decline over the last 20 years, others, such as Chlamydia (*Chlamydia trachomatis*) have increased in prevalence (Centers for Disease Control and Prevention, 2001). Furthermore, some STDs on the verge of elimination have made dramatic reversals, re-establishing themselves as endemic infections (Williams et al., 1999).

The control and eventual eradication of STDs is an important public health goal. Both mathematical and statistical models of infectious disease processes have proven to be invaluable tools for infectious disease epidemiology (Anderson and Garnett, 2000; Foulkes, 1998). However, developing useful models for STDs presents a number of challenges. Prominent among these is characterizing population heterogeneity in sexual behavior. The average behavior of most populations is not sufficient either to allow an epidemic or maintain an endemic STD infection. Mathematical formalizations of STD infection dynamics indicate that

☆Research supported by Grant R01-DA012831 from NIDA and grants R01-HD034957 and R01-HD41877 from NICHD.

*Corresponding author. Fax: +1-360-365-6324.
E-mail addresses: handcock@stat.washington.edu (M.S. Handcock), jhj1@stanford.edu (J.H. Jones).

0040-5809/\$ - see front matter © 2004 Elsevier Inc. All rights reserved.
doi:10.1016/j.tpb.2003.09.006

heterogeneity in sexual behavior allows STDs, which would otherwise fade out given average behavior, to persist (Hethcote and Yorke, 1984; Anderson and May, 1991; Jones and Handcock, 2003a). The importance of behavioral heterogeneity has focused attention on the properties of the distribution of sexual partner number. In this paper, we will present an analysis of the statistical properties of empirical sexual partnership distributions. These distributions play a key role in the mathematical theory of STD transmission, and may hold the key to their control (Morris, 1991, 1995).

1.1. *Mathematical epidemiology of STDs*

In the standard theory of infectious disease transmission dynamics (Bailey, 1975; Anderson and May, 1991), the force of infection is an increasing function of the size of the population, and populations will exhibit threshold sizes below which an epidemic is impossible. These behaviors are a consequence of the traditional assumption of mass action. However, Anderson and May (1991) note that there is no reason to assume that the number of intimate contacts will increase with increasing population size.

In addition to threshold population sizes for epidemics, the dynamics of an epidemic are governed by a threshold parameter, R_0, the basic reproductive number. R_0 represents the expected number of secondary cases produced by a single index case in a population of susceptibles. In the case of an unstructured population, R_0 is simply the product of three quantities: (1) the transmissibility of the pathogen, (2) the duration of infectiousness, and (3) the contact rate between susceptible and infectious individuals. For more complexly structured models, the calculation of R_0 can be generalized in a fairly straightforward manner (Diekmann et al., 1990), though the interpretation becomes more difficult.

Behavioral heterogeneity has been incorporated into the formulation of R_0 by Anderson et al. (1986). Assuming random mixing with respect to the degree distribution in a population structured by sexual activity, they show that R_0 increases linearly with the variance of the degree distribution of the population sexual contact graph.

Since surveys of sexual behavior reveal that the great majority of people have one partner or fewer in the last year, (Laumann et al., 1994; Lewin, 1996) the driving factor for STDs is clearly the tail of the degree distribution, and this is where the emphasis for inference typically focuses (May and Lloyd, 2001; Liljeros et al., 2001).

1.2. *Social networks*

By definition, socially communicable diseases are transmitted from person to person. An intuitive and mathematically convenient means of representing social contacts is a graph (Wasserman and Faust, 1994). The nodes of the graph represent individual people and the edges represent contact. The number of edges adjacent to a particular node is its degree, and the collection of nodal degrees is the degree distribution of the population.

Graph-theoretic network models have been used to describe a wide variety of relational data (Borgatti and Everett, 1992) including friendship networks among children (Moody, 2001), scientific collaboration networks (Newman, 2001), social and economic exchange networks (Bearman, 1997), and contact networks for the spread of infectious disease (Morris, 1993a, 2004).

1.2.1. *Models for degree distributions*

Empirical degree distributions for sexual partnership networks are highly skewed (Jones and Handcock, 2003a). The modal yearly degree is $k = 1$ for nearly all large representative surveys (e.g., Laumann et al., 1994; Lewin, 1996; Hubert et al., 1998; Aral, 1999; Youm and Laumann, 2002). The tremendous skew of sexual degree distributions has, through analogy to a variety of physical systems, suggested the possibility of power-law scaling (Lloyd and May, 2001). Networks exhibiting power-law scaling have been referred to as "scale-free" networks in Amaral et al. (2000) and subsequent publications. This attribution is associated with properties of the implicit underlying stochastic mechanism, and is often used loosely.

Let K be the degree of a randomly sampled person from the population. Recent empirical work (Amaral et al., 2000; Liljeros et al., 2001) has claimed that some sexual network degree distributions have a probability mass function (PMF) for network degree of the form, $P(K = k) \approx k^{-\rho}, k \gg 1$, where $P(K = k)$ is the probability of observing exact degree k and ρ is referred to as a scaling parameter. Let f and g be two functions with support the whole numbers. We take $f(k) \asymp g(k)$ to mean that there exist constants c_1, c_2 such that $0 \leqslant c_1 < f(k)/g(k) \leqslant c_2 < \infty$ for $k = 1, \ldots$. We then say that $P(K = k)$ has *power-law behavior* if $P(K = k) \asymp k^{-\rho}$.

Inference on the scaling parameter ρ of a power-law model typically involves fitting a regression line through the apparently linear region of a plot of the survival function of the degree distribution plotted against the distribution on double logarithmic axes (Amaral et al., 2000; Liljeros et al., 2001). The measurement of uncertainty is then taken as the standard error of the estimated slope. This methodology is inappropriate for the inference problem, yielding (1) biased estimates of the scaling parameter, and (2) greatly underestimated model uncertainty (Jones and Handcock 2003a, b).

Even for very large surveys (such as NHSLS in the USA (Laumann et al., 1994)), the number of observations in the tail of the distribution is very small. The information contained in the tail of a degree distribution is therefore low. Consequently, the precision of inferences on the tail is extremely low.

The degree distribution, of course, was generated by the behavior of the individual actors. The distribution at any time point is probably best thought of as representing a dynamic equilibrium of underlying social and epidemiological processes, rather than simply a static pattern of sexual behavior (Kendall, 1961). Specifying a plausible stochastic process by which the observed data can have been generated provides a better strategy for investigating the properties of sexual networks than considering mathematical distributions with weak proximate mechanisms. The equilibria of stochastic models of network formation can be fit to empirical data using likelihood techniques, allowing both the estimation of parameters and the assessment of goodness-of-fit to the data.

1.3. Stochastic models for network formation

We will discuss three general classes of stochastic process model for sexual network formation: (1) non-homogenous Poisson, (2) preferential attachment, and (3) "vetting" models.

Non-homogeneous Poisson models: Consider the population of individuals with at least one partner in a given time period. Suppose that the number of additional partners $K-1$ that the person has in the time period follows a Poisson distribution with expected value λ. There are many proximate mechanisms for this (for example, the partners are accumulated at a constant rate in time).

The assumption that all individuals have identical propensities to form partnerships is unrealistic. Individuals differ by gender, age, marital status, attractiveness, and other fundamental characteristics that greatly influence partnership formation. To model within-population heterogeneity, we can represent the individual expected values λ as independent draws from a distribution $P(\lambda)$.

There are myriad reasonable models for $P(\lambda)$. One flexible choice is the Gamma distribution. A Poisson distribution with Gamma-distributed rate parameter λ is a classical hierarchical model, with marginal distribution of negative binomial (Johnson et al., 1992).

We employ a model in which $\lambda_i + 1$ is the expected number of sexual partners of the ith individual, and refer to the resulting partner distribution the *shifted negative binomial distribution.* The shifted negative binomial distribution can have quite a long left tail. Nonetheless, the variance is finite. The negative binomial distribution is flexible enough to model a variety of shapes of degree distributions and is widely used in ecology and epidemiology (Martin and Katti, 1965; Alexander et al., 2000).

Preferential attachment models: In preferential attachment models, the probability that a contact is made with any particular individual is a function of that individual's current degree. Two models for preferential attachment are (1) the Yule distribution (Simon, 1955; Jones and Handcock, 2003a) and (2) the Waring distribution (Irwin, 1963).

The underlying stochastic model motivating the partnership distributions under the Yule begins with a network of r connections. Assume that (1) there is a constant probability $(\rho-2)/(\rho-1)$ that the $r+1$st partnership in the population will be initiated from a randomly chosen person to a previously sexually inactive person, and (2) otherwise the probability that the $r+1$st partnership will be to a person with exactly k partners is proportional to $kf(k|r)$, where $f(k|r)$ is the frequency of nodes with exactly k connections out of the r total links in the population. Simon (1955) called the limiting partnership distribution of this process the *Yule distribution*, following the pioneering work of Yule (1925).

Often a measure of passing time is associated with the growth model. In some variants a partnership is associated with each time increment, and in others a person is associated with each time increment. However, these formulations are equivalent in their essential representation of the network.

This stochastic process has been rediscovered, apparently without awareness, by several research groups (Barabási and Albert, 1999; Albert and Barabási, 2000; Dorogovtsev et al., 2000) and been used to characterize Internet growth. The special case with $\rho = 3$ has been proposed by (Barabási and Albert, 1999). It is also implicit in the analysis of the scaling properties of sexual contact networks (Liljeros et al., 2001). It has been described as the "rich get richer" model, as people in the network with many partners tend to accumulate partners faster than those with less. The resulting distribution under this model has the desired property that most people have very few partners, while a very few have many sexual partners. The PMF of the Yule distribution (Johnson et al., 1992) is

$$P(K=k) = \frac{(\rho-1)\Gamma(k)\Gamma(\rho)}{\Gamma(k+\rho)}, \quad k = 1,2,\ldots, \tag{1}$$

where $\Gamma(\rho)$ is the Gamma function of ρ. The Yule distribution has power-law behavior as $P(K=k) \asymp k^{-\rho}$.

The *Waring distribution* is a natural generalization of the Yule proposed by Irwin (1963). The motivating stochastic process is identical to that of Simon, with the exception that the probability that the $r+1$st partnership in the population will be initiated from a randomly chosen person to a previously sexually inactive person is $\frac{\rho-2}{\rho+\alpha-1}$. This model allows for the probability of *non-preferential* partnerships to be a separate parameter from that governing the preferential process. As $\alpha \downarrow -1$ the probability that the tie is to a previously sexually inactive person approaches unity. As $\alpha \to \infty$ the probability of approaches zero. The limiting distribution is given by

$$P(K=k) = \frac{(\rho-1)\Gamma(\rho+\alpha)}{\Gamma(\alpha+1)} \cdot \frac{\Gamma(k+\alpha)}{\Gamma(k+\alpha+\rho)}, \quad \alpha > -1. \tag{2}$$

The Waring distribution has apparently been derived independently by Levene et al. (2002) in the context of modeling growth of the Internet.

Vetting models: The general idea underlying the vetting models is that people form sexual partnerships based on a two-stage process. First, they generate an acquaintance list from which they, second, choose their sexual partners. This class of model is extremely flexible in that practically any probability distribution can be specified for both of these processes. This process focuses attention on the stopping rules that people employ when forming sexual partnerships.

The Yule-vetting models are generalizations of the Yule distribution that recognize that the formation of sexual partnerships is not cost-less. Suppose that sexual partners are chosen from a pool of acquaintances. First, individuals form a random number A acquaintances from a PMF $P(A=a)$. In many situations the process of acquaintance formation may be a relatively cost-less process and $P(A=a)$ may have power-law behavior. This process may represent social networking, geographic, or other processes. Second, suppose the potential number L of sex partners an individual has in the time period follows a distribution that is typically short-tailed. For example, it could be a geometric or negative binomial distribution. There are many proximate mechanisms for this (e.g., the partners are from a queue with constant rate). However, the actual number of partners K is bounded by the number of acquaintances the person has, in the sense that it cannot exceed that number: $K = \min(A, L)$.

The resulting distribution will resemble $P(L=k)$ (e.g., short-tailed) when L is stochastically much smaller than A. It will resemble $P(A=a)$ when L is stochastically much larger than A. These two situations correspond to relatively large and small acquaintance networks, respectively.

As before, the assumption that all individuals have identical propensities to form sex partnerships is unrealistic due to the individual characteristics such as gender, age, marital status, attractiveness. We can model this within-population heterogeneity, by independently drawing the individual expected values from a distribution $P(\lambda)$.

A distribution that may prove useful for generating long-tailed acquaintance lists is the *discrete Pareto* distribution (Johnson et al., 1992). Similarly, the acquaintance list distribution can be modeled as Yule, Waring, or negative binomial.

Vetting models represent many other degree distributions as special cases. On particular case of interest lies in the intermediate parameter range of models with power-law list generators in which it can be shown that $P(K=k) \asymp e^{-k/\kappa} k^{-\alpha}$, where $\kappa = 1/\log(1+1/\tau\mu)$. This is the power-exponential distribution used by Newman (2001) to model scientific collaboration networks and is frequently referred to as the *truncated power-law* model.

1.4. Differential tail behavior

It is probable that the behavior that governs the acquisition (and maintenance) of the first sexual partner a person has in a given time interval will differ from the process leading to the acquisition of further partners. For example, the process leading to the choice of marriage partner is likely to differ from the process leading to the extra-marital affairs.

To account for the possibility of substantial differences between the process at low and high network degree, we allowed for the inclusion of extra parameters to fit the low-degree observations (e.g., $k = 1, 2$). We evaluated the improvement of fit effected by the inclusion of such parameters in the model selection stage (see Section 2.2.1). Network data collected on sexual activity of ever-active people for a relatively short interval, such as a single year, are likely to contain a fraction of people for whom $k=0$. While it is traditional to discard these values before performing inference on the scaling behavior of the degree distribution (Amaral et al., 2000; Liljeros et al., 2001), the number of zeros in a population holds a tremendous amount of information about both the distribution and the process that generated it. All models that we fit contain a parameter for the frequency of degree $k=0$.

2. Methods

2.1. Data

There are a variety of forms of network data (Morris, 1997). Local network data result from collecting information on the number and attributes of a focal individual's sexual partners ascertained using an epidemiological/sociological survey instrument. For our analysis, we use local network data gathered from men and women as a part of three large, representative surveys of sexual behavior. For all surveys, we used the reported number of heterosexual partners in the last year as the estimate of individual network degree.

Table 1
Descriptive statistics for the three sexual history surveys

Survey	Sex	n	Mean	Variance	HIV prevalence (%)
Rakai	Women	803	0.89	0.27	16
	Men	621	1.28	1.23	
Sweden	Women	1335	1.01	0.88	0.08
	Men	1476	1.27	2.19	
USA	Women	1919	1.09	0.69	0.6
	Men	1506	1.41	1.42	

Means and variances are for the number of reported sexual partners in the last year. HIV prevalence data are for the year 2000 and come from Sewankambo et al. (2000) for Rakai and UNAIDS country fact sheets for Sweden and the United States.

Rakai project sexual network survey: The Rakai district is an administrative unit of southern Uganda with a mature AIDS epidemic and an HIV/AIDS prevalence of approximately 16%. As with all of Sub-Saharan Africa, the primary mode of HIV transmission in Rakai is believed to be heterosexual. Data were collected as part of a large international collaborative research project.

Sex in Sweden: Data from Sweden come from the 1996 "Sex in Sweden" survey based on a nationwide probability sample and financed by the Swedish National Board of Health (Lewin, 1996).

National health and social life survey: Data from the United States comes from the National Health and Social Life Survey (NHSLS) (Laumann et al., 1994). NHSLS was a national probability sample of the sexual behavior of Americans. Sexual partner count was ascertained by two different techniques for the NHSLS. First, respondents were asked to indicate the number of sexual partners they had had in the last year (and in their lives) in face-to-face interviews. Second, respondents were asked the same questions on a written instrument in which partner numbers were binned into categories of $k = \{0, 1, 2-5, 5-10, 10-20, 20-100\}$.

There is considerable evidence that the latter method generally yields more reliable results (Laumann et al., 1994; Lewin, 1996). We therefore use the responses to the written question for our model estimation. This introduces some complications to the likelihood function which are nonetheless easily handled (see below).

Neither Sweden nor the United States is characterized by a generalized HIV/AIDS epidemic with national prevalence for both countries less than 1%. Table 1 presents summary statistics for the three studies.

2.2. Statistical inference

We adopt a likelihood framework to estimate the model parameters and compare the different models against each other. The likelihood framework provides a set of powerful tools for inference. The method of maximum likelihood estimation has been traditionally regarded as optimal based on asymptotic arguments (Casella and Berger, 2002). For most models the MLE is approximately normally distributed (Johnson et al., 1992) and for small sample sizes the uncertainty of the MLE can be quantified by the bootstrap (Efron and Tibshirani, 1993). Here, we employ bootstrap methods to quantify the small sample properties of the MLEs and calculate confidence intervals.

Like any sample from a population, the samples obtained in the sexual surveys we analyze here are imperfect representations of their populations. Errors accrue due to sampling frame mis-specification, informant mis-report and non-response (Rubin, 1987; Morris, 1993b). Likelihood methods enjoy the tremendous advantage that the sampling design is "ignorable" for many standard (and non-standard) designs under the likelihood framework (Thompson and Seber, 1996). That is, the likelihood only depends on data in the sample, and not on unknown missing data.

Since our models allow for the decoupling of the behavior of the majority of observations and that of the tail, the complete data likelihood will be somewhat more complicated than a standard likelihood (Groeneboom and Wellner, 1992). Define k_{min} as the degree above which the parametric model (e.g., the Yule) is fit and denote $P(K = k) = \pi_k$ for $k \leqslant k_{min}$. Given n total observations with observed frequencies $n_0, n_1, \ldots$ on network degree $k = 0, 1, \ldots$, the log-likelihood of the data is

$$\begin{aligned}\mathscr{L}(\pi, \theta | K_1 = k_1, \ldots, K_n = k_n) \\ = \sum_{m=0}^{k_{min}} n_m \log(\pi_m) + \left(n - \sum_{m=0}^{k_{min}} n_m\right) \log\left(1 - \sum_{m=0}^{k_{min}} \pi_m\right) \\ + \sum_{m=k_{min}+1}^{\infty} n_m \log(P(K = m | K > k_{min})). \qquad (3)\end{aligned}$$

The last term is the contribution to the likelihood of the observed values given that they are in the tail of the distribution. From the form of (4) the observed values in the tail are sufficient for the parameter θ. Similarly, the MLEs of π_m are given by the sample proportions, $\hat{\pi}_m = n_m/n, m = 0, 1, \ldots, k_{min}$ (see Appendix A).

Special care was needed in calculating the likelihood for the NHSLS data, because the data on sexual partner count were grouped. We employed a mixed parametric likelihood approach to the problem (Groeneboom and Wellner, 1992). The full data likelihood for NHSLS is presented in the appendix.

2.2.1. Model selection

The different models we consider have different numbers of parameters. In general, a model with more parameters may be expected to fit data better than a model with few parameters. The likelihood $\mathscr{L}$ provides a measure of the goodness-of-fit of a model to the data that does not adjust for the complexity of the model. If the models are nested, in the sense that they form a sequence with each model being a subset of a previous more complex one, then likelihood ratio tests can be used. Many solutions have been proposed for non-nested situations, such as we are faced with here. We adopt two different approaches: (1) the Akaike information criterion (AIC) (Akaike, 1974; Burnham and Anderson, 2002) and (2) the Bayesian information Criterion (BIC) (Raftery, 1995). For a simple random sample of n people with data $K_1, \ldots, K_n$, the AIC is defined as $\mathrm{AIC} = -2\mathscr{L}(\hat{\theta}|K_1 = k_1, \ldots, K_n = k_n) + 2d$, and $\mathrm{BIC} = -2\mathscr{L}(\hat{\theta}|K_1 = k_1, \ldots, K_n = k_n) + \log(n)d$.

A model is judged better than another model if it has a smaller AIC (or BIC) value. Both AIC and BIC have solid theoretical foundations: Kullback–Leibler distance in information theory (for AIC), and integrated likelihood in Bayesian theory (for BIC). The BIC approach has the benefit of incorporating model uncertainty and sample size into the decision. The AIC has the advantage of efficiency. That is, for large sample size, it is the best approximation to the "true" model (Burnham and Anderson, 2002). If the complexity of the true model does not increase with the size of the data set, the BIC is usually preferred, otherwise AIC is preferred. However, both criteria should be used for guidance and not used to unilaterally exclude models solely based on ranking.

3. Results

The results of the model fits are presented in Table 2. We have listed the five best-fitting models, by AIC. Four out of the six networks were best fit by the shifted

Table 2
Top five best fitting models for each network ordered by the value of the AIC

Population	Sex	Model	k_{min}	d	$\mathscr{L}$	AIC	BIC	ρ/Mean	Scale/s.d.
Uganda	F	nb	2	4	−525.17	1058.35	1077.10	0.27	1.90
		yule	3	4	−525.32	1058.64	1077.39	3.68	NA
		geo	3	4	−525.45	1058.91	1077.66	2.67	2.67
		nb	3	5	−524.89	1059.79	1083.23	1.88	0.11
		tpl	3	5	−525.28	1060.57	1084.01	2.59	14.84
	M	geo	4	5	−781.10	1573.10	1595.26	2.72	2.72
		zero-nb	4	6	−780.90	1573.80	1600.38	6.29	0.93
		tpl	4	6	−781.02	1574.03	1600.62	4.54	0.88
		nb	4	6	−781.22	1574.44	1601.03	3.58	0.52
		yule	2	3	−784.25	1574.50	1587.79	5.43	NA
Sweden	F	nb	1	3	−1068.45	2142.90	2158.27	0.38	3.62
		yule	2	3	−1068.64	2143.27	2158.64	4.23	NA
		waring	2	4	−1067.95	2143.91	2164.39	6.53	1.87
		g-y	2	4	1068.16	2144.31	2164.80	2.89	4.62
		yule	3	4	−1068.17	2144.33	2164.82	4.68	NA
	M	nb	1	3	−1509.17	3024.34	3040.07	0.66	2.47
		g-y	2	4	−1508.21	3024.43	3045.39	2.47	7.35
		tpl	2	4	−1508.22	3024.44	3045.41	1.82	6.75
		nb-y	1	4	−1508.27	3024.55	3045.51	1.51	1.23
		nb-y0	1	4	−1508.37	3024.55	3045.51	1.51	1.23
USA	F	yule	2	3	−1600.74	3207.48	3224.03	3.84	NA
		waring	1	3	−1601.34	3208.67	3225.23	3.11	−0.68
		tpl	4	6	−1598.64	3209.27	3242.38	33.07	0.14
		pois	4	5	−1599.69	3209.39	3236.98	0.69	NA
		yule	3	4	−1600.71	3209.42	3231.49	3.91	NA
	M	nb	1	3	−1599.14	3204.28	3220.08	0.78	0.26
		nb	2	4	−1605.31	3218.63	3239.70	1.43	1.75
		geo	2	3	−1608.33	3222.66	3238.46	1.39	1.39
		nb	3	5	−1606.58	3223.15	3249.49	1.65	2.09
		nb	4	6	−1607.11	3226.21	3257.81	1.95	2.67

k_{min} is the cutoff, d is the number of parameters of the model, log $\mathscr{L}$ is the log-likelihood of the model, AIC is Akaike's information criterion, BIC is the Bayesian information criterion, ρ is the scaling parameter for power distributions or the mean of the non-power distributions (e.g., negative binomial, geometric), scale is the second parameter of the distribution if applicable. Models: nb = shifted negative binomial, geo = geometric, tpl = truncated power law, zero-nb = zero negative binomial, g-y = geometric-yule, pois = Poisson, nb-y = negative binomial yule.

negative binomial model and one was best fit by the special case of the negative binomial, the geometric. For all networks but one, the best-fitting model had low k_{min} (i.e., 1 or 2), whereas for the Ugandand men, $k_{min} = 4$.

For all the networks but one, the top five best-fitting models had similar values of the likelihood (AIC/BIC). However, for the American men's network, the negative binomial model fit dramatically better than the next best model.

For all networks in which a power-law model (i.e., Yule, Waring) appeared, the value of the scaling parameter $\rho > 3$, indicating that all networks were characterized by finite variance (Jones and Handcock, 2003a).

4. Discussion

A wide range of potential models for the degree distributions exist and empirical data are often limited. Frequently, many models will fit the empirical information approximately equally well, at least superficially. Thus sophisticated statistical methodology should be used to assess the quality of the fit of proposed models, and the plausibility of stochastic mechanisms underlying the model must be weighed. Kendall (1961) emphasized this point in his 1960 inaugural presidential address to the Royal Statistical Society. He argued that for statistical modeling in the social sciences to mature as a scientific discipline, it must move beyond simple curve fitting exercises and into tests of process models. While revealing regular patterns in social systems is an important first step in scientific understanding, their existence does not entail a causal mechanism. Since the observed patterns in the social world were generated by the behavior of individual actors, specifying plausible, testable stochastic processes by which larger patterns emerge is essential in the scientific understanding of social phenomena. Unfortunately, this message has often been lost in the passage of time and the segmentation of scientific enterprise.

Our results indicate that no unitary stochastic process easily explains the formation of sexual networks in the populations we have examined. It has been suggested that sexual networks may display power-law behavior (Liljeros et al., 2001; Dezső and Barabási, 2002). While some of the power-law models appear to fit the women's networks, both the men's and women's networks were generally better fit by the negative binomial and variants. Preferential attachment is one stochastic mechanism for the evolution of networks characterized by power-law degree distributions (Albert and Barabási, 2000; Pastor-Satorras and Vespignani, 2001). Lloyd and May (2001) suggested that such processes might be relevant for human sexual contact networks and this hypothesis found putative empirical support by Liljeros et al. (2001) in the Swedish sexual networks analyzed in this paper. We found the support for preferential attachment to be mixed overall. While a preferential attachment model (i.e., Yule, Waring) fell into the top five best-fitting models for all networks but the American and Swedish men, it was the top model only in the case of American women.

Despite their appealing stochastic mechanism, the vetting models performed poorly relative to the simpler models.

As noted in Jones and Handcock (2003a), many of these degree distributions are essentially L-shaped for all three populations. That is, the distributions fall off very rapidly from the mode of $k = 1$, with a very small number of observations in the tail. The L-shaped degree distributions of these networks may favor the power-law models, such as the Yule or discrete Pareto. However, models fit by maximum likelihood are sensitive to outliers induced by reporting error. While the great majority of the data may indicate a certain range for the parameter(s) of interest, an outlier, by definition, will have very low probability at these parameter values. Consequently, a distribution contaminated by reporting errors may yield a maximum likelihood estimator that is worse for more of the data.

4.1. *The importance of heterogeneity*

The stochastic mechanism most clearly supported by the model selection procedure was individual heterogeneity in the propensity for forming partnerships. This is reflected in the negative binomial fitting best for five of the six networks, and in the magnitude of the likelihood criteria supporting this model for the American men.

One interpretation for such a model is that each individual person has a constant hazard of forming a new sexual partnership, but that this hazard varies from individual to individual. For the great majority of people, this hazard is very low.

4.2. *Limitations of exclusively degree-based models*

Research on social networks with application to epidemiology has focused on degree distribution as the primary property of interest (Newman, 2002a, b; Liljeros et al., 2001). This is a natural starting point, both because of the theoretical focus on heterogeneity in sexual activity (Anderson et al., 1986), and the relative availability of local network data. However, there are other features of sexual networks which have a bearing on epidemic processes. Two such features that have received a great deal of attention recently are clustering and minimum path length, the determinants of the so-called "small world" network effect. Small-world networks can arise from power-law models of the degree distribution (Amaral et al., 2000). Small world graphs tend to have larger final epidemic sizes than other sparse networks because of their relatively high connectedness (Newman, 2002b).

Nodal attributes, such as gender, ethnicity or marital status, are also of fundamental importance for the formation of networks (Morris, 1991). The probability that an actor will form a sexual contact with an alter is a function of both the actor's and the alter's nodal attributes, separately from their degree (Wasserman and Pattison, 1996). Differential selectivity in interaction is not easily accommodated in standard degree-based epidemic theory. The standard model of heterogeneity (Anderson and May, 1991) assumes random mixing conditional on activity class. More recent formal treatments, which expressly deal with network structure, deploy similar assumptions (Newman, 2002a).

In addition to the nodal attributes, other network properties, which are not necessarily related to degree, play a fundamental role in epidemics on networks. One property of particular importance is concurrency. Concurrency denotes the propensity to form simultaneous partnerships (Morris and Kretzschmar, 1995). In both simulation (Morris and Kretzschmar, 1995, 1997) and empirical studies, (Potterat et al., 1999) concurrency has been demonstrated to have significant effects on STD epidemics independent of the number of sexual partnerships.

An alternative to strictly degree-based network models that can accommodate all these features are the exponential random graph models (ERGMs), also known as p^* models (Frank and Strauss, 1986). While the specification of such models presents a host of technical challenges, recent advancements in statistical computing have put them within reach. Specifically, Markov–Chain Monte Carlo (MCMC) simulation provides a means both for routinely calculating the likelihoods of large ERGMs and simulation networks with given nodal and structural characteristics. Recent work into the properties of MCMC solutions to ERGM models promises to overcome some of the traditional difficulties in calculating likelihood estimates (Snijders, 2002; Handcock, 2002).

4.3. Limitations

Inference about the properties of the sexual network degree distribution is limited by data quality. There are three problematic features of available network data: (1) potentially sizable sampling errors associated with sexual history surveys (Brewer et al., 2000), (2) data coarsening (particularly for NHSLS), (3) censoring of very high-degree individuals due to AIDS mortality (King Holmes, personal communication).

While these factors all limit our ability to infer network properties, they are certainly not unique to the likelihood-based procedures we have employed. The likelihood framework, in fact, can deal with issues associated with measurement error and bias more effectively than previous methods used in this context (Jones and Handcock, 2003a). However, the intrinsic data limitations warrant caution for the interpretation of stochastic models. Any stochastic model will be a caricature of reality, and it is the modeler's responsibility to decide to what degree of precision a useful model must conform.

A potentially greater problem for scientific progress in this field is the dearth of behavioral data on how sexual partnership networks are formed. In their critique of the common practice of using observational data to simultaneously choose a model and estimate its parameters, Burnham and Anderson (2002) note that the *scientific* work in multi-model inference lies in the development of a priori mechanistic models. These models can be derived either empirically or deductively from theory. In developing alternative models for this paper, we have attempted to specify plausible behavioral mechanisms by which networks form, allowing us to confront existing models (e.g., Liljeros et al., 2001) with alternatives. Progress on this front is clearly predicated on careful behavioral work such as that of Gorbach et al. (2002) on the formation of concurrent sexual partnerships.

Table 2 reveals substantial model uncertainty. Different models yield similar AIC values for the same data. This result suggests two important features of the inference for sexual networks. First, it highlights the need both for models based on plausible stochastic processes and for behavioral data to allow us to distinguish models based on mechanism. Second, it indicates that model uncertainty should be accounted for when attempting to make predictions regarding the behavior of sexual networks or epidemics thereupon. This uncertainty can be dealt with through, for example, Bayesian model averaging (Hoeting et al., 2000).

4.4. Conclusion

The results we have presented here indicate that a general behavioral model for the formation of human sexual contact networks is still lacking. Preferential attachment is one model that has been previously suggested, and the assumptions of preferential attachment have been incorporated into subsequent work. However, this mechanism does not perform especially well when confronted with alternative models. We suggest that incorporating actor heterogeneity and dependence is essential for future network models in epidemiology. Furthermore, if we are to move beyond ad hoc curve fits of network degree distributions and make real progress in understanding the stochastic mechanisms which generate empirical networks, two points are essential: (1) we must recognize that there is much more to sexual networks than degree distributions, and (2) collaboration between network modelers, epidemiologists, and behavioral scientists is essential.

Acknowledgments

We gratefully acknowledge the critical feedback and support we have received from Martina Morris, King Holmes, Julian Besag, Adrian Raftery, Steve Goodreau, Mark Newman, Richard Hayes, Roy Anderson, and Bob May. Virginia Rutter suggested the term "vetting model." Frederik Liljeros generously made available the Swedish data used in Liljeros et al. (2001) We especially wish to thank Dr. Bo Lewin, Professor of Sociology, Uppsala University and head of the research team responsible for the "Sex in Sweden" study for providing the data used in this study. This research supported by Grant R01-DA012831 from NIDA and Grants R01-HD034957 and R01-HD41877 from NICHD.

Appendix A. Maximum likelihood estimator for π_m

Eq. (4) presents the full data log-likelihood for the models we use. Here we verify that the maximum likelihood estimator $\hat{\pi}_m$ is simply the sample proportion of m, n_m/n.

The maximum likelihood estimator is given by,

$$\frac{\partial \mathscr{L}(\pi, \theta | K_1 = k_1, \ldots, K_n = k_n)}{\partial \pi_k} = \frac{n_k}{\pi_k} - \frac{n - \sum_{m=0}^{k_{min}} n_m}{1 - \sum_{m=0}^{k_{min}} \pi_m},$$

which is zero, only if $\hat{\pi}_m = \frac{n_m}{n} \forall m = 0, \ldots$ This partial is not defined for $\pi_k = 0$ or $\sum_{m=0}^{k_{min}} \pi_m = 1$. However, in these special cases the likelihood is maximized (with probability 1) by the sample proportions.

The Hessian matrix of Eq. (4),

$$\left[\frac{\partial^2 \mathscr{L}(\pi, \theta | K_1 = k_1, \ldots, K_n = k_n)}{\partial \pi_k \partial \pi_j} \right]$$

is negative definite indicating that the MLE is the unique global maximum.

Appendix B. Likelihood for NHSLS data

The NHSLS data for annual sexual partner count were binned into categories $k = \{0, 1, 2-5, 6-10, 11-20, 21-100\}$. Suppose the observations C are a categories $\{1, \ldots, \mathscr{C}\}$. The probability of an observation in category c with inclusive range $[l_c, u_c]$ is

$$P_\theta(C = c) = P_\theta(l_c \leqslant K \leqslant u_c) = \sum_{m=l_c}^{u_c} P(K = m),$$

where $P_\theta(K = m)$ is the log-likelihood. The exact observation of the degree is the special case where the categories correspond to a single degree.

Suppose we observe categories $C_1 = c_1, \ldots, C_n = c_n$. The full data likelihood for the grouped NHSLS data is then

$$\mathscr{L}(\theta | C_1 = c_1, \ldots, C_n = c_n) = \sum_{m=1}^{n} \log(P_\theta(C = c_m)).$$

References

Akaike, H., 1974. A new look at statistical model identification. IEEE Trans. Autom. Control AU-19, 716–722.

Albert, R., Barabási, A.L., 2000. Topology of evolving networks: local events and universality. Phys. Rev. Lett. 85 (24), 5234–5237.

Alexander, N., Moyeed, R., Stander, J., 2000. Spatial modelling of individual-level parasite counts using the negative binomial distribution. Biostatistics 1 (4), 453–463.

Amaral, L.A.N., Scala, A., Barthelemy, M., Stanley, H.E., 2000. Classes of small-world networks. Proc. Nat. Acad. Sci. USA 97 (21), 11149–11152.

Anderson, R.M., Garnett, G.P., 2000. Mathematical models of the transmission and control of sexually transmitted diseases. Sexually Transmitted Dis. 27 (10), 636–643.

Anderson, R.M., May, R.M., 1991. Infectious Diseases of Humans: Dynamics and Control. Oxford University Press, Oxford.

Anderson, R., Medley, G., May, R.M., Johnson, A., 1986. 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.

Aral, S.O., 1999. Sexual network patterns as determinants of STD rates: paradigm shift in the behavioral epidemiology of STDs made visible. Sexually Transmitted Dis. 26 (5), 262–264.

Bailey, N., 1975. The Mathematical Theory of Infectious Disease. Hafner Press, New York.

Barabási, A.L., Albert, R., 1999. Emergence of scaling in random networks. Science 286 (5439), 509–512.

Bearman, P., 1997. Generalized exchange. Amer. J. Sociology 102, 1383–1415.

Borgatti, S.P., Everett, M.G., 1992. Notions of position in network analysis. Sociological Methodology 22, 1–36.

Brewer, D.D., Potterat, J.J., Garrett, S.B., Muth, S.Q., Roberts, J.M., Kasprzyk, D., Montano, D.E., Darrow, W.W., 2000. Prostitution and the sex discrepancy in reported number of sexual partners. Proc. Nat. Acad. Sci. USA 97 (22), 12385–12388.

Burnham, K., Anderson, D., 2002. Model Selection and Inference: A Practical Information-Theoretic Approach, 2nd Edition. Springer, New York.

Casella, G., Berger, R., 2002. Statistical Inference, 2nd Edition. Duxbury, Pacific Grove.

Centers for Disease Control and Prevention, 2001. Sexually transmitted disease surveillance 2000 supplement: chlamydia prevalence monitoring project annual report. Technical Report, Department of Health and Human Services, Centers for Disease Control and Prevention.

Dezső, Z., Barabási, A.L., 2002. Halting viruses in scale-free networks. Phys. Rev. E 65, art. no. 055103.

Diekmann, O., Heesterbeek, J.A.P., Metz, J.A.J., 1990. On the definition and the computation of the basic reproduction ratio R_0 in models for infectious-diseases in heterogeneous populations. J. Math. Biol. 28 (4), 365–382.

Dorogovtsev, S.N., Mendes, J.F.F., Samukhin, A.N., 2000. Structure of growing networks with preferential linking. Phys. Rev. Lett. 85 (21), 4633–4636.

Efron, B., Tibshirani, R.J., 1993. An Introduction to the Bootstrap. Chapman & Hall, New York.

Foulkes, M.A., 1998. Advances in HIV/AIDS statistical methodology over the past decade. Statist. Med. 17 (1), 1–25.

Frank, O., Strauss, D., 1986. Markov graphs. J. Amer. Statist. Assoc. 81 (395), 832–842.

Gorbach, P.M., Stoner, B.P., Aral, S.O., Whittington, W., Holmes, K.K., 2002. "It takes a village": understanding concurrent sexual partnerships in seattle, washington. Sexually Transmitted Dis. 29 (8), 62–453.

Groeneboom, P., Wellner, J., 1992. Information Bounds and Nonparametric Maximum Likelihood Estimation. Birkhäuser, Basel.

Handcock, M., 2002. Assessing degeneracy in statistical models of social networks. Working paper, Center for Statistics and the Social Sciences, University of Washington.

Hethcote, H.W., Yorke, J.A., 1984. Gonorrhea: transmission dynamics and control. Lecture Notes in Biomath. 56, 1–105.

Hoeting, J.A., Madigan, D., Raftery, A.E., Volinsky, C.T., 2000. Bayesian model averaging: a tutorial. Statist. Sci. 15 (3), 193–195.

Hubert, M., Bajos, N., Sandfort, T., 1998. Sexual Behaviour and HIV/AIDS in Europe: Comparisons of National Surveys. UCL Press, London.

Irwin, J., 1963. The place of mathematics in medical and biological statistics. J. Roy. Statist. Soc. Ser. A (General) 126 (1), 1–45.

Johnson, N., Kotz, S., Kemp, A., 1992. Univariate Discrete Distributions, 2nd Edition. Wiley Series in Probability and Mathematical Statistics. Wiley, New York.

Jones, J., Handcock, M.S., 2003a. An assessment of preferential attachment as a mechanism for human sexual network formation. Proc. Roy. Soc. London, B 270, 1123–1128.

Jones, J., Handcock, M.S., 2003b. Sexual contacts and epidemic thresholds. Nature 423 (6940), 605–606.

Kendall, M., 1961. Natural law in the social sciences: presidential address. J. Roy. Statist. Soc. Ser. A-Statist. Soc. 124 (1), 1–16.

Laumann, E., Gagnon, J., Michael, T., Michaels, S., 1994. The Social Organization of Sexuality: Sexual Practices in the United States. University of Chicago Press, Chicago.

Levene, M., Fenner, T., Loizou, G., Wheeldon, R., 2002. A stochastic model for the evolution of the web. Comput. Networks 39 (3), 277–287.

Lewin, B. (Ed.), 1996. Sex in Sweden, National Institute of Public Health, Stockholm.

Liljeros, F., Edling, C.R., Amaral, L.A.N., Stanley, H.E., Åberg, Y., 2001. The web of human sexual contacts. Nature 411 (6840), 907–908.

Lloyd, A.L., May, R.M., 2001. Epidemiology—how viruses spread among computers and people. Science 292 (5520), 1316–1317.

Martin, D., Katti, S., 1965. Fitting of some contagious distributions to some available data by maximum the likelihood method. Biometrics 21, 34–48.

May, R.M., Lloyd, A.L., 2001. Infection dynamics on scale-free networks. Phys. Rev. E 64 (6), 066112.

Moody, J., 2001. Race, school integration, and friendship segregation in america. Amer. J. Sociol. 107 (3), 679–716.

Morris, M., 1991. A log-linear modeling framework for selective mixing. Math. Biosci. 107 (2), 349–377.

Morris, M., 1993a. Epidemiology and social networks: modeling structured diffusion. Sociological Methods and Research 22 (1), 99–126.

Morris, M., 1993b. Telling tails explain the discrepancy in sexual partner reports. Nature 365 (6445), 437–440.

Morris, M., 1995. Data driven network models for the spread of disease. In: Mollison, D. (Ed.), Epidemic Models: Their Structure and Relation to Data. Cambridge University Press, Cambridge, pp. 302–322.

Morris, M., 1997. Sexual networks and HIV. AIDS 11, S209–S216.

Morris, M., 2004. Network Epidemiology: A handbook for survey design and data collection. International Studies in Demography Series. Oxford University Press, Oxford.

Morris, M., Kretzschmar, M., 1995. Concurrent partnerships and transmission dynamics in networks. Soc. Networks 17 (3–4), 299–318.

Morris, M., Kretzschmar, M., 1997. Concurrent partnerships and the spread of HIV. AIDS 11 (5), 641–648.

National Institute of Medicine, 1997. The Hidden Epidemic: Confronting Sexually Transmitted Diseases. National Academy Press, Washington, DC.

Newman, M.E.J., 2001. The structure of scientific collaboration networks. Proc. Nat. Acad. Sci. USA 98, 404–409.

Newman, M., 2002a. Random graphs as models of networks. Working Paper 02-02-005, Santa Fe Institute.

Newman, M.E.J., 2002b. Spread of epidemic disease on networks. Phys. Rev. E 66 (1), art. no. 016128.

Pastor-Satorras, R., Vespignani, A., 2001. Epidemic dynamics and endemic states in complex networks. Phys. Rev. E 63 (6), art. no–066117.

Potterat, J.J., Zimmerman-Rogers, H., Muth, S.Q., Rothenberg, R.B., Green, D.L., Taylor, J.E., Bonney, M.S., White, H.A., 1999. Chlamydia transmission: concurrency, reproduction number, and the epidemic trajectory. Amer. J. Epidemiol. 150 (12), 1331–1339.

Raftery, A.E., 1995. Bayesian model selection in social research. Sociol. Methodol. 25, 111–163.

Rubin, D., 1987. Multiple Imputation for Nonresponse in Surveys. Wiley, New York.

Sewankambo, N.K., Gray, R.H., Ahmad, S., Serwadda, D., Wabwire-Mangen, F., Nalugoda, F., Kiwanuka, N., Lutalo, T., Kigozi, G., Li, C.J., Meehan, M.P., Brahmbatt, H., Wawer, M.J., 2000. Mortality associated with HIV infection in rural Rakai District, Uganda. AIDS 14 (15), 2391–2400.

Simon, H., 1955. On a class of skew distribution functions. Biometrika 42 (3–4), 425–440.

Snijders, T.A.B., 2002. Markov chain monte carlo estimation of exponential random graph models. J. Social Structure 3 (2).

Thompson, S.K., Seber, G.A., 1996. Adaptive Sampling. Wiley, New York.

Wasserman, S., Faust, K., 1994. Social Network Analysis: Methods and Applications. Structural Analysis in the Social Sciences. Cambridge University Press, Cambridge.

Wasserman, S., Pattison, P., 1996. Logit models and logistic regressions for social networks: an introduction to Markov graphs and p^*. Psychometrika 61 (3), 401–425.

Williams, L.A., Klausner, J.D., Whittington, W.L., Handsfield, H.H., Celum, C., Holmes, K.K., 1999. Elimination and reintroduction of primary and secondary syphilis. Amer. J. Public Health 89 (7), 7–1093.

Youm, Y., Laumann, E., 2002. Social network effects on the transmission of sexually transmitted diseases. Sexually Transmitted Dis. 29 (11), 689–697.

Yule, G., 1925. A mathematical theory of evolution based on the conclusions of Dr. J.C. Willis FRS. Philos. Trans. Roy. Soc. London Ser. B-Biol. Sci. 213, 21–87.