Statistical Evidence Tells Tails of Human Sexual Contacts

James Holland Jones[1]
Center for AIDS & STD, University of Washington, Seattle, WA 98104

Mark S. Handcock[2]
Center for Statistics and the Social Sciences,
University of Washington, Seattle, WA 98195

Working Paper no. 21
Center for Statistics and the Social Sciences
University of Washington
Box 354322
Seattle, WA 98195-4322, USA

May 2002

[1]email: jameshj@u.washington.edu
[2]email: handcock@stat.washington.edu

There has been a growing interest in the application of social network theory to the epidemiology of sexually-transmitted diseases (STD).[1] This interest arises from recognition that STDs are transmitted through binary contacts and, consequently, predictions about STD epidemics are unlikely to be robust to the mass-action assumptions of classical mathematical epidemiology.[2] Substantial attention has recently been given to the possibility that human sexual networks exhibit scale-free behavior.[3, 4] We consider scale-free networks that arise when the probability mass function for partner number (i.e., network degree) is given by a power law, $P(k) \propto k^{-\alpha}$, with $2 < \alpha \leq 3$. Such networks are characterized by a degree distribution with infinite variance, and since the epidemic threshold parameter increases linearly with the variance under behavioral heterogeneity,[2] the epidemic threshold should always be exceeded in such a scale-free system. We derive the maximum likelihood estimator for the scaling exponent in the discrete power-law distribution, and show that the statistical fit of the power-law model is very poor for a large local network data set from Uganda. This finding suggests that human sexual networks may be more complex than simple analogy to computer networks might suggest, and that a more flexible, actor-based approach to modeling sexual networks is required to understand and control STD epidemics.

We use data from the Rakai Project Sexual Network Survey.[5] The sample consists of the elicited counts of sexual partners for the last year and lifetime of 1353 men (573) and women (780) ages 15-49. From these data, we calculated the maximum likelihood estimates of the discrete version of the probability model presented by May and Lloyd.[4] This model is

typically known as the zeta distribution,[6] which specifies that the probability of observing a node degree of exactly k is given by,

$$P(K = k) = \frac{1}{\zeta(\alpha)k^{\alpha}}, \tag{1}$$

where $\zeta(\alpha)$, the Riemann zeta function of α.

The maximum likelihood estimator (MLE) for the scaling parameter α in equation 1 is the solution $\hat{\alpha}$ of:

$$-\frac{\zeta'(\alpha)}{\zeta(\alpha)} = \overline{\log(k)} \tag{2}$$

where $\overline{\log(k)}$ is the mean of the log-degree distribution. The uncertainty of the MLE can be estimated by the bootstrap [7] and is approximately normally distributed.[6]

In figures 1 and 2, we plot the observed frequency of degree k against k together with the maximum likelihood fits to the model with 95% confidence intervals. While the fit is somewhat better for men than it is for women, both fits are clearly very bad. This result may appear somewhat surprising given that a fit by eye yields the expected slope for a scale free network ($2 < \alpha \leq 3$). However, neither visual nor OLS regression [8] fits to such plots are appropriate since the values of a distribution function will be highly correlated and inhomogeneous in variance.

The scaling parameter for the population of women in Uganda is outside of the range of scale-free behavior (CI $[4.21, 5.09]$; $\widehat{\alpha} = 4.659$), while the scaling parameter for men falls within the region of scale-free behavior (CI $[2.38, 2.69]$; $\widehat{\alpha} = 2.54$). When we estimate the scaling parameter conditional on observations of degree $k_{min} \geq 2$, the lower bound of the

confidence interval is above the range of scale-free networks. In fact, higher values of k_{min} infer increasing values. As the power law scaling is supposed to capture the behavior of the tail of the degree distribution, this finding suggests that this sexual network is not scale-free. Respondent reporting of the number of sexual partners in the extreme upper tail is notoriously suspect.[9] Our fits of power law models that are robust to this reporting bias tend to support somewhat lower values of the scaling parameter. This suggests that upper tail misreporting should receive special attention when assessing the statistical validity of power law models to sexual network data.

We applied these methods to 1996 Swedish data derived from figure 2a of Liljeros et al.[3] with similar results. Higher values of k_{min} infer more divergent values for the scaling again suggesting a poor fit of the scale-free model to the observed local network structure. The version of the power law model that represents k as a continuous variable infers values of α above the scale-free range for both Uganda and Sweden.

The censoring of partners makes the fitting of models to reported lifetime partner count data problematic. Cross-sectional survey data on the degree distribution contain considerable hidden reporting heterogeneity as the cumulative lifetime partner count is likely to increase with age. To assess this quantitatively, we fit a Poisson generalized linear model, using age as a covariate. Age has a significant effect on the expected lifetime reported partner count in both men (Poisson GLM, $\beta_{\text{age}} = 0.069, z = 41.43, p < 0.001$) and women (Poisson GLM, $\beta_{\text{age}} = 0.017, z = 41.43, p < 0.001$). Given the exceptionally poor fit of the yearly degree scaling, together with the censoring, we must treat the fit of the lifetime degree scaling with skepticism. The epidemiological significance of cumulative lifetime networks can also be questioned since these are not the structures that an STD pathogen must navigate in order

to persist in a population. The actual network over which a pathogen passes is extremely sparse, a fact which has made statistical estimation of sexual networks extremely challenging.

Our results have important implications for the interpretation of random graph models of social networks. The power law model predicts that there will only be an epidemic transition in a small region of the parameter space of α. Our statistical estimates of α place it well outside of these limits, and in a region where an epidemic on a scale-free network is impossible.[10] Given the fact that Uganda has one of the highest HIV incidence rates in the world, the scale-free power law model clearly fails to describe Ugandan sexual networks.

Public health policy aimed toward eradication of STD pathogens can benefit tremendously from theory derived from mathematical models of underlying biological and behavioral processes.[11] The continued development of a theory of STD epidemiology should be grounded in the behavior of the individual actors who comprise the social networks along which pathogens propagate. The development of a rigorous theory of STD epidemiology will benefit from combining traditional mathematical approaches with the inferential tools of statistical science.[12]

Acknowledgements

We wish to thank Martina Morris for useful discussion. This work supported by grants from NICHD.

Direct correspondence to: James Holland Jones, Center for Statistics and the Social Sciences, University of Washington Box 354320, Seattle, WA, USA 98195-4320.

1. Morris, M. Sexual networks and HIV. *AIDS* **11**, S209–S216 (1997).

2. Anderson, R. M. and May, R. M. *Infectious diseases of humans: Dynamics and control*. Oxford University Press, Oxford, (1991).

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

4. May, R. M. and Lloyd, A. L. Infection dynamics on scale-free networks. *Physical Review E* **64**(6), 066112 (2001).

5. Wawer, M. HIV prevention study. Research Grant R01HD028886, National Institute of Child Health and Human Development, (1992).

6. Johnson, N., Kotz, S., and Kemp, A. *Univariate discrete distributions*. Wiley series in probability and mathematical statistics. Wiley, New York, 2nd edition, (1992).

7. Efron, B. and Tibshirani, R. J. *An introduction to the bootstrap*. Chapman and Hall, New York, (1993).

8. Axtell, R. L. Zipf distribution of us firm sizes. *Science* **293**(5536), 1818–1820 (2001).

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

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

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

12. Poole, D. and Raftery, A. E. Inference for deterministic simulation models: The Bayesian melding approach. *Journal of the American Statistical Association* **95**(452), 1244–1255 (2000).

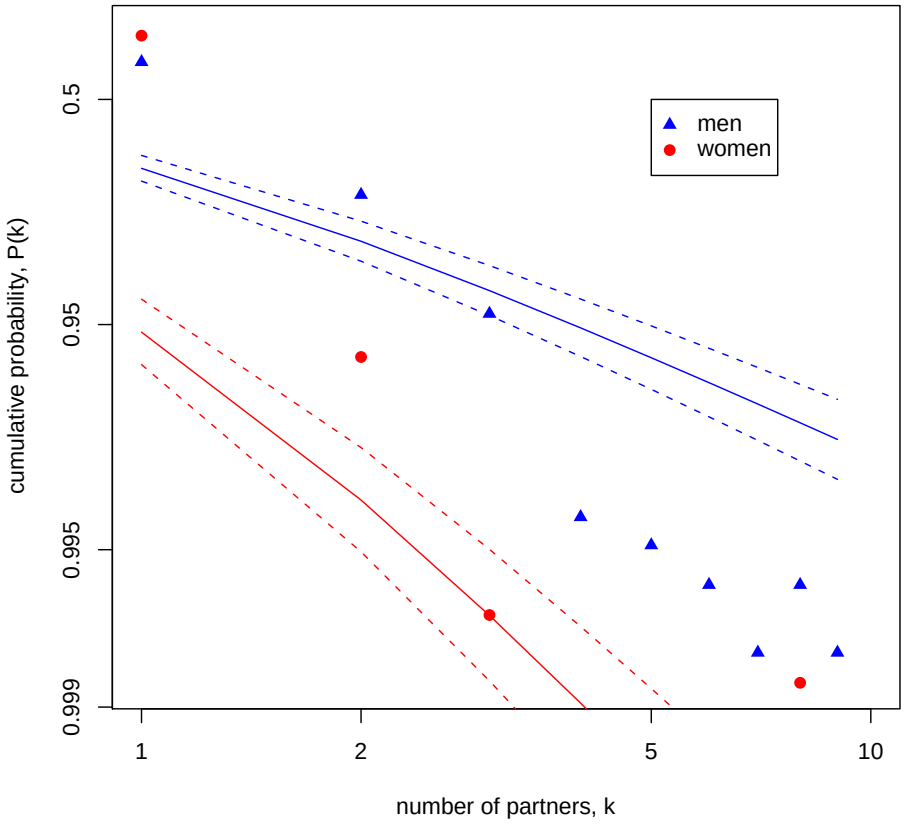

Figure 1: Plot of empirical degree distributions for men and women together with maximum likelihood estimates (with 95% confidence bands) of the power law model. For cutoff degree of $k_{min} = 1$. The MLE of the scaling parameter is $\widehat{\alpha} = 4.65$ $(4.21, 5.09)$ for women and $\widehat{\alpha} = 2.54$ $(2.38, 2.69)$ for men. Note the discrepancy of the fits based on the statistical evidence from the fits "by eye" or based on naïve least-squares regression.

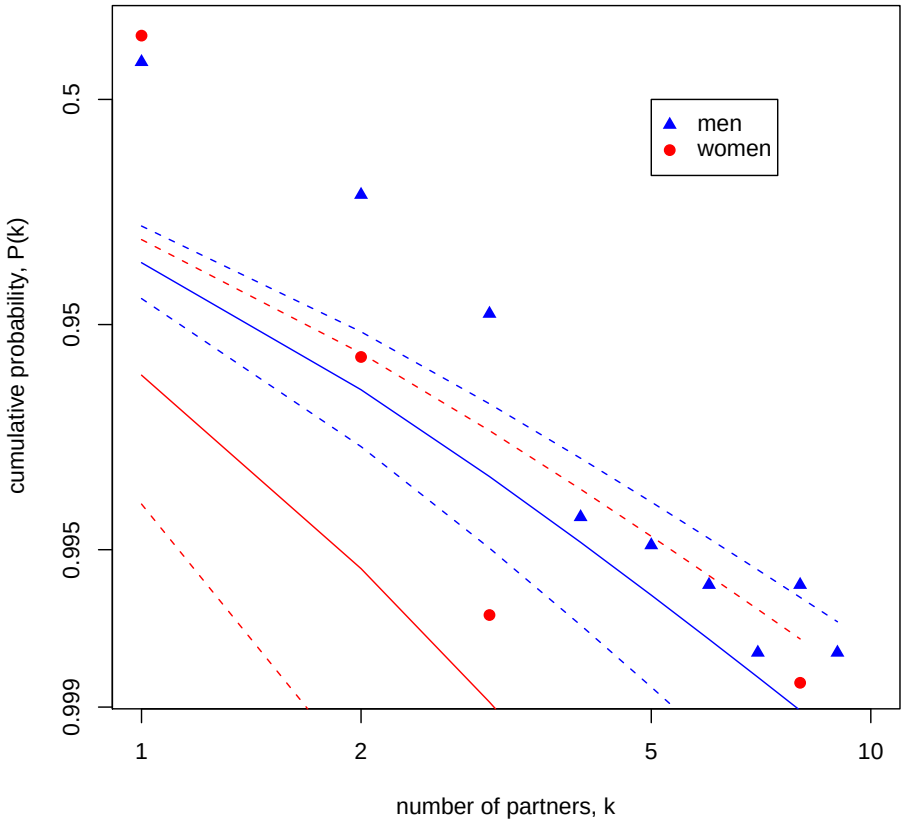

Figure 2: Plot of empirical degree distributions for men and women together with maximum likelihood estimates (with 95% confidence bands) of the power law model. For cutoff degree of $k_{min} = 2$. The MLE of the scaling parameter is $\widehat{\alpha} = 5.24$ $(3.43, 7.05)$ for women and $\widehat{\alpha} = 3.73$ $(3.25, 3.73)$ for men. Note the discrepancy of the fits based on the statistical evidence from the fits "by eye" or based on naïve least-squares regression.