Published in final edited form as:

Stat Med. 2018 January 30; 37(2): 218–235. doi:10.1002/sim.7393.

Inference on Network Statistics by Restricting to the Network Space: Applications to Sexual History Data

Ravi Goyal^{1,*} and Victor De Gruttola²

¹Mathematica Policy Research Inc Cambridge Office, MA, U.S.A.

²Department of Biostatistics, Harvard School of Public Health. Boston, MA, U.S.A.

Abstract

Analysis of sexual history data intended to describe sexual networks presents many challenges arising from the fact that most surveys collect information on only a very small fraction of the population of interest. In addition partners are rarely identified and responses are subject to reporting biases. Typically each network statistic of interest, such as mean number of sexual partners for males or females, is estimated independently of other network statistics. There is, however, a complex relationship among networks statistics; and knowledge of these relationships can aid in addressing concerns mentioned above. We develop a novel method that constrains a posterior predictive distribution of a collection of network statistics in order to leverage the relationships among network statistics in making inference about network properties of interest. The method ensures that inference on network properties is compatible with an actual network. Through extensive simulation studies, we also demonstrate that use of this method can improve estimates in settings where there is uncertainty that arises both from sampling and from systematic reporting bias compared to currently available approaches to estimation. To illustrate the method, we apply it to estimate network statistics using data from the Chicago Health and Social Life Survey.

Keywords

Sexual Networks; Network Sampling; Reporting Error; Bayesian; Egocentric Data; Constrained Distribution

1. Introduction

Sexual relationships play a fundamental role within a society by driving birth rates, influencing social stratification, and spreading sexually transmitted diseases. Therefore, the structure of a sexual network—the collection of sexual relationships within the population—has important public policy consequences, particularly regarding public health initiatives. For example, understanding the structure of sexual networks is useful in developing preventative strategies for sexually transmitted infection control programs as mathematical

*Correspondence to: Ravi Goyal, Mathematica Policy Research Inc 955 Massachusetts Avenue, Cambridge, MA 02139, U.S.A., ravi.goyal@mail.harvard.edu.

Conflict of Interest: None declared.

models have shown that the network structure has the potential to greatly influence the rate of propagation of a disease [1, 2, 3, 4] and the efficacy of interventions designed to mitigate the epidemic [5, 6]. However, characterizing such sexual networks is challenging because, in most settings, it is only feasible to collect data on the connections between sexual partners from surveys, in particular, egocentric study designs; these refer to studies wherein respondents provide information on their partners (including such information as education level, community of residence, and age), but the respondents' partners are not surveyed. Egocentric survey designs have been implemented in many important studies on sexual behavior including National Health and Social Life Survey (NHSLs)[7], Chicago Health and Social Life Survey (CHSLs) [8], and Sex in Sweden [9]. An additional challenge in characterizing the network is dealing with reporting biases that commonly arise in self-reported data [10]. For example, even when questionnaires elicit only information on heterosexual relationships, many surveys show large discrepancies between female number of partners and their male counterparts; this phenomenon has been noted in diverse geographic settings including locations in the United States [7], Sweden [9], Zimbabwe [11, 12], Uganda [13, 11] and South Africa [11]. Even with these challenges, estimates derived from such sexual behavior surveys are often critical inputs to epidemic models such as those used to design HIV trials [6, 14, 15].

It is often useful to constrain an estimate to be consistent with the space of possible values associated with the quantities of interest; it not only helps to assure validity but also aids with interpretation. A common example is bounding an estimate of a probability between 0 and 1. For a simple network example, estimates for the total number of relationships reported by each gender in a closed heterosexual population can only have non-zero probability mass on integers between 0 and the maximum number of possible connections and on values where the two total numbers of partners (for males and females) are equal; the probability mass must be zero for all other real values. Investigators have employed various approaches to ensure valid estimates for network properties. In the modeling of heterosexual networks, often these approaches were used to handle discrepancies between the reporting of males and females. To simplify models or correct for reporting discrepancies, one common approach is to assume males and females have identical distribution of number of partners or identical proportions of males and females in risk groups associated with partnership change [16]. Another approach is to assign excess male reported partnerships to female sex workers [17]. Even when the total number of partnerships between males and females are consistent, there may be discrepancies in the reported characteristics of those relationships. For example, a study investigating sexual partnerships in Zimbabwe observed a disparity in the reported number of cohabiting and non-cohabiting partnerships between the genders and had to re-label some partnerships reported as co-habiting by females as non-cohabiting (and vice versa for males) to obtain consistent estimates to conduct simulations [12].

The proposed method ensures that the estimated posterior predictive distribution of network properties, such as the number of partners reported by males and females and mixing patterns by individual and partnership characteristics, will be zero for values that are not compatible with an actual network. The method has advantages over existing methods in that it allows investigators to avoid making potentially overly restrictive assumptions and makes use of all available data. For example, the proposed approach does not require disregarding

reported values for one gender or assuming that the degree distributions of males and females are identical. Instead our method shrinks the resulting posterior predictive distribution towards network-compatible values; the amount of shrinkage depends on the amount of available information. Furthermore, the proposed method does not require explicit constraints on network properties to ensure validity; for many network properties, such constraints can be complex and infeasible to specify [18, 19, 20, 21]. Our method automatically applies the necessary constraints, which is accomplished by extending an existing network sampling method [22].

The method we present was used in the design and monitoring of a large randomized community controlled trial, the Botswana Combination Prevention Program (BCPP), which investigates whether implementation of a combination of prevention interventions reduces HIV incidence [6]. As limited and sometimes conflicting network data exist regarding sexual networks in Botswana, we were faced with the challenge of applying statistical methods that would generate estimates of network properties that were compatible with actual networks; this was necessary for the development of epidemic simulations that served as a basis for investigation of the plausible effect sizes of the proposed interventions as well as power to detect them (see discussion for additional details). As the HIV epidemic in Botswana appears to be driven by heterosexual relationships, we were interested in estimating properties of bipartite networks, networks where individuals are divided into two disjoint sets (e.g., male and females), and connections exist only between the sets.

As the modeling choices for the likelihood function of parameters representing network properties, denoted as θ , depend on both the setting [23] and the specific technique used for network data collection, we propose a method that allows flexible specification of the model by the investigator, and only constrains the posterior predictive distribution to have zero probability mass on non-network compatible values; this constraint arises from the assumption that the data summarize properties of a network. Therefore, the resulting posterior predictive distribution will have the desired property of zero probability mass on non-network compatible values without requiring additional restrictions on selecting a model. Equation 1 presents the general formula for the posterior predictive distribution, which is referred to as the unconstrained posterior predictive distribution (P_{ucon})—as it does not necessarily conform to the complex constraints of networks:

$$P_{ucon}(\tilde{Y}|Y) \propto \int_{\theta} P(\tilde{Y}|\theta, Y) \times P(\theta|Y) d\theta, \quad (1)$$

where Y is the data collected from a population, $P(\theta|Y)$ is the posterior distribution of θ , and $P(\tilde{Y}|\theta, Y)$ is the probability model for the data. We propose to constrain the posterior predictive distribution in equation 1 by imposing that the values are network-compatible as shown below:

$$P_{con}(\tilde{Y}|Y) \propto \int_{\theta} P(\tilde{Y}|\theta, Y) \times P(\theta|Y) \times I_{\{\tilde{Y} \in \Omega_{\tilde{Y}}\}} d\theta, \quad (2)$$

where $\Omega_{\tilde{Y}}$ is the set of network compatible values for the network properties that comprise $\tilde{Y}$.

We develop two simulation studies to demonstrate the method's ability to improve network property estimates by imposing the constraint described above on the network values. The first investigates scenarios where only sampling uncertainty exists; the second investigates settings where there is uncertainty due both to sampling and to systematic reporting bias associated with the respondent's gender. As there is no standard approach to adjust for network property discrepancies, we assess our method by comparing a constrained posterior predictive distribution to its corresponding unconstrained distribution. This comparison demonstrates the value of restricting estimates to network-compatible values. Both simulation studies demonstrate significant improvement in network property estimates for the constrained posterior predictive distribution compared to the unconstrained distribution. We note that it would be easy to engineer a simulation study to produce major differences between our method that ensures valid network values and existing approaches that simply equate the male (or female) degree distribution to its complement or that disregard information entirely. For example, if the male and female degree distributions were very different, setting one equal to the other would obviously lead to bias.

We apply the proposed method to estimate network statistics, including male and female degree distributions and mixing by educational level, using data from the Chicago Health and Social Life Survey (CHSLS)—an egocentric survey containing a section on respondent's sexual history including questions regarding their sexual partners. An individual's level of education influences a range of outcomes, including health; examples include the observed association between education and cardiovascular disease in the US [24] and sexual transmitted infection (STI) risk in young South African women [25]. Though we focused on education level, the method can investigate the role of personal characteristics of interest on partner selection.

Section 2 introduces network terminology, an evaluation metric, and the network sampling method used to constrain the posterior predictive distribution. Section 3 investigates properties of the proposed method from simulation studies, while section 4 applies those methods to data from the Chicago Health and Social Life Survey. Section 5 discusses the advantages and limitations of the procedure and needs for further research. An R library, CCMnet, to use the proposed methods is available on R CRAN.

2. Methods

This section introduces network terminology used in this paper and provides details on the modeling of network properties and the metric used to compare the unconstrained posterior predictive distributions to the constrained posterior predictive distributions. It also describes the the network sampling method used to constrain the posterior predictive distribution.

2.1. Network Terminology

We represent a binary network, $g = (V, U, E)$, as a binary adjacency matrix with dimensions equal to the sizes of the disjoint sets V and U . Therefore, g has dimensions $|V| \times |U|$, where $|Z|$ (and n_Z) denote the size of set Z . Let $\{v_1, \dots, v_{|V|}\}$ and $\{u_1, \dots, u_{|U|}\}$ denote the individuals

in sets V and U , respectively; throughout this paper, set V can be thought of consisting of males while set U consists of females. Let $g[i, j] = 1$ indicate that there is a relationship between male $i \in \{v_1, \dots, v_V\}$ and female $j \in \{u_1, \dots, u_U\}$, while $g[i, j] = 0$ indicates that there is no relationship. Let $\eta(i)$ be the individuals with connections to individual i , i.e., $\eta(i) = \{j: g[i, j] = 1\}$ if $i \in V$ and $\eta(i) = \{j: g[j, i] = 1\}$ if $i \in U$.

We investigate two network properties: degree distribution and mixing patterns, though the proposed method can be applied to a broader set of properties (see appendix). The degree of individual i in set t , denoted as, $d_i^t(g)$, is the number of relationships the individual has with other individuals in the population; therefore $d_i^t(g) = \sum_j g[i, j]$ if $t = V$ and $d_i^t(g) = \sum_j g[j, i]$ if $t = U$. Let $d^t(g) = (d_1^t(g), \dots, d_{n_t}^t(g))$ represent the vector of degrees for nodes in set t , commonly referred to as a degree sequence. The degree distribution, denoted as $D^t(g)$, is a vector representing the number of these degrees over all individuals in set t ; the k^{th} entry represents the number of individuals from set t having degree k , i.e., $D_k^t(g) = \sum_{i=1}^{n_t} I_{\{d_i^t(g) = k\}}$. Let

$D[z]^t(g)$ denote the degree distribution restricted to only type t individuals in the set z , i.e., $D[z]^t_k(g) = \sum_{i \in z} I_{\{d_i^t(g) = k\}}$.

We also consider mixing in networks, which is the tendency for vertices in networks to be connected to other individuals that are like (or unlike) them based on their particular characteristics; we consider only discrete characteristics. Let $m_i^t(g)$ represent the vector of discrete characteristics for individual i of type t in network g ; this paper only considers a single characteristic, education; but the methods, formulas, and code in the CCMnet package apply to the more general setting of investigating multiple characteristics. Let

$m^t(g) = (m_1^t(g), \dots, m_{n_t}^t(g))$ be a vector containing the characteristics of all the individuals in set t . The characteristic distribution, denoted as $M^t(g)$, is a vector representing the number of individuals with these characteristics over all individuals in set t ; the k^{th} entry represents the number of individuals from set t having characteristic k , i.e., $M_k^t(g) = \sum_{i=1}^{n_t} I_{\{m_i^t(g) = k\}}$. Let

$M[z]^t(g)$ denote the characteristic distribution restricted to only type t individuals in the set z . Let $MM(g)$ be a matrix representing the mixing pattern of network g . $MM(g)$ is a $q \times q$ symmetric matrix, where q is the number of distinct characteristic patterns of interest in the population. The entry $MM_{k,l}(g)$ is the total number of relationships between any individual in set V with characteristic k and any individual in set U with characteristic l . For this paper, the homophily of network g is defined as the number of relationships $(i, j) \in E$ for which the characteristic of individual i equals the characteristics of individual j , i.e.,

$\sum_{(i,j) \in E} I_{\{m_i^V(g) = m_j^U(g)\}}$. Let $\{D^{V,1}, \dots, D^{V,p}\}$ and $\{D^{U,1}, \dots, D^{U,p}\}$ denote the degree

distributions stratified by individual-level characteristics, e.g., $D^{V,k} = D[z]^V_k$ where z is the set of characteristic k males. For clarity, we boldface the text to denote a random variable when necessary.

2.2. Modeling of Network Properties

There exists a range of choices for modeling network properties that represent the degree distribution and mixing patterns; for example, the degree distribution can be modeled as a negative binomial, geometric or Poisson distribution [23]. These modeling choices are represented as $P(\theta|Y)$ and $P(\tilde{Y}|\theta, Y)$, in equations 1 and 2. As the proposed method places little restriction on modeling choice, therefore, the investigator may freely select the most appropriate model.

In the Introduction, we discuss necessary constraints on the degree distribution for bipartite networks; the inclusion of modeling mixing patterns adds further constraints. To illustrate, consider a subpopulation of individuals who report few partners; this condition requires that the estimates of mixing between this subpopulation and others be small. This additional constraint arises from the fact that individuals in such a subpopulation must have few partners from any subpopulation. The proposed method ensures that the constrained posterior predictive distribution has zero probability mass on non-network compatible values across all network properties regardless of the model choice.

2.3. Evaluation Metric

To investigate improvement in the estimate of $\tilde{Y}$, the posterior predictive expected squared error loss function from the constrained posterior predictive distribution, $P_{con}(\tilde{Y}|Y)$ (equation 2), is compared to the unconstrained distribution, $P_{ucon}(\tilde{Y}|Y)$ (equation 1); the formula for the posterior predictive expected squared error loss function is provided in equation 3:

$$\rho(\pi, a) = \int_{\tilde{Y}'} L(\tilde{Y}', a) d\pi(\tilde{Y}'), \quad (3)$$

where $L(\tilde{Y}', a)$ is the squared error loss function and π is a posterior predictive distribution of $\tilde{Y}$ (either $P_{ucon}(\tilde{Y}|Y)$ or $P_{con}(\tilde{Y}|Y)$ in this paper). To calculate $\rho(\pi, a)$, we let a be the true value of the property for the network and $L(\tilde{Y}, a)$ equal $(\tilde{Y} - a)^2$. The posterior predictive expected squared error loss function quantifies the amount by which the estimate deviates from the actual value. Therefore, the posterior predictive expected squared error loss is a summary metric of this deviation across all values of $\tilde{Y}$ that is weighted by its posterior predictive distribution.

As $\tilde{Y}$ denotes a vector in the settings investigated in this paper, the values of the posterior predictive expected squared error loss functions associated with each component of $\tilde{Y}$ need to be combined. To quantify the posterior predictive expected squared error function for a vector $\tilde{Y}$, we average across all values comprising $\tilde{Y}$. For settings in which accuracy of estimation of some network properties is more important than that for others, the method to combine the components of $\tilde{Y}$ can be appropriately modified. As an example, simulation studies have demonstrated that properties associated with across-community mixing patterns were very important in designing cluster randomized trials for HIV prevention [6]. However, in combining the components one should carefully consider their ranges and variances and

select an appropriate normalization method; for example, a small change in the number of individuals with a high number of partnerships can have a significantly larger impact on the network structure than similar changes to the number of individuals with low number of partnerships.

2.4. Network Sampling

As stated in the Introduction, sampling from the distribution shown in equation 2 is challenging because the network constraints can be complex. In order to sample from the constrained posterior predictive distribution, $P_{con}(\tilde{Y}|Y)$, we sample networks in such a way that the values of network properties associated with the sampled networks follow $P_{con}(\tilde{Y}|Y)$.

To explain the network sampling method, we need to introduce additional notation. Let $\mathcal{G}$ denote the space of possible bipartite networks of size $n_V \times n_U$. Define $f_{\tilde{Y}}(g)$ to be the function that maps a network g to the values of network properties associated with $\tilde{Y}$. Let $C_x = \{g: f_{\tilde{Y}}(g) = x, g \in \mathcal{G}\}$. Finally, let $P_{\mathcal{G}}(G)$ be the probability mass function on network space $\mathcal{G}$ for a random variable G .

In order to sample networks from $\mathcal{G}$ such that $f_{\tilde{Y}}(G) \sim P_{con}(\tilde{Y}|Y)$, we only need the following to hold:

$$\sum_{g \in C_x} P_{\mathcal{G}}(G = g) = P_{con}(\tilde{Y} = x|Y). \quad (4)$$

We can satisfy equation 4 by defining $P_{\mathcal{G}}(G = g)$ to be the following,

$$P_{\mathcal{G}}(G = g) = \left(\frac{1}{|C_{f_{\tilde{Y}}(g)}|} \right) P_{con}(\tilde{Y} = f_{\tilde{Y}}(g)|Y), \quad (5)$$

where $|C_x|$ is the number of networks in C_x . As the constraint on the posterior predictive distribution, shown in equation 2, equivalently can be viewed as the following:

$$P_{con}(\tilde{Y}|Y) \propto \begin{cases} P_{ucon}(\tilde{Y}|Y) & \text{if } \tilde{Y} \in \Omega_{\tilde{Y}} \\ 0 & \text{else,} \end{cases} \quad (6)$$

defining,

$$P_{\mathcal{G}}(G = g) \propto \left(\frac{1}{|C_{f_{\tilde{Y}}(g)}|} \right) P_{ucon}(\tilde{Y} = f_{\tilde{Y}}(g)|Y), \quad (7)$$

would also satisfy equation 4. The probability distribution on $\mathcal{G}$ defined in equation 7 is the probability distribution underlying the network sampling model by Goyal et al. [22], which we refer to as the Congruence Class Model (CCM). As we are interested in heterosexual relationships, in the appendix, we extend the CCM to include bipartite networks; the original paper on CCM only sampled simple networks. Using this extension, we are able to sample networks from $\mathcal{G}$ such that $f_{\tilde{Y}}(G) \sim P_{con}(\tilde{Y}|Y)$ by sampling networks based on the probability distribution on $\mathcal{G}$ defined in equation 7.

3. Simulation Studies

To evaluate our method, we devise two types of simulation studies that investigate the improvement of the constrained posterior predictive distribution, equation 2, compared to the unconstrained distribution, equation 1. For each of the simulations, we generate networks with a population of 1000 individuals that includes 500 males and 500 females. The 1000 individuals are randomly assigned levels of education, (a) no high school diploma, (b) high school diploma or vocational degree or associate degree, and (c) bachelor's degree or higher. In this paper, we often refer to these three educational levels as low, middle (shortened to mid in formulas), and high. For the simulation studies, we randomly assigned individuals to these three levels, 334 individuals to low, 334 individuals to middle, and 332 individuals to high. The relationships in the sexual network are generated under the exponential random graph model (ERGM) [26]—a common network model. Specifically, we generate a bipartite network based on network statistics for degree distribution and homophily. The statistics for the degree distribution used in the simulation studies are presented in Table 1, which are close to the observed distribution from the CHSLS study of the neighborhood in Chicago described in section 4. For the simulations, we vary the value of the homophily statistic to generate networks with low and high levels of homophily.

In our simulations, a fraction of males and females were sampled at random, but at different rates based on gender. Let s_V and s_U equal the number of individuals sampled from V and U , respectively; let V_s and U_s be the sets of sampled individuals. Let $V_s^k(U_s^k)$ be the subset of sampled males (females) with characteristic k . For each sampled individual, their gender, degree, and education level were recorded as were the educational level of their partners. For a sampled individual i , let $y_i = (t, d_i^t(g), m_i^t(g), M[\eta(i)]_{low}^t(g), M[\eta(i)]_{mid}^t(g), M[\eta(i)]_{high}^t(g))$ be the observed information. For these simulations, the degree(s) of individual i 's partner(s) is not observed as this information is often unreliable; however, the appendix and code in the CCMnet package allow for investigators to incorporate such data. Also, no identifying information was recorded about the partners; therefore, it is not possible to link two individuals as sexual partners. As discussed in the introduction, this type of design is

common for collecting sexual history due to cost and privacy concerns and is the designed used in the BCPP and CHSLS studies. The first simulation uses information on degree and education; values provided by the sampled respondents are correctly reported. The second simulation uses information only on the degree, but systematic biases are generated in the reported number of partners depending on the respondent's gender.

3.1. Simulation Study 1: Sampling Variability

In this first simulation study, we evaluate the method's ability to estimate network properties in settings where there is uncertainty due to sampling a fraction of the population. To investigate the improvement of the constrained posterior predictive distribution, P_{con} , compared to the unconstrained distribution, P_{ucon} , in terms of their associated squared error loss, we first need to define our unconstrained posterior predictive distribution by specifying the network properties we are interested in modeling as well as the likelihood and prior distribution—all of which are specified by the investigator.

In this simulation study, our model consists of parameters associated with degree distributions, stratified by gender and education, i.e., $\{D^{V,low}, D^{V,mid}, D^{V,high}, D^{U,low}, D^{U,mid}, \text{ and } D^{U,high}\}$ as well as mixing patterns, i.e., $MM(g)$. We represent our observed data, $\{y_i: i \in V_s \text{ or } i \in U_s\}$, as nine vectors: the observed degree distribution of males of low education, middle education, and high education; the observed degree distribution of females of low education, middle education, and high education; and the observed number of edges to females of low, middle, and high education from males of low education, middle education, and high education. Therefore, the data used for modeling is the following:

$$Y = (D[V_s^{low}], D[V_s^{mid}], D[V_s^{high}], D[U_s^{low}], D[U_s^{mid}], D[U_s^{high}], M[V_s^{low}], M[V_s^{mid}], M[V_s^{high}]);$$

The mixing pattern data, $M[V_s^{low}]$, $M[V_s^{mid}]$, and $M[V_s^{high}]$, are the rows of the mixing pattern matrix $MM(g)$. In this simulation study, there is no reporting error; this is investigated in the next simulation.

3.1.1. Model: Degree Distributions and Mixing Patterns.—We model each of the six degree distributions for a specific gender and educational level as multinomial distributions because they are more flexible than other commonly used distributions, such as negative binomial, geometric, or Poisson distributions. In order to use the multinomial distribution for the degree distributions, a maximum degree, d_{max} , needs to be specified. In the neighborhood of interest in the CHSLS study (section 4), more than 98% of subjects reported two or fewer partners in the past year; therefore, our model for our simulation studies restrict the number of partners to be three or less, i.e., $d_{max} = 3$ (in section 4, we set d_{max} equal to the maximum observed degree in the neighborhood of interest, which is six). Let $\theta_{d^{V,low}}$ be a vector of parameters associated with probability that a male in the lowest education category has 0, 1, ..., d_{max} females partners. Let $\theta_{d^{V,mid}}$ and $\theta_{d^{V,high}}$ represent similar vectors of parameters for middle and highest education category for males,

respectively. Similarly, let $\theta_{d^{U,low}}$, $\theta_{d^{U,mid}}$ and $\theta_{d^{U,high}}$ be the parameter vectors associated with the probability that a female in the lowest, middle, or highest education category has 0, 1, ..., d_{max} males partners. The following likelihood are proposed for modeling each male degree distribution: Similarly, we propose the following likelihood for modeling each female degree distribution:

$$\mathbf{D}[\mathbf{V}_s^k] \sim \text{Multinomial}(\left|V_s^k\right|, \theta_{d^{V,k}}). \quad (8)$$

Similarly, we propose the following likelihood for modeling each female degree distribution:

$$\mathbf{D}[\mathbf{U}_s^k] \sim \text{Multinomial}(\left|U_s^k\right|, \theta_{d^{U,k}}). \quad (9)$$

We also model $\mathbf{M}[\mathbf{V}_s^{\text{low}}]$, $\mathbf{M}[\mathbf{V}_s^{\text{mid}}]$, and $\mathbf{M}[\mathbf{V}_s^{\text{high}}]$ as multinomial random variables, i.e., three different multinomial distributions represent the proportion of low, middle, and high educated male relationships with low, middle, and high educated females. Let $\theta_{m^{low}}$, $\theta_{m^{mid}}$, and $\theta_{m^{high}}$ be parameter vectors associated with the probability of a male having a relationships with a female for each educational level for low, middle, and high educational level males; therefore,

$$\mathbf{M}[\mathbf{V}_s^k] \sim \text{Multinomial}(D[V_s^k] \cdot (0, \dots, d_{max})^T, \theta_{m^k}). \quad (10)$$

For this simulation, we model the joint distribution of the nine components as a product distribution. The following likelihood is proposed for the observed data:

$$\begin{aligned} \text{Likelihood: } Y \sim & \prod_{k \in \{low, mid, high\}} \text{Multinomial}(\left|V_s^k\right|, \theta_{d^{V,k}}) \times \text{Multinomial}(\left|U_s^k\right|, \theta_{d^{U,k}}) \\ & \times \text{Multinomial}(D[V_s^k] \cdot (0, \dots, d_{max})^T, \theta_{m^k}). \end{aligned} \quad (11)$$

In order to include a more complex covariance structure between these nine multinomial models, we would need to make additional assumptions about the network generating process; the underlying theory does not preclude making such assumptions to develop a complex covariance structure. However, for clarity and to maintain the focus on the proposed methods, we used the product distribution. For the prior distribution, we use a product distribution of nine Dirichlet distributions. The following prior is proposed for observed data:

$$\text{Prior: } \theta^{M1} \sim \prod_{k \in \{low, mid, high\}} \text{Dirichlet}(\alpha_{d^V, k}^0) \times \text{Dirichlet}(\alpha_{d^U, k}^0) \times \text{Dirichlet}(\alpha_{m^k}^0), \quad (12)$$

where $\theta^{M1} = (\theta_{d^V, low}, \dots, \theta_{d^U, high}, \theta_{m^low}, \dots, \theta_{m^high})$. Let

$\alpha^0 = (\alpha_{d^V, low}^0, \dots, \alpha_{d^U, high}^0, \alpha_{m^low}^0, \dots, \alpha_{m^high}^0)$. Therefore, the unconstrained posterior predictive distribution, denoted as $P_{ucon}^{M1}(\tilde{Y}^{M1}|Y)$, is the following:

$$\begin{aligned} \text{Posterior Predictive: } \tilde{Y}|Y, \alpha^0 \sim & \prod_{k \in \{low, mid, high\}} \text{DirMult}(\alpha_{d^V, k}) \times \text{DirMult}(\alpha_{d^U, k}) \\ & \times \text{DirMult}(\alpha_{m^k}) \end{aligned} \quad (13)$$

where $\alpha_{d^V, k} = \alpha_{d^V, k}^0 + |V_s^k|$, $\alpha_{d^U, k} = \alpha_{d^U, k}^0 + |U_s^k|$, and $\alpha_{m^k} = \alpha_{m^k}^0 + D[V_s^k]$ for $k \in \{low, mid, high\}$. For the results presented in the next section, the prior distribution parameters are set to a vector of ones. The constrained posterior predictive distribution, denoted as $P_{con}^{M1}(\tilde{Y}|Y)$, is computed by constraining equation 13 to network compatible values.

3.1.2. Results.—Six different settings were simulated to evaluate the proposed method; each setting was repeated 100 times. All the simulations in each setting were based on generating networks with 500 males and 500 females and assigning, at random, individuals to the three levels of education. The relationships were generated using the ERGM based on the degree distribution in Table 1 and a homophily parameter value, which was allowed to vary as described below. Each of the six simulations sampled s_V and s_U number of males and females, respectively, from the population. From the sampled individuals, the respondent's gender, number of partners, and education level were recorded as well as the educational level of their partners. For simulations 1–3, $s_V = 100$ (20% of the males) and $s_U = 148$; while for simulations 4–6, $s_V = 50$ (10% of the males) and $s_U = 74$. These choices for s_V and s_U maintain the ratio of males to females sampled identical to those of the Westside neighborhood data in the CHSLS, which is the specific neighborhood studied in our application described below. For pairs of simulations (1,4), (2,5), and (3,6) the homophily parameter was set at 100, 200, and 300, respectively.

Figure 1 shows the results of 100 replicates of the average posterior predictive expected squared error for P_{ucon} and P_{con} across all network properties that comprise Y for the six simulations. Figure 1 shows a clear reduction in squared error loss; the average reduction for simulations 1–3 and 4–6 are approximately 30.6% and 23.6%, respectively. The average posterior predictive expected squared error of P_{ucon} or P_{con} is approximately half as large for the settings where $s_V = 100$ compare to $s_V = 50$, but there does not appear a strong relationship between the error and amount of homophily.

Figures 2 and 3 show the posterior predictive expected squared error for each network property that comprises Y for both P_{ucon} (red box plots) and P_{con} (blue box plots). Figure 2 shows the first three simulations where 20% of the males were sampled, while figure 3 shows the last three simulations where 10% of the males were sampled. The rows are results for simulations where the homophily value was set to 100, 200, and 300, respectively. The box plots in the first column represent properties $D^{U,low}$ [denoted as L0–L3], $D^{U,mid}$ [denoted as M0–M3], and $D^{U,high}$ [denoted as H0–H3], while the box plots in the second column represent properties $D^{V,low}$ [denoted as L0–L3], $D^{V,mid}$ [denoted as M0–M3], and $D^{V,high}$ [denoted as H0–H3]; the mixing properties M^{low} [denoted as LL–LH], M^{mid} [denoted as ML–MH], and M^{high} [denoted as HL–HH] are depicted in the third column. The figures demonstrate that improvement in estimation is driven mostly by improvement of the mixing properties (third column); there are more moderate improvements for the properties associated with the degree distributions (first and second column).

3.2. Simulation Study 2: Sampling Variability and Systematic Reporting Bias

In simulations conducted in this section, systematic reporting bias by gender is introduced. We use a tilde above the network properties to denote the observed value; the actual and reported values may not agree, for example, $\widetilde{d_i^I(g)}$ denotes the reported degree of respondent i . To induce reporting bias, the reported degrees of sampled individuals are either increased (for sampled males) or decreased (for sampled females) by a specified percentage, denoted as δ_m and δ_f for males and females, respectively. For example, if $\delta_m = 50\%$ and $\delta_f = 25\%$ then the number of partners reported by a male is, on average, 50% higher than his true number of partners, and the number of partners reported by a female is, on average, 25% lower than her true number. We specify a broad range for these values in the simulation study below. Specifically, for a sampled male i , the reported number of partners, $\widetilde{d_i^V(g)}$ is the following:

$$\widetilde{d_i^V(g)} = d_i^V(g) + B_i, \quad (14)$$

where

$$\mathbf{B}_i \sim \text{Binomial}\left(\frac{1}{s_V}, \delta_m \times \sum_{k \in V_s} d_k^V(g)\right), \quad (15)$$

and subject to the following:

$$\sum_{k \in V_s} \mathbf{B}_k = \delta_m \times \sum_{k \in V_s} d_k^V(g). \quad (16)$$

For a sampled female j , the reported number of partners, $\hat{d}_j^U(g)$, is the following:

$$\overline{d_j^U(g)} = d_j^U(g) - C_j, \quad (17)$$

where

$$\mathbf{C}_j \sim \text{Bionomial}\left(\frac{1}{s_U}, \delta_f \times \sum_{k \in U_s} d_k^U(g)\right), \quad (18)$$

and subject to the following:

$$\sum_{k \in U_s} \mathbf{C}_k = \delta_f \times \sum_{k \in U_s} d_k^U(g). \quad (19)$$

The following models and results assume no knowledge about the mechanism that induces systematic biases, and only make use of the reported data, which consist of respondents' genders and reported numbers of partners, i.e., the mechanisms causing reporting bias, including the values of δ_m and δ_f are assumed unknown and not used in the estimation procedure.

3.2.1. Model: Degree Distribution.—For clarity of illustration, we investigate only the network property of degree distribution stratified by gender, but the previous simulation studies and appendix provide details on including additional network properties. The data we use for modeling are $Y = (\overline{D[V_s]}, \overline{D[U_s]})$, i.e., reported degree distribution of males and females, respectively. As in the previous simulation study, we use a multinomial distribution to model the degree distributions. Let θ_{d_V} be the parameter vector associated with probability that a male has $0, 1, \dots, d_{\max}$ females partners; let θ_{d_U} represent a parameter vector for females. The following likelihoods are proposed for the degree distributions:

$$\mathbf{D}[V_s] \sim \text{Multinomial}(s_V, \theta_{d_V}), \quad (20)$$

and

$$\mathbf{D}[U_s] \sim \text{Multinomial}(s_U, \theta_{d_U}). \quad (21)$$

We model the joint distribution of the male and female degree distributions as a product distribution. The following likelihood is proposed for our data:

$$\text{Likelihood: } Y \sim \prod_{t \in \{U, V\}} \text{Multinomial}(s_t, \theta_t). \quad (22)$$

For the prior distribution, we use a product distribution of two Dirichlet distributions. The following prior is proposed for our data:

$$\text{Prior: } \theta^{M2} \sim \prod_{t \in \{U, V\}} \text{Dirichlet}(\alpha_t^0), \quad (23)$$

where $\theta^{M2} = (\theta_{dV}, \theta_{dU})$. Let $\alpha^0 = (\alpha_{dV}^0, \alpha_{dU}^0)$. Therefore, the unconstrained posterior predictive distribution, denoted as $P_{ucon}^{M2}(\tilde{Y}^{M2}|Y)$, is the following:

$$\text{Posterior Predictive: } \tilde{Y}^{M2}|Y, \alpha^0 \sim \prod_{t \in \{U, V\}} \text{DirMult}(\alpha_t^0) \quad (24)$$

where $\alpha_{dV}^0 = \alpha_{dV}^0 + \overline{D[V_s]}$ and $\alpha_{dU}^0 = \alpha_{dU}^0 + \overline{D[U_s]}$. For the results presented, α_{dV}^0 and α_{dU}^0 are set to a vector of ones. For the constrained posterior predictive distribution, denoted as $P_{con}^{M2}(\tilde{Y}^{M2}|Y)$, we again constrain equation 24 to network compatible values.

3.2.2. Results.—In this simulation study, we evaluated the proposed method in the presence of both sampling uncertainty and systematic reporting bias by gender. As mentioned above, the mechanisms causing reporting bias, including the values of δ_m and δ_f are assumed unknown and not used in the estimation procedure. Instead the proposed method shrinks the resulting posterior predictive distributions towards those that are network-compatible with the amount of shrinkage depending on the amount of available information. Once again, the simulations were based on networks with 500 males and 500 females. The relationships were generated using the ERGM based on the degree distribution in Table 1. For this simulation study $s_1 = s_2 = 100$, i.e., 100 males and females were sampled from the population. As with the previous stimulation study, we compare P_{ucon} and P_{con} in terms of their associated squared error.

The first plot in Figure 4 shows the average percent reduction in posterior predictive expected squared error for values of $\delta_m, \delta_f \in \{0, 2.5\%, 5\%, \dots, 47.5\%, 50\%\}$ when P_{con} is compared to P_{ucon} . For each pair of δ_m and δ_f 40 simulations were conducted. The plot shows that as the proposed method decreases the expected squared error of P_{con} compared to P_{ucon} by 16–28%. For example, when there is no reporting bias, i.e., $\delta_m = 0\%$ and $\delta_f = 0\%$ (lower left corner of the first plot), there is approximately a 24.6% reduction in posterior predictive expected square error in P_{con} compared to P_{ucon} ; this is similar to the improvement shown from the first simulation. The results of applying our method suggest that the larger the systematic bias, the larger the gain in accuracy. Increases in bias in both

the constrained (second plot) and unconstrained (not shown) posterior predictive distribution are, however, associated with increases in the posterior predictive expected squared error.

4. Application

4.1. Chicago Health and Social Life Survey

We investigate the association between educational level and selection of heterosexual partners using the Chicago Health and Social Life Survey (CHSLS). The CHSLS contains responses from $N = 2114$ individuals across five survey data sets; one represents the entire city of Chicago along with adjacent suburban communities ($N=890$). The other four are sampled from ethnically and socially distinctive urban neighborhoods within Chicago that have been assigned the following pseudonyms: Shoreland ($N=358$), an area with high concentration of young single heterosexuals and gay male, middle to upper-middle-class; Erlinda ($N=210$), a mixed Hispanic, working-class neighborhood; Southtown ($N=307$), a neighborhood of mixed lower, working and middle-class African-American residents; and Westside ($N=349$), a predominately working-class Mexican immigrant and Mexican-American neighborhood—see Laumann et al. [8] for a detailed description of the study and neighborhoods.

Our focus is on the structure of the heterosexual network of the Westside neighborhood. The Westside community was chosen for analysis because the sexual network was the “most geographically circumscribed” of the surveyed neighborhoods [8] (pg. 81); therefore, we expect a greater percentage of relationships is to be contained within the Westside community compared to the other surveyed neighborhoods. Since the proposed method restricts the network statistics to those that are graphical, interpretability of results improves with the degree to which the network boundary contains both individuals in each relationship. The complete CHSLS data set contains information to identify the partner’s community of residence; therefore it is possible to exclude relationships not contained within a community and make the network assumption more pertinent. However those fields are not available in the publicly available data set. In order to ensure our results are easily reproducible by all researchers, we only use the publicly available data set and provide an R library, CCMnet.

We investigate only the sexual relationships present during a 12 month period—a common time frame used in sexual behavior studies. We consider only heterosexual relationships and exclude five individuals who claimed a same-sex relationship as one of their last two partnerships. As the vast majority of individuals claimed two or fewer partners in the past 12 months this would remove all or nearly all same-sex relationships. Tables 2 and 3 show the number of partners for Westside respondents stratified by education level for males and females, respectively; Table 4 shows the mixing pattern between males and females stratified by educational level for Westside respondents.

Due to the personal nature of sexual history, not all information is available in the publicly accessible CHSLS data set. In particular, sizes of both the total population and sub-populations defined by gender and educational level are unavailable; in the following illustration of our method we consider the community to have 1000 adults (500 males and

500 females) and sub-population defined by education to have sizes proportional to the sampled individuals for each gender.

We explore the impact of constraining our sample space to network compatible values associated with degree distributions, stratified by gender and education, i.e., $\{D^{V,low}, D^{V,mid}, D^{V,high}, D^{U,low}, D^{U,mid}, \text{ and } D^{U,high}\}$ as well as mixing patterns, i.e., $MM(g)$ —the same network properties from the first simulation study; however, we increase d_{max} to 6. We also use the same formulation for the likelihood, prior, posterior and posterior predictive distributions as shown in equations 11, 12, and 13, respectively.

Based on the data presented in Tables 2, 3, and 4, there appears to exist reporting bias by gender, i.e., $\delta_m = 0$ or $\delta_f = 0$ or both. As with the simulations, we shrink the posterior predictive distributions toward network-compatible values; the amount of shrinkage depends on the amount of available information. If information is known about the reporting error mechanism, it can be incorporated into the likelihood.

Figure 5 shows the log of the posterior predictive distribution for each component of Y for both P_{ucon} (red box plots) and P_{con} (blue box plots). The first row of box plots represent the posterior predictive distributions for female degree distributions. The first set of seven box plots represent the posterior predictive distribution of $D^{U,low}$ (the degree distribution [0–6] for females with low education); in particular, the first set is the posterior predictive distribution for the estimate of the proportion of females with low education to have 0 partners. The second set of seven box plot pairs summarize the posterior predictive distributions for $D^{U,mid}$ (the degree distribution for females with medium education) and the third set of pairs are for $D^{U,high}$ (the degree distribution for females with high education). The second row of box plots represent similar posterior predictive distributions for male degree distributions, i.e., $D^{V,low}, D^{V,mid}, \text{ and } D^{V,high}$. The posterior predictive distributions for mixing properties are shown in the bottom row; the posterior predictive distribution for M^{low} (proportion of low education males with each category [low, medium, high] of females) are represented by the first three paired box plots; posterior predictive distribution for M^{mid} (proportion of medium education males with each category [low, medium, high] of females) are represented by the second paired set of three box plots; and posterior predictive distribution for M^{high} (proportion of high education males with each category [low, medium, high] of females) are represented by the last three paired box plots.

Figure 5 demonstrates the shrinkage toward network-compatible values discussed in the Introduction. The figure shows that the overall mean number of partners for low and medium educated males is lower (box plots in the second row associated with $D^{V,low}$ and $D^{V,mid}$ are lower for degrees 2–6 in the constrained distribution), while the overall mean number of partners for low, medium, and high educated females is higher (box plots in the first row are greater for degrees 2–6 in the constrained distribution) when these properties are constrained by being required to correspond to network compatible values. This shift in mean number of partners is necessary as the males reported greater numbers of partners than do the females, but the quantities must be equal in order to be represented by a network. It also shows the variability in the amount of shift, which reflects the available information. The estimates for having 3, 4, 5 or 6 partners under the constraint are greater for high educated males and

females compared to the unconstrained model (see distributions associated with properties $D^{V,high}$ and $D^{U,high}$). The estimates for mixing patterns for P_{ucon} and P_{con} are similar (the box plots in each pair in the bottom row are close to each other).

5. Discussion

In order to investigate the spread of infectious diseases, researchers collect data regarding sexual networks using novel methods (e.g. cell phone record) and multiple sources. To be maximally useful, flexible statistical inference techniques must be developed to analyze such data. In this paper, we develop a flexible method that permits the inclusion of knowledge about the data collection process and the data themselves and ensures that resulting estimates conform to the complex set of network constraints. The constraints ensure that the posterior predictive distribution of network properties has zero probability mass on all values that are not compatible with an actual network. Therefore, the method assures that posterior predictive estimates will conform to the complex constraints imposed by a network structure.

The method provides additional benefits besides those described above. Through simulation studies, we demonstrate that the proposed method can improve estimation of bipartite network property estimates in the presence of uncertainty that arises from both sampling and systematic reporting bias. The amount of improvement depends on many factors including the choice of model for the network properties, the sampling scheme, and the type of reporting bias. In the simulation studies, we assumed no knowledge about the causes of the systematic reporting bias. Should an investigator have this additional information, however, it could be incorporated into the likelihood function.

The method presented in this paper played a critical role in designing the BCPP trial described in the Introduction. The method enabled our colleagues and us to conduct agent-based network simulation studies, which took into consideration available information on sexual network characteristics, such as mixing within and between communities as well as target coverage levels for different prevention modalities in the combination prevention package. As network-based interventions grow in popularity and the benefits for these type of simulations increase, there will be additional need for producing estimates of network properties that are consistent with network constraints.

Acknowledgments

This research is supported by grants from the National Institute of Health (R37 AI-51164). We also acknowledge the Chicago Health Study for the data set.

Funded by

- National Institute of Health. Grant Number: R37AI-51164
- Chicago Health Study

6.: Appendix: Methods for Bipartite Network Construction

Let C_g and $|C_g|$ denote the congruence class containing network g and the number of networks in C_g , respectively. As closed form expressions for $|C_g|$ are not known, sampling from $\mathcal{G}$ according to the probability mass function in equation (2) is performed by a Metropolis-Hastings algorithm (MH)—a type of Markov Chain Monte Carlo (MCMC) procedure. The MH algorithm exploits relationships between the congruence classes to calculate ratios of volume factors. Given a current network g_t , the algorithm proposes a network, p_t , based on g_t by toggling the existence of an edge at random. If the proposed network is accepted, based on equation (25) below, $g_{t+1} = p_t$ otherwise $g_{t+1} = g_t$ see [22] for derivation of equation (25). The algorithm continues for a set number of proposals and produces an irreducible Markov chain. The equation for the acceptance probability for the MH algorithm is the following:

$$P(\text{Accept } gp_{t+1}|g_t) = \min\left(1, \frac{f(C_{gp_{t+1}}, C_{g_t})}{f(C_{g_t}, C_{gp_{t+1}})} \times \frac{P_C(C_{gp_{t+1}})}{P_C(C_{g_t})}\right). \quad (25)$$

where $f(C_g, C_h)$ is the average number of networks in C_h that are valid proposals from an element $g \in C_g$. The nonstandard acceptance probability formula in the MH algorithm arises because the CCM's focus on congruence classes. Equation (25) is identical to the acceptance probability derived in [22] except in that the congruence classes are defined on the space of bipartite networks.

Once the network space and network proposal method have been selected, only the functions $P_C(C_g)$ and $f(C_g, C_h)$ need to be specified to compute equation 25. In all scenarios, the network space consists of all bipartite networks of a particular size N . As mentioned above, edge toggling is used to propose networks (described in [22]), but is restricted so that only edges with opposite types of nodes can be endpoints—a requirement of bipartite networks. As the probability mass function, P_C , on congruence classes is set by the investigator, in this section we only derive $f(C_g, C_h)$ for bipartite network properties that are estimable from egocentric data and are of interest. Let C_g and C_h denote the congruence classes for g and h . Let the edge, (i, j) , between node i and node j be the connection that is toggled to move from g to h and back. Without loss of generality, let $(i, j) \in h$ but $(i, j) \notin g$.

Describing the method requires defining terminology and notation. Let nodes be designated as either type 1 or 2 with edges only between distinct node types. Let vector $M^t(g)$ denote the covariate pattern distribution for nodes of type $t \in \{1, 2\}$ of a graph g , where the i^{th} entry of $M^t(g)$, $M_i^t(g)$, is the number of type t nodes with covariate pattern i . Let $m^t(g)$ represent the vector of covariate patterns for type t nodes in network g , where the i^{th} entry, $m_i^t(g)$, is the covariate pattern of the i^{th} type t node. Let $MM(g)$ be a matrix representing the mixing pattern of graph g . The entry $MM_{k,l}(g)$ is the number of edges from a type 1 node with covariate pattern k to a type 2 node with covariate pattern l . We use slightly different notation for similar quantities describing nodal degrees. Let $D^t(g)$ and $d^t(g)$ represent the

degree distribution and degree sequence for nodes of type t . $DMM(g)$ represents degree mixing matrices, where entry $DMM_{ij}(g)$ is the number of edges from a type 1 node of degree i to a type 2 node of degree j .

6.1. Topological Properties

The equations for bipartite networks for network density are unchanged from the uni-mode case described in Goyal et al. [22], therefore we provide only details for degree distribution and the degree mixing matrix as higher-order bipartite network properties are not estimable from egocentric sample data.

Degree Distribution.

For degree distribution, congruence classes are sets of networks with identical type 1 node and type 2 node degree distributions. Thus, networks g_1 and g_2 are in the same congruence class if and only if $D_k^2(g_1) = D_k^2(g_2)$ and $D_k^1(g_1) = D_k^1(g_2) \forall k$. Recall the only difference between the graph g and h is edge $(i, j) \in h$ but $(i, j) \notin g$.

The number of edge toggles from a graph h to any graph in C_g is equal to the number of edges in h that have endpoint degrees of $d_i^1(h)$ and $d_j^2(h)$, $DMM_{d_i^1(h), d_j^2(h)}^{(h)}$. As $f(C_h, C_g)$ is equal to the average number of such edge toggles for all $h' \in C_h$

$$f(C_h, C_g) = \frac{1}{|C_h|} \sum_{h' \in C_h} DMM_{d_i^1(h), d_j^2(h)}^{(h')}. \quad (26)$$

Following arguments from Newman [5], the probability that a node's neighbor will have degree k is proportional to $k * D_k$, therefore,

$$f(C_h, C_g) \approx \frac{D_{d_i^1(h)}^1(h) * d_i^1(h) * D_{d_j^2(h)}^2(h) * d_j^2(h)}{\sum_l D_l^1(h) * l}. \quad (27)$$

The number of edge toggles from a graph g to any graph in C_h is equal to the number of possible non-loop edges with the type 1 node endpoint degree as $d_i^1(g)$ and type 2 node endpoint degree as $d_j^2(g)$ minus the number of edges that will generate a multi-edge; the number of such toggles that will result in a multi-edge is $DMM_{d_i^1(g), d_j^2(g)}^{(g)}$. Therefore,

$$f(C_g, C_h) = D_{d_i^1(g)}^1(g) * D_{d_j^2(g)}^2(g) - \alpha_1, \quad (28)$$

where,

$$\alpha_1 = \frac{1}{|C_g|} \sum_{g' \in C_g} DMM_{d_i^1(g), d_j^2(g)}(g') \quad (29)$$

$$\approx \frac{D_{d_i^1(g)}^1(g) * d_i^1(g) * D_{d_j^2(g)}^2(g) * d_j^2(g)}{\sum_l D_l^1(g) * l}, \quad (30)$$

by following the same arguments from Newman [5].

Degree Mixing and Degree Distribution.

We consider a partition of $\mathcal{G}$ such that networks g_1 and g_2 are in the same congruence class if and only if $D_x^1(g_1) = D_x^1(g_2)$ and $D_x^2(g_1) = D_x^2(g_2) \forall x$ and $DMM_{x,y}(g_1) = DMM_{x,y}(g_2) \forall x, y$.

Since the degree mixing matrix uniquely determines the degree distributions for type 1 and type 2 nodes, an identical partition is defined when networks g_1 and g_2 are in the same congruence class if and only if $DMM(g_1) = DMM(g_2)$. Thus, the probability mass function can be defined using the degree mixing matrix. Toggling the edge (x, y) modifies the degree mixing matrix by changing the entry corresponding to (x, y) along with the all the entries corresponding to the edges connected to nodes x and y . The number of edge toggles from a graph $h \in C_h$ to any graph in C_g is equal to $DMM_{d_i^1(h), d_j^2(h)}^1$ multiplied by the probability that

the neighbors of the edge's endpoints have appropriate degrees. Similar logic holds for $f(C_g, C_h)$, leading to the following expressions for $f(C_g, C_h)$ and $f(C_h, C_g)$:

$$f(C_h, C_g) = DMM_{d_i^1(h), d_j^2(h)}^1(h) * \beta_{(i,j)}^0(h). \quad (31)$$

$$f(C_g, C_h) = \left[D_{d_i^1(g)}^1(g) * D_{d_j^2(g)}^2(g) - \alpha_2 \right] * \beta_{(i,j)}^1(g). \quad (32)$$

where $\alpha_2 = DMM_{d_i^1(g), d_i^1(g)}^1(g)$,

$$\beta_{(i,j)}^s(z) \approx \frac{\Pi_k \left(\frac{DMM_{d_i(z),k}(z) * |E_z| - I_{\{d_j(z)=k\}} * s}{n_i^k - I_{\{d_j(z)=k\}} * s} \right)}{\left(\frac{\sum_k DMM_{d_i(z),k}(z) * |E_z| - I_{\{d_j(z)=k\}} * s}{d_i(z) - s} \right)} \quad (33)$$

$$* \frac{\Pi_k \left(\frac{DMM_{k,d_j(z)}(z) * |E_z| - I_{\{d_i(z)=k\}} * s}{n_j^k - I_{\{d_i(z)=k\}} * s} \right)}{\left(\frac{\sum_k DMM_{k,d_j(z)}(z) * |E_z| - I_{\{d_i(z)=k\}} * s}{d_j(z) - s} \right)}$$

and n_i^k and n_j^k denote the number of neighbors of i and j with degrees equal to k .

6.2. Nodal Covariates

Nodal Covariate Mixing and Degree Distribution.

To model degree distributions and mixing patterns based on nodal covariates, we first let p be the number of distinct nodal covariate patterns of interest in the population. A common scenario is one in which not only mixing patterns between covariate patterns but also the degree distributions of type 1 and type 2 nodes, $\{D^{1,1}, \dots, D^{1,p}\}$ and $\{D^{2,1}, \dots, D^{2,p}\}$, are observed for each covariate pattern. The following approach can be simplified for settings wherein individual covariate pattern degree distributions are not observed. In order to incorporate covariate information, knowledge of the number of individuals with covariate pattern k for each node type, M_k^1 and M_k^2 , is required for each k .

For nodal covariate mixing and degree distribution, the congruence classes contain networks with identical numbers of nodes, degree distributions, and nodal covariate mixing matrices. Thus, networks g_1 and g_2 are in the same congruence class if and only if $D_x^{1,k}(g_1) = D_x^{1,k}(g_2)$ and $D_x^{2,k}(g_1) = D_x^{2,k}(g_2) \forall x, k$ and $MM_{k,l}(g_1) = MM_{k,l}(g_2) \forall k, l$.

Let the matrix entry $DMM_{x,y}^{k,l}(g)$ represent the number of edges where type 1 node has covariate pattern k and degree x , while the type 2 node has covariate pattern l and degree y in a graph g . For each covariate degree distribution, one entry in the mixing matrix is fixed. Therefore, given degree distribution estimates for each of the covariate patterns, the probability mass function can only be specified for the degree distributions and the entries above the diagonal in the mixing matrix, and our calculations below are restricted to when $k = l$. Using the setup from the previous section, we let the edge set of g and h be identical except that $(i, j) \notin g$ and $(i, j) \in h$, and let nodes i and j have covariate patterns m_i^1 and m_j^2 ,

respectively. The number of edge toggles from a graph $h \in C_h$ to any graph in C_g is equal to the number of edges in h where one endpoint has degree $d_i^1(h)$ and type m_i^2 and the other endpoint, $d_j^2(h)$ and type m_j^2 . The proportion of edges where both endpoints are specified as type m_i^1 and type m_j^2 compared to edges where one endpoint is required to be of type m_i^1 is $\frac{MM_{i,j}}{\sum_z MM_{i,z}}$. Using arguments similar to those above, we can calculate the average value across all $g' \in C_{g'}$ of the degree mixing matrix at index (x, y) where only edges between nodes of type 1 with covariate k and nodes of type 2 with covariate l are considered; this quantity is denoted as $E(DMM_{x,y}^{k,l} | C_{g'})$ and expressed below:

$$E(DMM_{x,y}^{k,l} | C_{g'}) \approx \frac{D_x^{1,k,l}(g') * x * D_y^{2,k,l}(g') * y}{\sum_z D_z^{1,k,l}(g') * z} \quad (34)$$

where,

$$D^{(1,k,l)}(g') = D^{1,k}(g') * \frac{MM_{k,l}}{\sum_z MM_{k,z}} \quad (35)$$

and

$$D^{(2,k,l)}(g') = D^{2,l}(g') * \frac{MM_{k,l}}{\sum_z MM_{z,l}}. \quad (36)$$

Thus,

$$f(C_h, C_g) = E(DMM_{d_i^1(h), d_j^2(h)}^{m_i^1, m_j^2} | C_h) \quad (37)$$

and,

$$f(C_g, C_h) = M_{m_i^1}^1 * M_{m_j^2}^2 - \alpha_3 \quad (38)$$

$$\text{Where } \alpha_3 = E(DMM_{d_i^1(g), d_j^2(g)}^{m_i^1, m_j^2} | C_g).$$

Nodal Covariate Mixing, Degree Mixing, and Degree Distribution.

In a similar fashion as above the proposed method can be extended to include degree mixing.

References

- [1]. Ghani A, Swinton J, Garnett G. The role of sexual partnership networks in the epidemiology of gonorrhea. *Sexually Transmitted Diseases* 1997; 24(1):45–56. [PubMed: 9018783]
- [2]. Ghani A, Garnett G. Risks of acquiring and transmitting sexually transmitted diseases in sexual partner networks. *Sexually Transmitted Diseases* 2000; 27(10):579–587. [PubMed: 11099073]
- [3]. Kretzschmar M, Morris M. Measures of concurrency in networks and the spread of infectious disease. *Mathematical Biosciences* 1996; 133:165–195. [PubMed: 8718707]
- [4]. Morris M, Kurth A, Hamilton D, Moody J, Wakefield S. Concurrent partnerships and hiv prevalence disparities by race: Linking science and public health practice. *American Journal of Public Health* 2009; 99(6):1023–1031. [PubMed: 19372508]
- [5]. Newman M Assortative mixing in networks. *Physical Review Letters* 2002; 89(20):208701. [PubMed: 12443515]
- [6]. Wang R, Goyal R, Lei Q, Essex M, De Gruttola V. Sample size considerations in the design of cluster randomized trials of combination hiv prevention. *Clinical Trials* 2014;:1740774514523351.
- [7]. Laumann EO. The social organization of sexuality: Sexual practices in the United States. University of Chicago Press, 1994.
- [8]. Laumann E, Ellingson S, Mahay J, Paik A, Youm Y. The sexual organization of the city. University of Chicago Press, 2004.
- [9]. Lewin B, Fugl-Meyer K, Helmius G, Lalos A, Månsson S. Sex in sweden. Stockholm: National Institute of Public Health 1996;.
- [10]. Helleringer S, Kohler H, Kalilani-Phiri L, Mkandawire J, Armbruster B. The reliability of sexual partnership histories: implications for the measurement of partnership concurrency during surveys. *AIDS (London, England)* 2011; 25(4):503.
- [11]. Todd J, Cremin I, McGrath N, Bwanika J, Wringe A, Marston M, Kasamba I, Mushati P, Lutalo T, Hosegood V. Reported number of sexual partners: comparison of data from four african longitudinal studies. *Sexually transmitted infections* 2009; 85(Suppl 1):i72–i80. [PubMed: 19307344]
- [12]. Goodreau S, Cassels S, Kasprzyk D, Montañó D, Greek A, Morris M. Concurrent partnerships, acute infection and hiv epidemic dynamics among young adults in zimbabwe. *AIDS and Behavior* 2012; 16(2):312–322. [PubMed: 21190074]
- [13]. Morris M, Kretzschmar M. A microsimulation study of the effect of concurrent partnerships on the spread of hiv in uganda. *Mathematical Population Studies* 2000; 8(2):109–133.
- [14]. Boily M, Masse B, Alsallaq R, Padian N, Eaton J, Vesga J, Hallett T. Hiv treatment as prevention: Considerations in the design, conduct, and analysis of cluster randomized controlled trials of combination hiv prevention. *PLoS Med* 2012; 9(7):e1001250. [PubMed: 22807657]
- [15]. Boren D, Sullivan P, Beyrer C, Baral S, Bekker L, Brookmeyer R. Stochastic variation in network epidemic models: implications for the design of community level hiv prevention trials. *Statistics in medicine* 2014;.
- [16]. Cori A, Ayles H, Beyers N, Schaap A, Floyd S, Sabapathy K, Eaton JW, Hauck K, Smith P, Griffith S, et al. Hptn 071 (popart): a cluster-randomized trial of the population impact of an hiv combination prevention intervention including universal testing and treatment: mathematical model. *PloS one* 2014; 9(1):e84511. [PubMed: 24454728]
- [17]. Brewer DD, Potterat JJ, Garrett SB, Muth SQ, Roberts JM, Kasprzyk D, Montano DE, Darrow WW. Prostitution and the sex discrepancy in reported number of sexual partners. *Proceedings of the National Academy of Sciences* 2000; 97(22):12385–12388.
- [18]. Blitzstein J, Diaconis P. A sequential importance sampling algorithm for generating random graphs with prescribed degrees. *Internet Mathematics* 2010; 6(4):487–520.
- [19]. Newman M Spread of epidemic disease on networks. *Physical review E* 2002; 66(1):016128.
- [20]. Amanatidis Y, Green B, Mihai M. Graphic realizations of joint-degree matrices. Unpublished 2008;.

- [21]. Goyal R, Blitzstein J, De Gruttola V. Simulating bipartite networks to reflect uncertainty in local network properties. Harvard University Biostatistics Working Paper Series 2013; Working Paper 137.
- [22]. Goyal R, Blitzstein J, DeGruttola V. Sampling networks from their posterior predictive distribution. Network Science 2014; 2(01):107–131. [PubMed: 25339990]
- [23]. Handcock M, Jones J. Likelihood-based inference for stochastic models of sexual network formation. Theoretical Population Biology 2004; 65(4):413–422. [PubMed: 15136015]
- [24]. Winkleby M, Jatulis D, Frank E, Fortmann S. Socioeconomic status and health: how education, income, and occupation contribute to risk factors for cardiovascular disease. American journal of public health 1992; 82(6):816–820. [PubMed: 1585961]
- [25]. Pettifor A, Levandowski B, MacPhail C, Padian N, Cohen M, Rees H. Keep them in school: the importance of education as a protective factor against hiv infection among young south african women. International journal of epidemiology 2008; 37(6):1266–1273. [PubMed: 18614609]
- [26]. Frank O, Strauss D. Markov graphs. Journal of the American Statistical Association 1986; 81:832–842.

Figure 1. Simulation Study 1: Sampling Variability. The average of posterior predictive expected squared error of P_{ucon} and P_{con} across all parameters that comprise Y for each of the 6 simulations.

Figure 2.

Simulation Study 1: Sampling Variability. The posterior predictive expected squared error loss function of P_{ucon} (red box plots) and P_{con} (blue box plots) for each of the parameters comprising θ . The results are for simulations 1–3, where 20% of the males were sampled. The rows are results for simulations where the homophily value was set to 100, 200, and 300, respectively. The box plots in the first column represent network properties $D^{U,low}$ [denoted as L0–L3], $D^{U,mid}$ [denoted as M0–M3], and $D^{U,high}$ [denoted as H0–H3], while the box plots in the second column represent properties $D^{V,low}$ [denoted as L0–L3], $D^{V,mid}$ [denoted as M0–M3], and $D^{V,high}$ [denoted as H0–H3]; the mixing properties M^{low} [denoted as LL–LH], M^{mid} [denoted as ML–MH], and M^{high} [denoted as HL–HH] are depicted in the third column.

Figure 3.

Simulation Study 1: Sampling Variability. The posterior predictive expected squared error loss function of P_{ucon} (red box plots) and P_{con} (blue box plots) for each of the parameters comprising θ . The results are for simulations 4–6, where 10% of the males were sampled. The rows are results for simulations where the homophily value was set to 100, 200, and 300, respectively. The box plots in the first column represent network properties $D^{U,low}$ [denoted as L0–L3], $D^{U,mid}$ [denoted as M0–M3], and $D^{U,high}$ [denoted as H0–H3], while the box plots in the second column represent properties $D^{V,low}$ [denoted as L0–L3], $D^{V,mid}$ [denoted as M0–M3], and $D^{V,high}$ [denoted as H0–H3]; the mixing properties M^{low} [denoted as LL–LH], M^{mid} [denoted as ML–MH], and M^{high} [denoted as HL–HH] are depicted in the third column.

Figure 4.

Simulation Study 2: Sampling Variability and Systematic Reporting Bias. The first plot shows the reduction in posterior predictive expected squared error between the unconstrained and constrained distributions. The second plot shows the posterior predictive expected squared error loss for the constrained posterior predictive distribution.

Figure 5.

Application. Box plots depicting the log of the posterior predictive distribution for each component of Y for the constrained to unconstrained posterior predictive distributions. The first row of box plots represent the posterior predictive distribution for female degree distributions; the second row represents similar parameters for male degree distributions. The posterior predictive distribution associated with mixing properties are shown in the bottom row.

Table 1.

Degree distribution for simulation studies

Degree	0	1	2	3	Mean Degree
Both Genders	29	166	7	2	0.91

Table 2.

Westside Degree Distribution for Males by Education

Education Level	Degree	0	1	2	3	4	5	6	Total
	< High School	6	53	6	1	1	0	0	67
	High School or Vocational or Associate	12	37	8	1	1	0	0	59
	>= Bachelors	2	9	1	0	0	0	0	12
Total		20	99	15	2	2	0	0	138

Table 3.

Westside Degree Distribution for Females by Education

Education Level	Degree	0	1	2	3	4	5	6	Total
	< High School	16	105	3	0	0	0	0	124
	High School or Vocational or Associate	12	54	2	1	0	0	0	69
	>= Bachelors	1	7	2	0	0	0	1	11
Total		29	166	7	1	0	0	1	204

Table 4.

Westside Mixing Patterns by Education

		Female		
		< High School	High School or Vocational or Associate	>= Bachelors
Male	< High School	110.00	39.00	2.00
	High School or Vocational or Associate	58.00	63.00	13.00
	>= Bachelors	3.00	10.00	8.00