# An Assessment of Preferential Attachment as a Mechanism for Human Sexual Network Formation

James Holland Jones,[1,2,3*] Mark S. Handcock,[1,2]

[1]Center for Statistics and the Social Sciences
University of Washington, Seattle, WA, 98195
[2]Center for Studies in Demography and Ecology
University of Washington, Seattle, WA 98195
[3]Center for AIDS and Sexually Transmitted Diseases
University of Washington, Seattle, WA

[*]To whom correspondence should be addressed; email: `jameshj@stat.washington.edu`

1

Recent research into the properties of human sexual contact networks has suggested that the degree distribution of the contact graph exhibits power-law scaling. One notable property of this power-law scaling is that for a range of scaling exponents, the variance of the degree distribution is infinite. This property is of fundamental significance for the control of sexually transmitted diseases (STDs) such as HIV/AIDS since infinite variance of the degree distribution implies no epidemic threshold, and that an STD can persist regardless of its transmissibility. A stochastic process, known as preferential attachment, that yields one form of power law scaling has been suggested to underlie the scaling of sexual degree distributions. The limiting distribution of the preferential attachment process is the Yule distribution, which we fit using maximum likelihood (ML) to local network data for samples of three populations: (1) the Rakai District, Uganda, (2) Sweden, and (3) USA. For all local networks but one, our interval estimates of the scaling parameters do not overlap the range in which Yule distribution has infinite variance. The exponent for male networks in the USA is close to the infinite-variance range, but the preferential attachment model is a very poor fit to these data. We conclude that epidemic thresholds exist in both single-sex and two-sex epidemic model formulations. A strong conclusion we derive from these results is that public health interventions aimed at reducing the transmissibility of STD pathogens, such as implementing condom use or high activity anti-retroviral therapy (HAART), have the potential of bringing a

population below the epidemic transition, even in populations exhibiting large degrees of behavioral heterogeneity.

# Introduction

The course of an epidemic of an infectious disease is governed by a threshold parameter, $R_0$, the basic reproductive number (Anderson and May, 1991). $R_0$ represents the expected number of secondary cases produced by a single index case in a population of susceptibles. In its most general formulation, $R_0$ is a product of the transmissibility of the infectious agent, the duration of the infection, and some measure of the contact rate between susceptible and infected individuals. Public health strategies for control and eradication are based on reducing transmissibility, shortening the duration of infection, and reducing the contact rate between susceptibles and infecteds. A puzzle in sexually transmitted disease epidemiology has been how epidemics are maintained given the the relatively small number of sexual contacts people have (e.g., relative to the number of contacts for non sexually-transmitted infections such as measles or influenza). The answer to this puzzle is that heterogeneity in sexual activity can drive an STD epidemic (Hethcote and Yorke, 1984).

In single-sex models with heterogeneous levels of sexual activity, $R_0$ increases approximately linearly with the variance in sex partner number (Anderson and May, 1991). Analogous results have been derived for two-sex models (Newman, 2002). Information on heterogeneity in sexual activity is typically estimated from local network data (Morris, 1997) gathered in sexual history surveys. In sexual network analysis, sexual contact networks are represented as random graphs, where the nodes of the graph represent individual people and the edges represent sexual contact. The number of edges adjacent to a particular node is its degree, and the collection of nodal degrees is the degree distri-

bution of the population (Wasserman and Faust, 1994). It is the variance of this degree distribution which plays such an important role in determining the threshold reproduction number for an STD. An understanding of the degree distribution of a sexually active population, and the micro forces which generate this distribution is an important step toward designing public health interventions to eradicate STDs.

Representative surveys of sexual behavior reveal that the typical person has very few sexual partners in the course of a year (Laumann et al., 1994; Wawer, 1992; Lewin, 1996). Given this observation, concern clearly focuses on the statistical properties of the tails of the degree distribution. Recent work on the properties of human sexual contact networks has suggested that they are characterized by power-law decay of their tails (Liljeros et al., 2001). These networks are described as "scale free" in the recent network literature. The key scientific question which arises in this work is not whether a network is scale free or not, but whether the network's degree distribution has infinite variance, a phenomenon occurring in a specific range of the scaling exponent $\rho$ of the power law. A distribution characterized by a scaling parameter in this range places significant probability on very large degrees. Consequently, there is no epidemic threshold in a population characterized by an infinite-variance degree distribution, (Pastor-Satorras and Vespignani, 2001; Lloyd and May, 2001; Newman, 2002) allowing a pathogen of arbitrarily small transmissibility to be maintained (Lloyd and May, 2001).

The intuition underlying this surprising result is that a network which is simultaneously consistent with (1) the low mean degree characteristic of human sexual behavior, and (2) the power-law decay of the tail of the degree distribution, will exhibit large connected components. Randomly infecting a node in such a network is therefore likely to yield a large epidemic. Figure 2 illustrates this idea with a simulated 50-actor (mostly) heterosexual network with infinite-variance degree distribution ($\rho = 2.5$). The algorithm

for generating this network was essentially that of Molloy and Reed (1995). The giant connected component suggests that a expected size of an epidemic started by randomly infecting a single node would be large.

In recent work on the scaling of a variety of systems with possible power-law distributions, the scaling exponent has been inferred from the plot of the empirical cumulative distribution against degree (or frequency) on double-logarithmic axes. A theoretical curve is then fit to the apparently linear region of this empirical plot, either "by eye" or using a curve fitting algorithm such as least-squares regression (Axtell, 2001). The scaling exponent is estimated from the slope of the line. If least-squares is used, the standard error of the slope estimates is used as a measure of the uncertainty of the scaling exponent.

This is a very poor statistical approach to estimation of the scaling exponent as the assumptions justifying least-squares regression do not hold. First, the empirical values are highly correlated (typical sequential correlations are 0.7 or higher). This issue is especially true for the values for higher degree where the sequential correlation approaches unity. The additional information in the latter points is very small and visual trends are as likely to be due to the high correlations as to be real. For this reason considering only the upper tail of the distribution and inferring a pattern is a very dubious practice. Second, the statistical variation in the values is not constant and increases rapidly with the degree (typically by an order of magnitude). This is due to the logarithmic nature of the plot and the decreasing probabilities. Third, it is usually the procedure to exclude values from the plot that correspond to zero frequencies (e.g., see Figure 2 in Liljeros et al. (2001)). These points contain a great deal of information on the degree distribution and their exclusion introduces bias into the estimates. Fourth, the high degree frequencies are sensitive to misreporting and population heterogeneity (Morris, 1993, e.g.). While these can be adjusted for statistically, the least-squares and regression approaches are overly

influenced by them. Finally, accessible statistical methodology, such as the likelihood approach applied here, exists that does not suffer from these defects.

To date, the only empirical estimates of scaling parameters in human sexual networks come from an analysis of sexual history survey data done in Sweden, a country with an HIV/AIDS prevalence of less than 1%, and this analysis was subject to the methodological problems described above. A critical test of the adequacy of the current formulation of sexual network scaling models therefore comes from estimating the scaling parameters in a population using robust, unbiased methodology in a variety of populations, including some with a clear epidemic. In this paper, we estimate the scaling parameters of the heterosexual contact network in the three populations: (1) Rakai District, Uganda, (2) Sweden, and (3) the USA.

## Methods

**Data**   We use local network data gathered from men and women in as a part of three large, representative surveys of sexual behavior. The Rakai district is an administrative unit of southern Uganda with a mature AIDS epidemic and an HIV/AIDS prevalence of approximately 16%. The primary mode of HIV transmission in Rakai is believed to be heterosexual. Data were collected as part of the Rakai Project Sexual Network Survey (Wawer, 1992). 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) Data from the United States comes from the National Health and Social Life Survey (NHSLS) (Laumann et al., 1994). Neither Sweden nor the United States is characterized by a generalized HIV/AIDS epidemic with national prevalence for both countries less than 1%. For all surveys, we used the reported number of sexual partners in the last year as the estimate of individual network degree. Sample sizes are given in

table 1.

The degree distributions for the three samples are plotted in figure 1.

**Stochastic Model**   The underlying stochastic model motivating the partnership distributions is essentially that of Simon (Simon, 1955). It is based on two assumptions: (1) a constant probability $(\rho - 2)/(\rho - 1)$ that the $r + 1$st partnership in the population is initiated with a previously sexually inactive person, and (2) 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 $k$ partnerships out of the $r$ total partnerships in the population. Simon called the limiting partnership distribution of this process the Yule distribution, following the pioneering work of Yule (Yule, 1924) The probability mass function (PMF) of the Yule distribution (Johnson et al., 1992) is:

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

where $\Gamma(\rho)$ is the Gamma function of $\rho$. The Yule distribution has power-law behavior in the sense that $P(K = k)/k^{-\rho}$ is approximately constant for large $k$. The stochastic formulation requires $\rho > 2$,, so the mean of the Yule distribution is defined. For $\rho \leq 3$ the variance of the Yule distribution is infinite.

**Statistical Inference**   Consider fitting a PMF $P_\theta(K = k)$ to survey information where $\theta$ is the parameter. For example, for the Yule model the parameter is $\rho$, the scaling exponent. 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. Given a random sample of $n$ individuals with reported degrees

$K_1, \ldots, K_n$ the likelihood of the model is

$$\mathcal{L}(\theta, k_{min}|K_1 = k_1, \ldots, K_n = k_n) \equiv \prod_{m=1}^{n} \log(P_\theta(K = k_m|K > k_{min})). \tag{2}$$

A maximum likelihood estimator (MLE) for the $\theta$ is a value $\hat{\theta}$ that maximizes (2) as a function of $\theta$. Formulae for the full data likelihoods are given in Jones and Handcock (2002$b$).

Although the statistical properties of the maximum likelihood estimator can be analyzed asymptotically, we employ bootstrap methods to quantify the small sample properties of MLEs and calculate confidence intervals (Efron and Tibshirani, 1993).

We adapt the model to allow for the possibility that the tail behavior (i.e., $k > 1$) of the degree distribution may differ fundamentally from the majority of the observations for which $k = 0$ or 1 (May and Lloyd, 2001). We generalize the Yule model to potentially include parameters to fit the probabilities of lower degree (Jones and Handcock, 2002$b$). To choose the best-fitting Yule model for the observed data, we employed a Bayesian Information Criterion (BIC) approach to model selection (Raftery, 1995). The BIC represents the integrated likelihood of a model and takes into account both the number of parameters a model uses as well as sample size. Given a random sample of size $n$, $(K_1, \ldots, K_n)$, the BIC is given by:

$$BIC = -2\mathcal{L}(\hat{\theta}, k_{min}|K_1, \ldots, K_n) + \log(n)(d + k_{min} + 1),$$

where $d$ is the dimension of $\theta$.

**Results** The results of the Yule model fits are given in table 1. For all estimates but Rakai women, the best-fitting model fit the proportions with degree zero and one separately.

8

For all models but one, the interval estimate of the scaling parameter fall above the range in which the Yule distribution has infinite variance (i.e., $\rho > 3$). The 95% confidence interval for $\rho$ for men from the USA NHSLS sample includes values within the infinite-variance region.

Elsewhere, we have shown the effect of conditioning on higher degree on the confidence intervals of the scaling parameter estimates for Swedish males (Jones and Handcock, 2002a). However, it is worth noting here that in addition to substantially reducing the "goodness-of-fit" and increasing the BIC, estimates based on high $k_{min}$ (e.g., 4 or 5 as in (Liljeros et al., 2001)) yield wildly increasing confidence intervals.

## Discussion

Using methods appropriate to the inference problem, we have estimated the scaling parameter of the Yule distribution, the limiting distribution for the preferential attachment process, for local sexual network data from three large datasets. The scaling parameter estimates indicate that the variance of the degree distribution for both sexes is finite in two out of the three populations, with the plausibility that $\rho < 3$ only for American men.

The estimate of the Yule scaling parameter for US men was 3.03 and the confidence interval overlaps the region of infinite variance. However, the Yule distribution was not the globally best-fitting model for the US data. In a separate paper (Jones and Handcock, 2002b), we have developed a variety of stochastic models for sexual network growth and estimated them using the same data analyzed here. For the US men, the best fitting model does not have a power tail, and therefore, has finite variance.

The predictions of the model depend on the form of the population degree distribution. The intuition underlying power-law scaling models is that the tails of the degree distribution in human sexual networks are long and decrease relatively slowly. However,

the extremely high values of the scaling exponents of the Yule model for most of the local networks indicate that the observed degree, in fact, falls off rapidly within the range of the data. Models with power tails fit the observed data because of the essentially L-shaped nature of degree distributions, where the great majority of people have low degree and a very small outlying fraction have high degree. This observation suggests that a unitary behavioral process, such as preferential attachment, is unlikely to underlie empirical sexual network degree distributions.

While the language of recent work may be novel in epidemiology, the interventions suggested by the putative power-law behavior of sexual networks are not particularly radical, as has been suggested (Liljeros et al., 2001; Dezső and Barabási, 2002). Behavioral heterogeneity was recognized as an important contributer at an early date in the HIV/AIDS epidemic (Anderson et al., 1986) and degree-based interventions proposed (Woolhouse et al., 1997). Targeting populations such as commercial sex workers (Ford and Koetsawang, 1999), truck drivers (Morris et al., 2000), army recruits (Nelson et al., 2002), or injection drug users (Neaigus, 1999) have a proven record in reducing disease incidence.

Our results suggest that efforts to reduce pathogen transmissibility are not wasted. A sexual network with finite variance will have an epidemic threshold for positive transmissibility. Indeed, public health efforts aimed at reducing the transmissibility of HIV have met with great success. Recently, Velesco-Hernandez et al. (2002) have argued that the use of HAART and other public health interventions in the San Francisco have brought the $R_0$ for HIV in gay men below threshold and, all things being equal, a slow endemic fade-out can be expected. Thailand's 100 per cent condom use intervention for commercial sex workers (CSWs) and army recruits has been a spectacular success in curbing an incipient generalized AIDS epidemic (Ford and Koetsawang, 1999; Nelson et al., 2002).

Much of the recent interest in the scaling of networks has focused exclusively on the behavior of the degree distribution (Pastor-Satorras and Vespignani, 2001; Liljeros et al., 2001; Dezső and Barabási, 2002; Newman, 2002), and some of this work proposes policy recommendations based on the inferred properties of the degree distribution (Liljeros et al., 2001; Dezső and Barabási, 2002). However, there are other features of networks which could have a substantial impact on epidemic processes. Two structural properties of networks that have received some attention are concurrency and local clustering. Morris and Kretzschmar (1995, 1997) have documented the impact of concurrency in sexual networks on the speed of epidemics and epidemic size. Networks characterized by moderate amounts of concurrency (holding degree distribution constant) produce larger epidemics faster. Watts (1999) has popularized the concept of "small world" networks, namely, those networks with high clustering and short minimum path length (relative to the Bernoulli graph). The joint effect of high clustering and short path length means that an epidemic could spread rapidly through a small world network. Amaral et al. (2000) note that power-law networks can be small world networks, but power law scaling of the degree distribution is not a necessary condition for the small world phenomenon.

The limitations of the exclusively degree-based perspective of Liljeros et al. (2001); Dezső and Barabási (2002, e.g.) are highlighted by the fact that infinite variance networks can have dramatically different structure depending on the values of other network parameters, and that these different structures should be expected to produce qualitatively different epidemic behavior. In figures 4 and 3, we again present simulated networks with the same degree distributions as that of figure 2. However, in both these networks, we further specified the the propensity to form three-paths (i.e., triangles), a measure of transitivity in networks (Wasserman and Faust, 1994). The algorithm used to generate these networks characterizes the network as an exponential random graph (Frank

and Strauss, 1986). Networks were simulated conditional on a the degree distribution used in figure 2 using Markov Chain Monte Carlo. Figure 3, shows an infinite variance network with a maximized propensity for forming triangles, whereas figure 4 presents an infinite variance network with minimized triangles. It seems highly likely that epidemic behavior on these networks, nonetheless characterized by the same infinite variance degree distribution, would be qualitatively different.

This observation highlights the need to exercise caution in developing public health policy from information on the degree distribution alone (Liljeros et al., 2001; Dezső and Barabási, 2002), regardless of the inferential proceedures employed to characterize the network.

The analysis we have provided here indicates that interventions aimed at reducing transmissibility still have the potential to eradicate STDs. Though sexual degree distributions may have long tails, the models analyzed here are characterized by finite variance. Both degree-based and transmissibility-reducing interventions have the possibility of lowering the reproductive rate of STD agents below epidemic threshold and should continue to be pursued in the quest for STD eradication.

# Acknowledgements

from NIDA.

# References

Amaral, L. A. N., A. Scala, M. Barthelemy, and H. E. Stanley. 2000. Classes of small-world networks. Proceedings of the National Academy of Sciences of the United States of America **97**:11149–11152.

Anderson, R., G. Medley, R. M. May, and A. Johnson. 1986. A preliminary study of the transmission dynamics of the Human Immunodeficiency Virus (HIV), the causative agent of AIDS. IMA Journal of Mathematics Applied in Medicine and Biology **3**:229–263.

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

Axtell, R. L. 2001. Zipf distribution of US firm sizes. Science **293**:1818–1820.

Dezső, Z. and A. L. Barabási. 2002. Halting viruses in scale-free networks. Physical Review E **65**:art. no. 055103.

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

Ford, N. and S. Koetsawang. 1999. A pragmatic intervention to promote condom use by female sex workers in Thailand. Bulletin of the World Health Organization **77**:888–94.

Frank, O. and D. Strauss. 1986. Markov graphs. Journal of the American Statistical Association **81**:832–842.

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

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

Jones, J. and M. S. Handcock, 2002$a$. Epidemic thresholds exist in human sexual contact networks.

Jones, J. and M. S. Handcock, 2002$b$. Likelihood-based inference for stochastic models of sexual network evolution.

Laumann, E., J. Gagnon, T. Michael, and S. Michaels. 1994. The social organization of sexuality: Sexual practices in the United States. University of Chicago Press, Chicago.

Lewin, B., editor. 1996. Sex in Sweden. National Institute of Public Health, Stockholm.

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

Lloyd, A. L. and R. M. May. 2001. Epidemiology - how viruses spread among computers and people. Science **292**:1316–1317.

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

Molloy, M. and B. Reed. 1995. A critical point for random graphs with a given degree sequence. Random Structures and Algorithms **6**:161–179.

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

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

Morris, M. and M. Kretzschmar. 1995. Concurrent partnerships and transmission dynamics in networks. Social Networks **17**:299–318.

Morris, M. and M. Kretzschmar. 1997. Concurrent partnerships and the spread of HIV. AIDS **11**:641–648.

Morris, M., M. J. Wawer, F. Makumbi, J. R. Zavisca, and N. Sewankambo. 2000. Condom acceptance is higher among travelers in Uganda. AIDS **14**:733–741.

Neaigus, A. 1999. The network approach and interventions to prevent HIV among injection drug users. Public Health **113**:140–150.

Nelson, K. E., S. Eiumtrakul, D. D. Celentano, C. Beyrer, N. Galai, S. Kawichai, and C. Khamboonruang. 2002. HIV infection in young men in northern Thailand, 1991-1998: Increasing role of injection drug use. Journal of Acquired Immune Deficiency Syndromes **29**:62–8.

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

Pastor-Satorras, R. and A. Vespignani. 2001. Immunization of complex networks. Physical Review Letters **86**:3200–3203.

Raftery, A. E. 1995. Bayesian model selection in social research. Sociological Methodology **25**:111–163.

Simon, H. 1955. On a class of skew distribution functions. Biometrika **42**:435–440.

Velesco-Hernandez, J., H. Gershengorn, and S. Blower. 2002. Could widespread use of combination antiretroviral therapy eradicate HIV epidemics? Lancet Infectious Disease **2**:487–493.

Wasserman, S. and K. Faust. 1994. Social network analysis: Methods and applications. Structural analysis in the social sciences, Cambridge University Press, Cambridge.

Watts, D. J. 1999. Small Worlds: The Dynamics of Networks between Order and Randomness. Princeton University Press, Princeton.

Wawer, M., 1992. HIV prevention study. Technical report, National Institute of Child Health and Human Development.

Woolhouse, M. E. J., C. Dye, J. F. Etard, T. Smith, J. D. Charlwood, G. P. Garnett, P. Hagan, J. L. K. Hii, P. D. Ndhlovu, R. J. Quinnell, C. H. Watts, S. K. Chandiwana, and R. M. Anderson. 1997. Heterogeneities in the transmission of infectious agents: Implications for the design of control programs. Proceedings of the National Academy of Sciences of the United States of America **94**:338–342.

Yule, G. 1924. A mathematical theory of evolution based on the conclusions of Dr. J.C. Willis FRS. Philosophical Transactions of the Royal Society of London Series B-Biological Sciences **213**:21.

Table 1: Results of statistical inference for Yule model. Estimates of $\rho$ are given with 95% bootstrap confidence intervals.

| Country | sex | $n$ | $k_{min}$ | BIC | $\rho$ | (95% CI) |
|---|---|---|---|---|---|---|
| **Uganda** | women | 803 | 0 | 1070.45 | 17.04 | (12.58, 25.19) |
| | men | 621 | 1 | 1587.79 | 5.43 | ( 4.54, 5.39) |
| **Sweden** | women | 1335 | 1 | 2158.64 | 4.23 | ( 3.60, 5.21) |
| | men | 1476 | 1 | 3041.55 | 3.25 | ( 3.01, 3.63) |
| **USA** | women | 1919 | 1 | 3224.03 | 3.84 | ( 3.34, 4.55) |
| | men | 1506 | 1 | 3267.56 | 3.03 | ( 2.80, 3.32) |
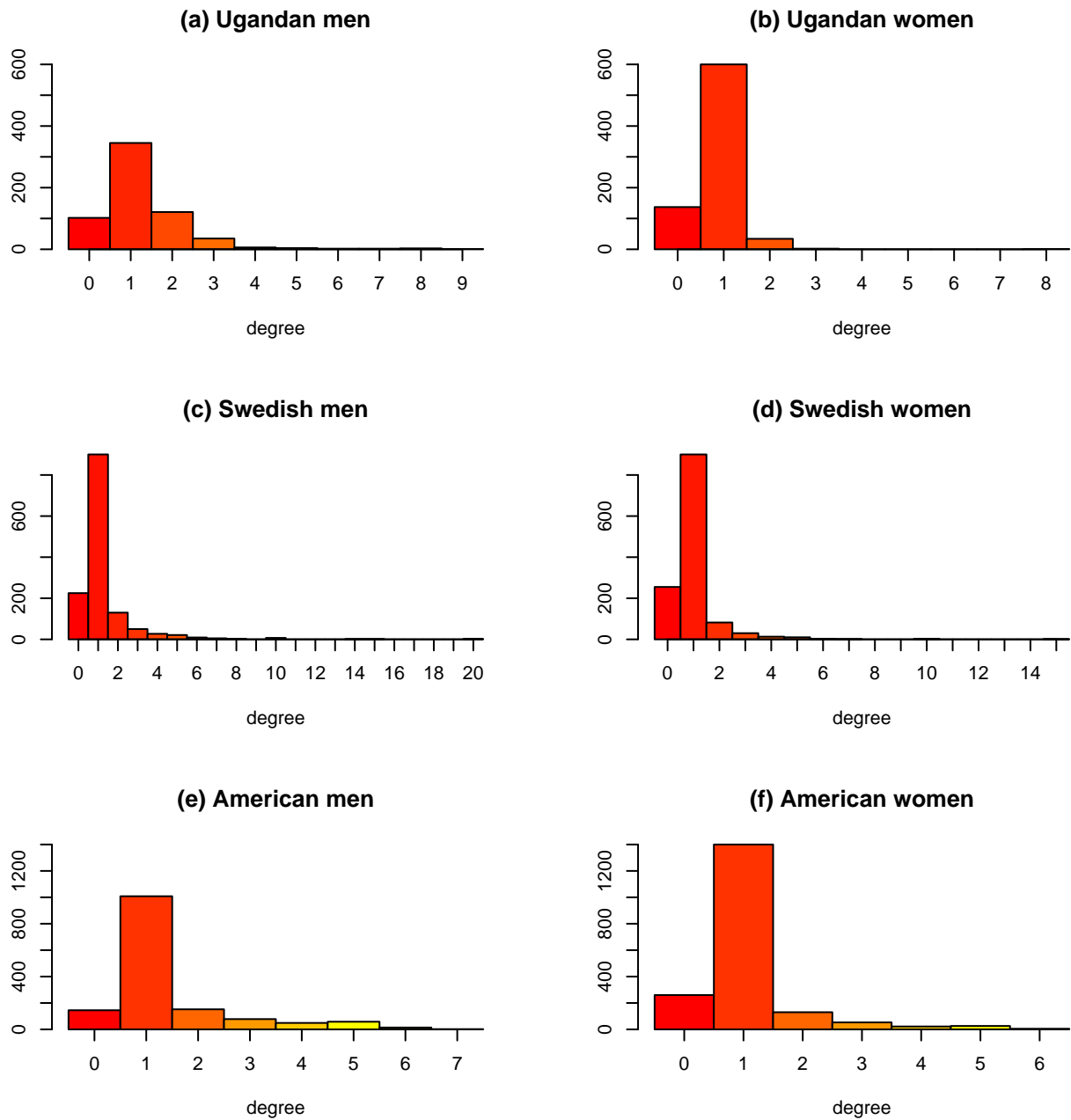
Figure 1: Reported degree distributions for men and women from the three population samples. The plots are histograms showing the absolute number of observed degree $k$ (including zeros). By row, top to bottom, the figures are: Uganda, Sweden, USA.
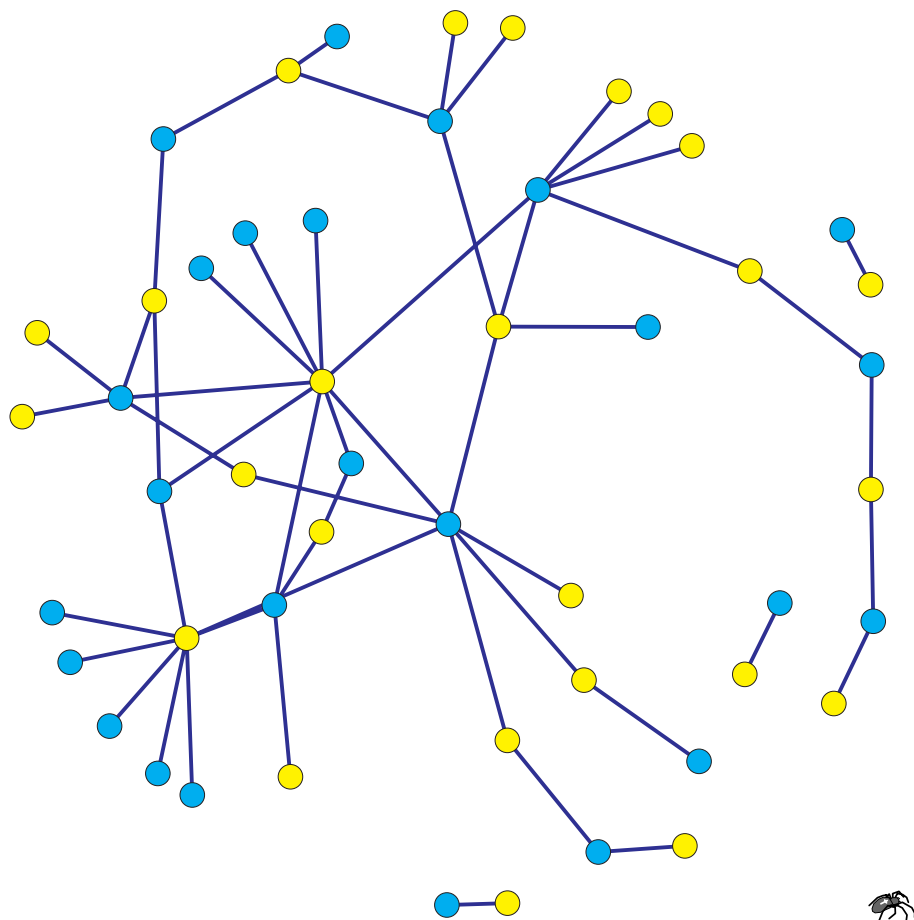
Figure 2: Simulated 50-actor, predominantly heterosexual network with infinite degree distribution. The underlying distribution is Yule with parameter $\rho = 2.5$. The algorithm producing this network is essentially that of Molloy and Reed (1995).
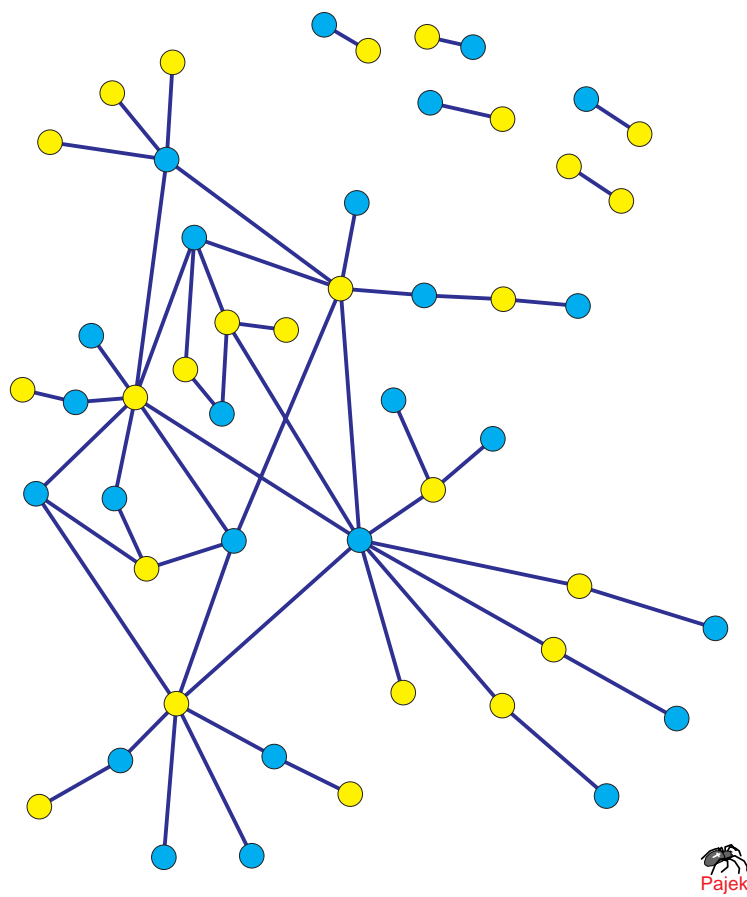
Figure 3: Simulated 50-actor, predominantly heterosexual network with infinite degree distribution and a maximal transitivity (measured by propensity to form triangles). The underlying distribution is identical to that in figure 2.
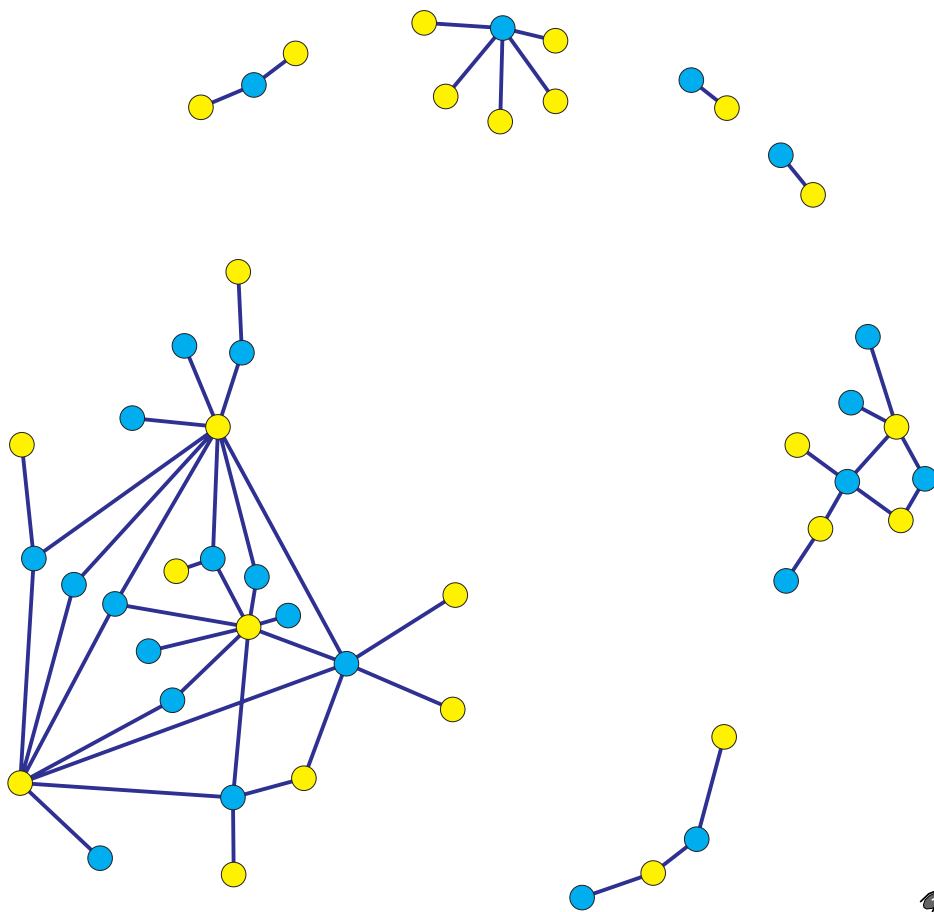
Figure 4: Simulated 50-actor, predominantly heterosexual network with infinite degree distribution and a minimal transitivity (measured by propensity to form triangles). The underlying distribution is identical to that in figure 2.