Harvard University
Harvard University Biostatistics Working Paper Series

Year 2013 *Paper* 137

Simulating Bipartite Networks to Reflect Uncertainty in Local Network Properties

Ravi Goyal* Joseph Blitzstein†
Victor De Gruttola‡

*Harvard University, rgoyal@fas.harvard.edu
†Harvard University, blitzstein@stat.harvard.edu
‡Harvard University, degrut@hsph.harvard.edu

Simulating Bipartite Networks to Reflect Uncertainty in Local Network Properties

Ravi Goyal
Department of Biostatistics, Harvard School of Public Health
and
Joseph Blitzstein
Department of Statistics, Harvard University
and
Victor De Gruttola
Department of Biostatistics, Harvard School of Public Health

December 12, 2013

Abstract

Computational methods are presented for generation of bipartite networks that are consistent with given probability distributions for important local network properties, including degree distribution and mixing patterns. To our knowledge, the proposed methods are the first to allow specification of the probability distributions of network properties rather than solely mean estimates for these properties. Our focus of interest is isolation of the effect of mixing patterns, which is achieved by constructing collections of bipartite networks with distinct probability distributions on parameters characterizing mixing patterns given a prescribed degree sequence. The proposed methods have been used in designing a large HIV randomized community prevention trial in Botswana; they were developed to investigate the implications of various spatial and degree mixing patterns for statistical power of proposed interventions. This setting is useful to illustrate our methods as modeling sexual disease transmission is an important research area that often has limitations of available information which could lead to wide probability distributions on network properties. In addition, we analyze various spatial and degree mixing patterns to investigate the complex relationship between global network topology and impact of the proposed HIV intervention.

Keywords: Network Generation, Disease Modeling, Degree Mixing, Spatial Mixing

1 Introduction

Bipartite networks are used to represent a myriad of complex systems such as sexual disease transmissions in heterosexual populations (Morris et al., 2009; Wang et al., 2013), social affiliations (Davis et al., 1941; Galaskiewicz, 1985), scientific collaborations (Ramasco et al., 2004), patient-provider interactions (Landon et al., 2012), and relationships between disorders and disease genes (Goh et al., 2007). Development of accurate representations, however, is complicated by the fact that network data are often incomplete or even incorrect, particularly in settings involving social interactions (Kossinets, 2006). The mechanisms that lead to such problems include the nature of study designs themselves (e.g. ego-centric and fixed choice), missing and inaccurate responses, and network boundary specification (i.e. inclusion and exclusion conditions for the study population). Kossinets (2006) and Shalizi and Rinaldo (2013) have shown that estimates for network properties can strongly depend on the type(s) of mechanism(s) leading to misspecified network data and the extent of misspecification. Under certain settings, methods exist to correct for biases in estimates based on sampled network data; however, these approaches can lead to large variability in network property estimates (Kolaczyk, 2009).

To accommodate varying degrees of uncertainty, we develop methods to generate bipartite networks of a fixed size that are consistent with a given probability distribution on local network parameters, e.g. degree distribution and mixing patterns. Current network generation methods for bipartite networks are based on specification of only the mean estimates of network properties (see, for example, Wang et al. (2012)). Our particular focus is development of a method to generate bipartite networks given a prescribed degree sequence and a probability distribution for parameters representing degree and non-degree mixing patterns. Our focus on mixing patterns reflects their presence in many real-world networks representing a range of systems from biological to technological (Newman, 2010). Furthermore, mixing has the ability to significantly alter processes operating on networks (Onnela et al., 2007; Newman, 2002); therefore, in certain settings (e.g. epidemic disease control) it is useful to generate networks that are consistent with an estimated distribution of network parameters in order to ensure reliable conclusions regarding the process operating on the network.

One important research area where models typically suffer from imprecise estimates due to misspecification and incomplete network data is in the modeling of sexual disease epidemics.

Such epidemic models are important in designing and evaluating the results of intervention studies to prevent or control disease (Boily et al., 2012; Wang et al., 2013). The methods we propose have already been utilized in designing a large HIV randomized community trial in Botswana, referred to the Botswana Combination Prevention Project (BCPP) (Wang et al., 2013). Below, we use the BCPP study to illustrate these methods. In particular, the methods were utilized in simulation studies to assess statistical power of the BCPP trial by modeling both epidemic spread and the impact on it of the study interventions; see Boily et al. (2012) and Wang et al. (2013) for additional details. As estimates of statistical power are sensitive to mixing patterns associated with the sexual contact network, we needed to address ways to take into account limitations and uncertainty in available network information.

Section 2 develops a method to model bipartite networks for a fixed degree sequence that incorporates the probability distribution associated with mixing parameter estimates. The supplement extends this method to settings where parameters characterize density, degree distribution and mixing patterns without requiring a prescribed degree sequence. Section 3 provides a summary of the BCPP trial in order to demonstrate the usefulness of the methods in designing network-based interventions. Section 4 illustrates the methods by generating bipartite networks based on spatial mixing patterns between two district communities. Section 5 investigates degree assortativity–one particularly important example of a sexual contact network property for which it can be difficult to obtain information to support estimation. We describe the ability of our methods to handle influential network properties for which estimates are not available by generating networks under a diffuse distribution–as such networks span a wide range of degree mixing patterns. Section 6 provides a discussion and further areas for research.

2 Network Generation

Our discussion of bipartite network construction extends the framework presented in Goyal et al. (2013), which provided a method to model degree distribution and both degree and non-degree mixing patterns for one-mode networks. The framework allowed different measures of uncertainty for each parameter as does the extension considered here. This section focuses on

a method to generate bipartite networks with prescribed degree sequence that are consistent with a probability distribution on degree or non-degree mixing parameters; see the supplement for extensions.

Describing the method for generating bipartite networks requires defining terminology and notation. Let nodes be designated as either type 1 or 2; as the networks are bipartite, edges only exist 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 $M^t(g)$, $M^t_i(g)$, 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, $m^t_i(g)$, 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. We use the notation $DMM(g)$ to represent degree mixing matrices, where entry $DMM_{i,j}(g)$ is the number of edges from a type 1 node of degree i to a type 2 node of degree j.

The collection of generated networks form a subset of networks from the space, $\mathscr{G}$, of bipartite graphs with n nodes and a prescribed degree sequence. To construct such a collection, we begin by partitioning $\mathscr{G}$ into congruence classes, $\{C_i\}_{i=1}^{\lambda}$, such that each network in the congruence class has the same values for all entries of the (degree) mixing matrix. Let C_g represent the congruence class containing network g. Networks g and h reside in the same congruence class if and only if $MM(g) = MM(h)$. Let $P_{\mathscr{C}}$ denote a probability mass function on the space of congruence classes where $P_{\mathscr{C}}(C_g)$ represents the probability of sampling a network from the congruence class C_g. Each network g in $\mathscr{G}$ is assigned a probability, $P_{\mathscr{G}}(g)$, of being selected into the collection based solely on the congruence class the network resides, i.e. if $C_g = C_h$ then $P_{\mathscr{G}}(g) = P_{\mathscr{G}}(h)$. Therefore, the network g has a probability mass equal to the probability of the congruence class C_g divided by the number of networks in C_g, $|C_g|$. So, we have the following probability mass function for $\mathscr{G}$:

$$P_{\mathscr{G}}(g) \propto \left(\frac{1}{|C_g|}\right) * P_{\mathscr{C}}(C_g). \tag{1}$$

The setup described above allows an investigator to control the probability of sampling a

network with specific values for network properties. For example, to construct networks with a discrete uniform (non-informative) distribution on mixing matrices, let the ratio $\frac{P_{\mathscr{C}}(C_g)}{P_{\mathscr{C}}(C_h)} = 1$ for all congruence classes; this is the probability distribution assigned to degree mixing matrices as described in section 5. To construct a collection of networks from an informative distribution, one would assign values other than 1 to the ratio; examples of such distributions are provided in section 4.

A Markov chain Monte Carlo (MCMC) procedure–in particular a Metropolis-Hastings algorithm–is used to generate a collection of networks, $\{g_1, \cdots, g_t\}$, that satisfy the probability distribution assigned to the congruence classes. In order to implement the Metropolis-Hastings algorithm, four aspects must be specified: 1) initial starting element, 2) target function, 3) proposal function, and 4) acceptance probability. Appendix B describes a method to construct an initial starting element and equation (1) describes the target function, so we discuss only 3) and 4) below.

2.1 Proposal Function

The algorithm for generating a proposal network, p_{t+1}, at time $t+1$ is based on a procedure called edge switching, which slightly modifies the current network, g_t, while preserving the degree sequence. The procedure selects two edges at random, (a, b) and (c, d), from g_t. If the edges (a, d) and (c, b) do not create multiple edges or self-loops, the network which is created by replacing edges (a, b) and (c, d) with (a, d) and (c, b) is proposed. Otherwise the proposed network is just g_t. To ensure that the edge switching procedure produces a bipartite network, nodes a and c must be of the same type, as must b and d. The algorithm produces an irreducible Markov chain among all graphs with fixed degree sequence. The chain also has equal forward and backward probabilities.

2.2 Acceptance Probability

Given a proposed network, p_{t+1}, the Metropolis-Hastings algorithm will either accept, $g_{t+1} = p_{t+1}$ or reject, $g_{t+1} = g_t$, the proposal. As derived in Goyal et al. (2013), the Metropolis-Hastings acceptance probability for the target and proposal function described above is the following:

$$P(\text{Accept } gp_{t+1}|g_t) = min\left(1, \frac{f(C_{gp_{t+1}}, C_{g_t})}{f(C_{g_t}, C_{gp_{t+1}})} * \frac{P_{\mathscr{C}}(C_{gp_{t+1}})}{P_{\mathscr{C}}(C_{g_t})}\right) \tag{2}$$

where $f(C_g, C_h)$ as the average number of elements in C_h that are valid proposals from an element $g \in C_g$. The value of $f(C_g, C_h)$ can be calculated from the mixing matrices, $MM(g)$ and $MM(h)$, associated with C_g and C_h. Since we only are interested in the ratio of $f(C_g, C_h)$ and $f(C_h, C_g)$, we will assume that $C_g \neq C_h$, otherwise the ratio is one. Given that $C_g \neq C_h$, $f(C_h, C_g) > 0$ only if $MM(g)$ and $MM(h)$ have exactly four distinct entries,$(i, j), (k, l), (k, j)$ and (i, l), such that the following relationships hold:

$$MM_{(r,s)}(h) = \begin{cases} MM_{(r,s)}(g) - 1 \text{ if } r = i \text{ and } s = j \text{ or } r = k \text{ and } s = l \\ MM_{(r,s)}(g) + 1 \text{ if } r = k \text{ and } s = j \text{ or } r = i \text{ and } s = l \\ MM_{(r,s)}(g) \text{ else.} \end{cases} \tag{3}$$

Using this insight, we are able to calculate an approximate acceptance probability, as described in the proposition below, for each proposal graph in our MCMC procedure.

Proposition: *If* $C_g \neq C_h$, $M_s^t \gg max(\{d_i : i \in g\})$, *and relationships outlined in equation (3) hold, then* $f(C_g, C_h) \approx MM_{(j,k)}(g) * MM_{(l,i)}(g) * (1 - P_1 - P_2 + P_1 * P_2)$, *where* $P_1 = \frac{(\gamma^1(i)-1)*(\gamma^2(j)-1)*MM_{(i,j)}(g)}{(\sum_r MM_{(i,r)}(g)-1)*(\sum_r MM_{(r,j)}(g)-1)}$, $P_2 = \frac{(\gamma^1(k)-1)*(\gamma^2(l)-1)*MM_{(k,l)}(g)}{(\sum_r MM_{(k,r)}g-1)*(\sum_r MM_{(r,l)}g-1)}$, *and* $\gamma^t(r)$ *is the average degree of a type* t *node with covariate pattern* r.

Details of proposition are located in Appendix A. Since the degree sequence is fixed, the degree of a node can be treated as a covariate; therefore, the following corollary provides the necessary formula to calculate the acceptance probability using congruence classes defined by degree mixing matrices.

Corollary: *If* $C_g \neq C_h$, $D_s^t \gg max(\{d_i : i \in g\})$, *and relationships outlined in equation (3) hold, then* $f(C_g, C_h) \approx DMM_{(j,k)}(g) * DMM_{(l,i)}(g) * (1 - P_1 - P_2 + P_1 * P_2)$, *where* $P_1 = \frac{(i-1)*(j-1)*DMM_{(j,i)}(g)}{(i*D_i^2-1)*(j*D_j^1-1)}$ *and* $P_2 = \frac{(k-1)*(l-1)*DMM_{(k,l)}(g)}{(k*D_k^2-1)*(l*D_l^1-1)}$.

3 BCPP Cluster Randomized Trial

A series of recent findings regarding the efficacy of antiviral treatment in reducing HIV transmission rates from infected people provided motivation for a large cluster-randomized trial, the Botswana Combination Prevention Project (BCPP). This study is designed to ascertain whether combination preventive modalities can bring about a marked reduction in HIV incidence. Complicating evaluation of properties of the study design, however, is the fact that only limited data are available for estimation of parameters associated with the sexual networks within and among the communities in the study. Because no reliable information to estimate the distribution of number of partnerships per individual is available in Botswana, investigators made use of information from a sexual network study in Likoma Island, Malawi–the most complete sexual network ever collected for Sub-Saharan Africa (Jones et al., 2007). Degree sequences representing number of sexual partners were generated by applying methods described by Handcock and Jones (2004) to data from this study. Even after fixing the degree sequence, however, there are many unknown or imprecise estimates for network properties, including spatial and degree mixing, that must be considered to calculate reliable estimates of statistical power of the study. The methods proposed in section 2 are useful in investigating the importance of two types of mixing, spatial and degree mixing, on statistical power. The bipartite networks we generated were utilized in simulations of both the epidemic spread and the effect on it of interventions to estimate power to detect this effect (Wang et al., 2013).

A complicating factor in design and analysis of cluster-randomized trials is cross-contamination of intervention and control clusters (Hayes et al., 2000). In network models, cross-contamination occurs when edges exist between communities randomized to distinct treatments; in the case of BCPP the treatments are standard of care (SOC) and combination prevention (CP). As shown in Wang et al. (2013), the power of the BCPP trial is significantly impacted by the sexual network–as the level of mixing, i.e. cross-contamination, between the communities assigned to CP and SOC is associated with the degree of attenuation of intervention effects. In section 4, we demonstrate the use of methods described in section 2 to characterize the impact of this form of spatial mixing.

In this paper, we also investigate degree mixing, sometimes referred to as mixing by activity level, i.e. number of sexual partners. Newman (2002) concluded that degree assortative

networks dissseminate disease more easily and are more robost to removal of their highest degree nodes than do disassortative networks. This insight may have important implications for the prevention interventions proposed for BCPP. A wide range of values for mixing by sexual activity levels have been used in epidemic simulations. Many mathematical models calibrated to historical HIV prevalence data select parameter values associated with assortative mixing in the population under study (Eaton et al., 2012), but other modeling studies used parameters associated with disassortitive mixing (Palombi et al., 2012). Morris et al. (2007) expressed the belief that people do not tend to select partners according to their degree, thereby implying the appropriateness of parameter values leading to random mixing by degree; such parameter values were used in a network model of HIV in the United States (Morris et al., 2009). Random degree mixing generates networks that are neither assortative or disassorative. However, many real-world networks do not exhibit random mixing (Newman, 2002). There have been relatively few sexual network studies where estimation of degree assortativity was possible; Bearman et al. (2004) studied student relationships and concluded that they were disassortative. But other studies conducted in STD clinics show evidence for assortivative mixing in populations in the United States and Sweden (Garnett et al., 1996; Granath et al., 1991). The Mochudi pilot program has only collected ego-centric data, and therefore, it is not possible to estimate degree mixing patterns. Network-centric data from Likoma Island, Malawi does provide an estimate for a network in Sub-Saharan Africa, but its relevance for Botswana is unknown. Due to the lack of consensus on degree mixing patterns, we used the proposed methods to generate networks over a wide range of possible degree mixing patterns in order to evaluate the intervention; details are provided in section 5.

4 Spatial Mixing Patterns

To assess whether cross-contamination between intervention and SOC clusters have any effect on the proposed intervention, we first generated bipartite networks from a range of spatial mixing values. After concluding that spatial mixing has a significant impact on estimates of intervention effect and statistical power of the trial, we generated networks using a probability distribution on spatial mixing based on self-reported data from the pilot study in Mochudi,

Botswana.

All the generated bipartite networks contained 1340 nodes that consisted of nodes of type male and type female. The nodes also were given a spatial covariate, either SOC or CP, detailing which community they reside. The nodes were equally labeled as one of four designations, female/SOC, male/SOC, female/CP, or male/CP, and all four designations had identical degree sequences. The congruence classes, as described in section 2, were defined by the mixing matrix. Since the degree sequence is fixed, the proportion of edges between nodes labeled SOC and nodes labeled CP, regardless of the gender label, creates an equivalent partition of the network space $\mathscr{G}$; therefore networks g and h exist in the same congruence class if and only if the proportion of edges between CP and SOC nodes are the same for both g and h. Equation (1) can be equivalently stated as the following:

$$P_{\mathscr{G}}(g) \propto \left(\frac{1}{|C_g|}\right) * P_{\eta}(\eta(g)), \tag{4}$$

where $\eta(g)$ denotes the proportion of mixing between SOC and CP in network g.

In the exploratory stage of investigating the impact of spatial mixing, a total of five collections of networks were generated. Each collection was centered at a distinct level of mixing between two communities. The level of mixing ranged from 0.1 to 0.5 (random mixing between communities); a mixing level of 0.2 denotes that 20% of all relationships are between an individual designated as SOC and an individual designated as CP. Bipartite networks were generated under the following five probability distributions on the congruence classes,

$$\eta \sim N(\mu, \sigma^2), \tag{5}$$

where $\mu \in \{.1, .2, .3, .4, .5\}$ and $\sigma^2 = (.02/1.96)^2$.

The MCMC algorithm outlined in section 2 was used to generate a chain of 5,050,000 networks for each simulation. The first 50,000 were discarded for MCMC burn-in. Of the remaining 5,000,000 networks, every thousandth network was used to calculate the proportion of mixing between the two communities. Figure 1a shows the trace and convergence plots of our MCMC algorithm along with the mixing values for each sampled network. Figure 1b plots the probability density distribution as described in equation (5) as a dashed red line, and the density associated with the proportion of mixing for the simulated networks as a dashed

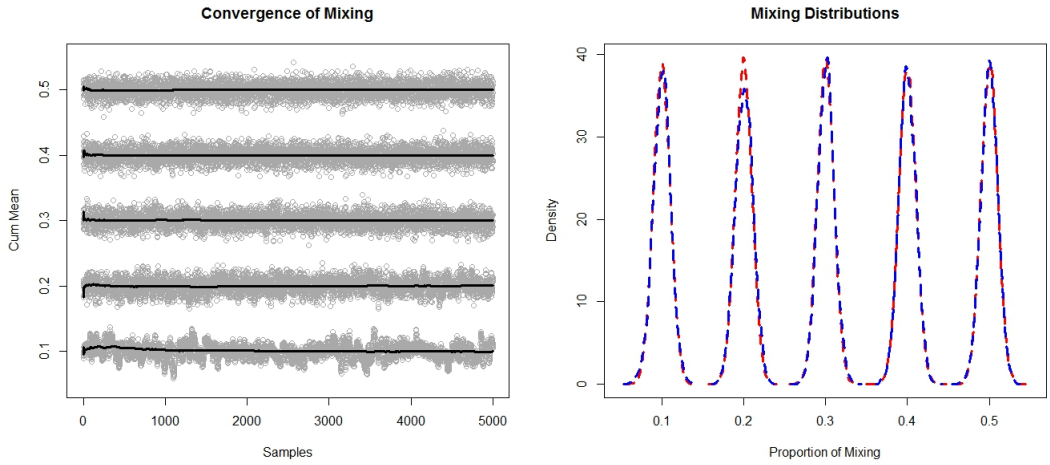

Figure 1: Convergence Plots. (a) Trace and cumulative means for five distinct network mixing patterns. (b) Probability density distribution plots of proportional of mixing described in equation (5) (dashed red line) and the corresponding distribution calculated for the collection of simulated networks (dashed blue line).

blue line for mixing values of {.1, .2, .3, .4, .5}. As all 5 simulated density curves match closely with the 5 corresponding target functions, this provides evidence that the methods described in section 2 generated the desired collection of networks.

Epidemic simulation models were run on each of the sampled networks; these simulation studies showed that difference in 3-year cumulative incidence between SOC and CP communities depended on the amount of mixing; see Wang et al. (2013) for further details regarding the epidemic simulations and results. In order to explore the reason for large variation in 3-year incidence associated with spatial mixing, we investigated how global properties of the networks are affected by changes in mixing. Figure 2 plots values of global network properties for level of mixing from 0.1 to 0.5. The global properties varied little over this range; therefore, we conclude that changes in global properties do not contribute greatly to the effect of mixing on the intervention proposed by BCPP.

As level of mixing across communities has a large effect on statistical power, we generate networks that reflect the uncertainty in this network property among villages that participate in the BCPP trial. Review of data from the pilot study in Mochudi, Botswana implied that about

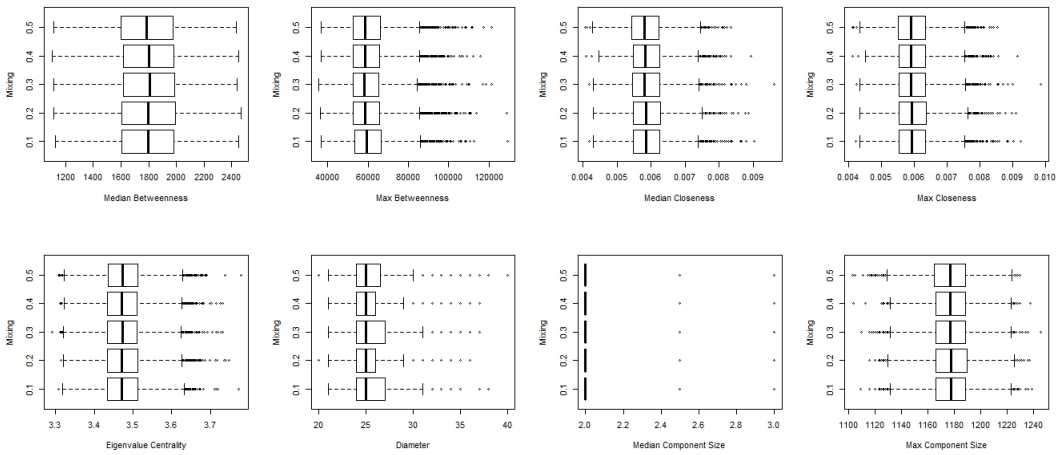

Figure 2: Global Network Property Distributions. Boxplots of global network properties for five distinct network mixing patterns. The first row of plots shows the mean and max of two different centrality measures, betweenness and closeness. The second row shows eigenvalue centrality, diameter, and mean and max component size.

20% of relationships (standard error 2.5%) would be expected to have partners assigned to a different treatment condition from that in their village; for further details regarding the spatial mixing estimates refer to Wang et al. (2013). This implies about 95% of sampled values will be between 15% to 25%; it was considered that a reasonable distribution for the amount of mixing is the following:

$$\eta \sim N(\mu = .2, \sigma^2 = (.05/1.96)^2). \tag{6}$$

The MCMC algorithm outlined in section 2 was used to generate 5 chains of 5,050,000 networks. The first 50,000 of each chain were discarded for MCMC burn-in. Of the remaining 5,000,000 networks, every thousandth network was used to calculate proportion of mixing. Figure 3a shows the trace and convergence plot for all 5 chains combined. The MCMC algorithm used equation (6) and equation (4) to define the target function. Figure 3b depicts a plot of the probability density distribution as described in equation (6) as a dashed red line, and the density associated with the proportion of mixing for the simulated networks as a dashed blue line; the spatial mixing values from the generated networks match the target function closely.

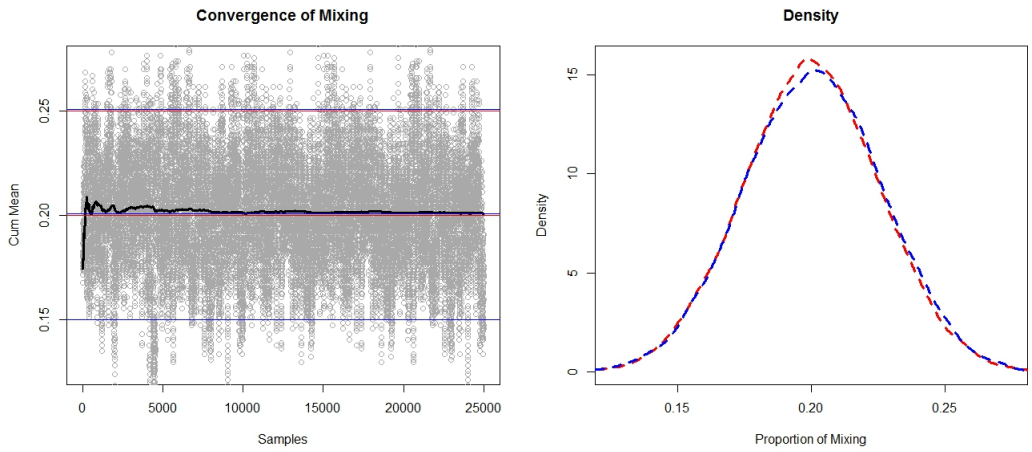

Figure 3: Convergence Plots. (a) Trace, cumulative mean, overall mean, and overall 2.5% and 97.5% quantiles for proportion of mixing for the simulated networks along with the mean and quantiles from the target distribution. (b) Probability density distribution plots of proportional of mixing described in equation (6) (dashed red line) and the corresponding distribution calculated for the collection of simulated networks (dashed blue line).

The increase in variance in equation (6) compared to equation (5) requires a longer MCMC chain to achieve convergence as seen by contrasting the cumulative mean curve in figure 3a to the curves in figure 1a.

5 Degree Mixing Patterns

This section addresses the scenario wherein no estimate exists for a possible influential network property, as is the case for degree mixing in villages that participate in the BCPP trial. To address this lack of information, we construct networks corresponding to a uniform distribution on degree mixing matrices. We note that our proposed approach differs from existing models that generate networks uniformly from the space of all networks conditional on prescribed degree sequence; such an approach produces networks with the undesired property that values for the degree assortativity coefficient will be near zero, see Newman (2002) for details regarding one-mode networks. By contrast, our use of a non-informative degree mixing distribution results in sampling each matrix with equal probability; such a sampling scheme is consistent

with an assumption that individuals can form partnerships based on the partner's degree, but that no information is known about the frequency about this partnership formation process.

To describe the process of constructing collections of networks, we first provide a theorem that states criteria for determining whether a degree mixing matrix *DMM* is graphical, i.e. correspond to a simple undirected bipartite network. This theorem is required because the probability distribution in equation (1) only applies to graphical matrices, and some degree mixing matrices are not graphical.

Theorem: *A matrix, DMM, is graphical by a bipartite undirected network if and only if the following four conditions are met :*

1. $D_i^1 := (\sum_j DMM_{(i,j)})/i \in \mathbb{Z}^+ \forall i$
2. $D_j^2 := (\sum_i DMM_{(i,j)})/j \in \mathbb{Z}^+ \forall j$
3. $DMM_{(i,j)} \leq D_i^1 * D_j^2$
4. $DMM_{(i,j)} \geq 0$.

Refer to Append B for the proof of Theorem 1.

Exactly as in to section 4, all the generated bipartite networks contained 1340 nodes that were equally designated as type male or type female. By setting $P_{\mathscr{C}}(C_g) \propto 1$ for all congruence classes $C_g \in \mathscr{C}$, we are able to generate networks with a uniform distribution on degree mixing matrices. Therefore, the collection of networks will reflect the large uncertainty in degree mixing patterns. The MCMC algorithm outlined in section 2 was used to generate a chain of 5,010,000 networks. The first 10,000 were discarded for MCMC burn-in. Of the remaining 5,000,000 networks, every thousandth network was used to calculate network properties.

To investigate the impact of degree mixing on power for the BCPP trial, the networks generated based on a uniform distribution on degree mixing matrices were used in the epidemic disease simulations, just as they were for the study of spatial mixing. Figures 4 and 5 show predicted cumulative incidence for each of the generated networks under conditions associated with SOC and CP, respectively. The x-axis is the degree assortativity coefficient, a summary

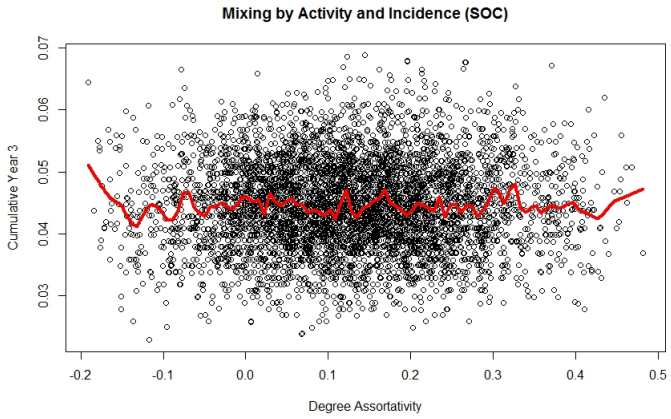

Figure 4: Cumulative Incidence for SOC Community. The x-axis is the degree assortativity coefficient and the red line depicts the lowess curve for the estimated 3-year cumulative incidence.

statistic of a degree mixing matrix. The red line in each of the figures depicts the lowess curve for the estimated 3-year cumulative incidence across the range of degree mixing values for the control and intervention communities. There appears to be no association between cumulative incidence and degree assortativity coefficient for either the SOC or CP communities.

Figure 6 contains scatter plots of degree mixing and global properties of the networks generated under the uniform degree mixing matrix distribution. The first row of plots shows the mean and max of two different centrality measures, betweenness and closeness. The second row shows eigenvalue centrality, diameter, and mean and max component size. The network properties are computed using the igraph library (Csardi and Nepusz, 2006) in R (R Development Core Team, 2011).

In contrast to little variation in global property distributions over different levels of spatial mixing, as shown in figure 2, global properties are highly correlated with degree mixing patterns, shown in figure 6. Nonetheless, the impact of the intervention proposed for BCPP does not appear to be modified by variability in degree mixing patterns.

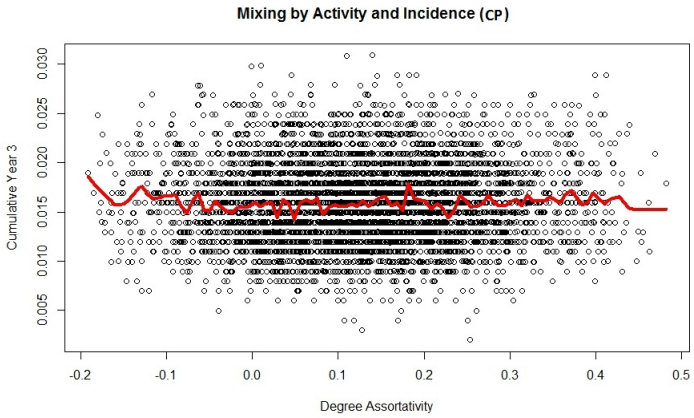

Figure 5: Cumulative Incidence for CP Community. The x-axis is the degree assortativity coefficient and the red line depicts the lowess curve for the estimated 3-year cumulative incidence.

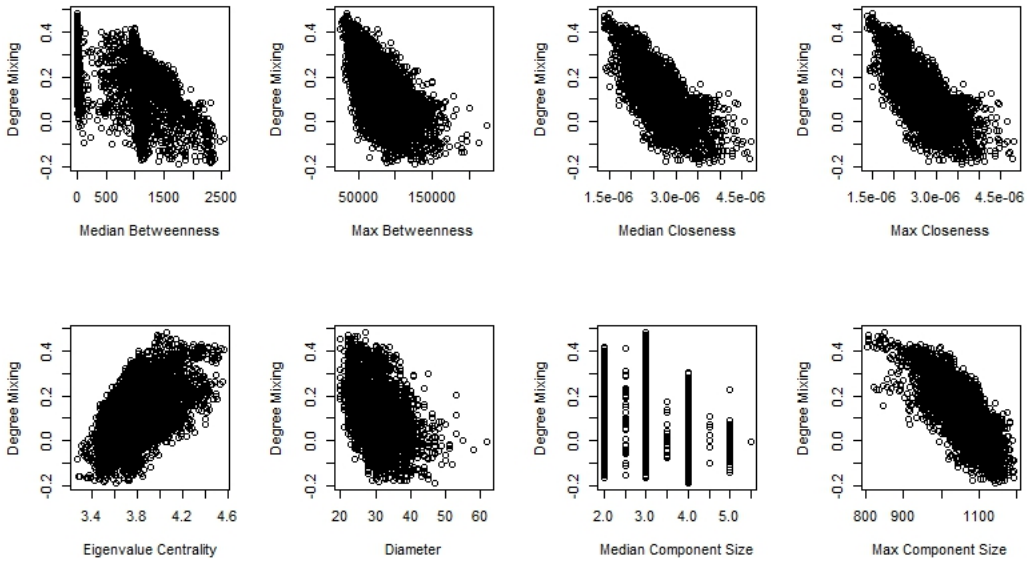

Figure 6: Global Network Property Distributions. Scatter plots of global network properties for range of network degree mixing patterns. The first row of plots shows the mean and max of two different centrality measures, betweenness and closeness. The second row shows eigenvalue centrality, diameter, and mean and max component size.

6 Discussion

We present novel methods to generate bipartite networks that incorporate probability distributions associated with network properties. These methods provide additional flexibility compared to current methods which are based on mean values for networks properties. Networks generated using the proposed methods demonstrated the importance of estimating spatial mixing in the design of the BCPP. In this paper, we make use of the same epidemic model used in the design of the trial, and demonstrate that knowledge regarding degree mixing patterns is not essential in estimating the impact of the trial.

We also investigate the distribution of global network properties of networks constructed over a range of degree and non-degree mixing patterns. Results from the former are associated with large variability in global network properties; nonetheless, this variability does not adversely affect the ability of the model to predict HIV incidence and the efficacy of proposed interventions on incidence. These results demonstrate the complex relationship between network interventions and network topology; therefore, network simulation studies have an important role in predicting the impact of network interventions. The methods presented are useful for generating networks that reflect a given level of uncertainty in order to simulate network processes under investigation. It is important to understand the ways in which network properties impact processes that operate on networks as well as the consequences of lack of information about specified network properties on the ability to predict impact of network interventions.

The supplement describes additional methods to calculate the accept-reject probability as described in equation 2 for network properties associated with density, degree distribution, and mixing patterns–without the requirement that the degree sequence is known. However, further research is necessary to generalize the methods to higher-order network properties.

Supplementary Materials

Additional Methods for Bipartite Network Construction: The supplement includes details to extend the proposed method to density, degree distribution and mixing patterns without requiring a prescribed degree sequence. (pdf file)

References

Bearman, P. S., Moody, J., and Stovel, K. (2004), "Chains of Affection: The Structure of Adolescent Romantic and Sexual Networks," *American Journal of Sociology*, 110(1), 44–91.

Boily, M.-C., Mâsse, B., Alsallaq, R., Padian, N., Eaton, J., Vesga, J., and Hallett, T. (2012), "HIV Treatment as Prevention: Considerations in the Design, Conduct, and Analysis of Cluster Randomized Controlled Trials of Combination HIV Prevention," *PLoS Med*, 9(7), e1001250.

Csardi, G., and Nepusz, T. (2006), "The igraph software package for complex network research," *InterJournal*, Complex Systems, 1695.
URL: *http://igraph.sf.net*

Davis, A., Gardner, B. B., and Gardner, M. R. (1941), *Deep south*, Chicago, IL: University of Chicago Press.

Eaton, J. W., Johnson, L. F., Salomon, J. A., Bärnighausen, T., Bendavid, E., Bershteyn, A., Bloom, D. E., Cambiano, V., Fraser, C., Hontelez, J. A. C., Humair, S., Klein, D. J., Long, E. F., Phillips, A. N., Pretorius, C., Stover, J., Wenger, E. A., Williams, B. G., and Hallett, T. B. (2012), "HIV Treatment as Prevention: Systematic Comparison of Mathematical Models of the Potential Impact of Antiretroviral Therapy on HIV Incidence in South Africa," *PLoS Med*, 9(7), e1001245.

Galaskiewicz, J. (1985), *Social organization of an urban grants economy: A study of business philanthropy and nonprofit organizations*, Vol. 80, Orlando, FL: Academic Press.

Garnett, G., Hughes, J., Anderson, R., Stoner, B., Aral, S., Whittington, W., Handsfield, H., and Holmes, K. (1996), "Sexual Mixing Patterns of Patients Attending Sexually Transmitted Diseases Clinics," *Sexually Transmitted Diseases*, 23(3), 248 – 257.

Goh, K.-I., Cusick, M. E., Valle, D., Childs, B., Vidal, M., and Barabasi, A.-L. (2007), "The human disease network," *Proceedings of the National Academy of Sciences*, 104(21), 8685–8690.

Goyal, R., Blitzstein, J., and Gruttola, V. D. (2013), "Sampling Networks from Their Posterior Predictive Distribution," *Under Review*, .

Granath, F., Giesecke, J., Scalia-Tomba, G., Ramstedt, K., and Forssman, L. (1991), "Estimation of a preference matrix for women's choice of male sexual partner according to rate of partner change, using partner notification data," *Mathematical Biosciences*, 107(2), 341 – 348.
URL: *http://www.sciencedirect.com/science/article/pii/0025556491900139*

Handcock, M. S., and Jones, J. H. (2004), "Likelihood-based inference for stochastic models of sexual network formation," *Theoretical Population Biology*, 65(4), 413ï£¡422.

Hayes, R., Alexander, N. D., Bennett, S., and Cousens, S. (2000), "Design and analysis issues in cluster-randomized trials of interventions against infectious diseases," *Statistical Methods in Medical Research*, 9(2), 95–116.

Jones, J., Helleringer, S., and Kohler, H. (2007), "Extended Abstract: Exponential Random Graph Models for Sexual Networks on Likoma Island, Malawi: Implications for Sexual Behavior and HIV Control," , .

Kolaczyk, E. (2009), "Sampling and Estimation in Network Graphs," in *Statistical Analysis of Network Data: Methods and Models*, New York: Springer Science+Business Media, LLC, chapter 5, pp. 123–152.

Kossinets, G. (2006), "Effects of missing data in social networks," *Social networks*, 28(3), 247–268.

Landon, B. E., Keating, N. L., Barnett, M. L., Onnela, J.-P., Paul, S., O'Malley, A. J., Keegan, T., and Christakis, N. A. (2012), "Variation in patient-sharing networks of physicians across the United States," *JAMA: the journal of the American Medical Association*, 308(3), 265.

Morris, M., Goodreau, S., and Moody, J. (2007), Sexual networks, concurrency, and STD/HIV,, in *Sexually Transmitted Diseases*, eds. K. Holmes, S. PF, and S. WE, McGraw-Hill International Book Co, New York, NY, USA, pp. 109–126.

Morris, M., Kurth, A., Hamilton, D., Moody, J., and Wakefield, S. (2009), "Concurrent Partnerships and HIV Prevalence Disparities by Race: Linking Science and Public Health Practice," *American Journal of Public Health*, 99(6), 1023–1031.

Newman, M. (2002), "Assortative mixing in networks," *Physical Review Letters*, 89(20), 208701.

Newman, M. E. (2010), *Networks An Introduction*, New York: Oxford University Press.

Onnela, J.-P., Saramäki, J., Hyvönen, J., Szabó, G., Lazer, D., Kaski, K., Kertész, J., and Barabási, A.-L. (2007), "Structure and tie strengths in mobile communication networks," *Proceedings of the National Academy of Sciences*, 104(18), 7332–7336.

Palombi, L., Bernava, G. M., Nucita, A., Giglio, P., Liotta, G., Nielsen-Saines, K., Orlando, S., Mancinelli, S., Buonomo, E., Scarcella, P., Altan, A. M. D., Guidotti, G., Ceffa, S., Haswell, J., Zimba, I., Magid, N. A., and Marazzi, M. C. (2012), "Predicting trends in HIV-1 sexual transmission in Sub-Saharan Africa through the Drug Resource Enhancement Against AIDS and Malnutrition model: Antiretrovirals for reduction of population infectivity, incidence and prevalence at the district level," *Clinical Infectious Diseases*, .

R Development Core Team (2011), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.
URL: *http://www.R-project.org*

Ramasco, J. J., Dorogovtsev, S. N., and Pastor-Satorras, R. (2004), "Self-organization of collaboration networks," *Physical review E*, 70(3), 036106.

Shalizi, C. R., and Rinaldo, A. (2013), "Consistency under sampling of exponential random graph models," *The Annals of Statistics*, 41(2), 508–535.

Wang, P., Pattison, P., and Robins, G. (2012), "Exponential random graph model specifications for bipartite networks-A dependence hierarchy," *Social Networks*, .

Wang, R., Goyal, R., Lei, Q., Essex, M., and De Gruttola, V. (2013), "Sample Size Considerations in the Design of Cluster Randomized Trials of Combination HIV Prevention," *Harvard University Biostatistics Working Paper Series*, 161, 1–29.
URL: *http://biostats.bepress.com/harvardbiostat/paper161*

7 Appendix A: Proof of Proposition

Proposition: *If* $C_g \neq C_h$, $M_s^t \gg max(\{d_i : i \in g\})$, *and relationships outlined in equation (3) hold, then* $f(C_g, C_h) \approx MM_{(j,k)}(g) * MM_{(l,i)}(g) * (1 - P_1 - P_2 + P_1 * P_2)$, *where* $P_1 = \frac{(\gamma^1(i)-1)*(\gamma^2(j)-1)*MM_{(i,j)}(g)}{(\sum_r MM_{(i,r)}(g)-1)*(\sum_r MM_{(r,j)}(g)-1)}$, $P_2 = \frac{(\gamma^1(k)-1)*(\gamma^2(l)-1)*MM_{(k,l)}(g)}{(\sum_r MM_{(k,r)}g-1)*(\sum_r MM_{(r,l)}g-1)}$, *and* $\gamma^t(r)$ *is the average degree of a type t node with covariate pattern r.*

Proof of Proposition: Let $T_1(g)$ be the edges in g where the endpoint of type 1 has covariate pattern i and the other endpoint of type 2 has covariate pattern j; so $|T_1(g)| = MM_{(i,j)}(g)$. Let $T_2(g)$ be the edges in g where the endpoint of type 1 has covariate pattern k and the other endpoint of type 2 has covariate pattern l, so $|T_2(g)| = MM_{(k,l)}(g)$. Let $h(e_1, e_2; g)$ denote the resulting graph after swapping the endpoints of edges e_1 and e_2; it is possible that $h(e_1, e_2; g)$ contains a multi-edge, and hence not a valid proposal graph, i.e $h(e_1, e_2; g) \notin \mathscr{G}$. The average number of valid proposals from a graph $g \in C_g$ to a graph in C_h can be expressed as the following,

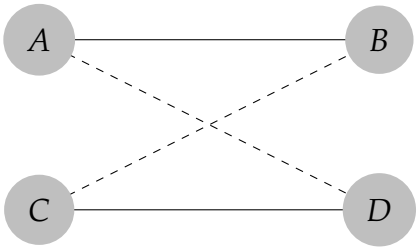

Figure 7: Edge switching by replacing edges (A,B) and (C,D), solid lines, with edges (A,D) and (C,B), dashed lines.

$$
\begin{aligned}
f(C_g, C_h) &= \frac{1}{|C_g|} \sum_{g \in C_g} \sum_{\substack{e_1 \in T_1(g) \\ e_2 \in T_2(g)}} I_{h(e_1,e_2;g) \in \mathscr{G}} \\
&= MM_{(i,j)}(g) * MM_{(k,l)}(g) - \frac{1}{|C_g|} \sum_{g \in C_g} \sum_{\substack{e_1 \in T_1(g) \\ e_2 \in T_2(g)}} I_{h(e_1,e_2;g) \notin \mathscr{G}} \\
&= MM_{(i,j)}(g) * MM_{(k,l)}(g) - \frac{1}{|C_g|} \sum_{g \in C_g} MM_{(i,j)}(g) * MM_{(k,l)}(g) * P(h(e_1, e_2; g) \notin \mathscr{G}) \\
&= MM_{(i,j)}(g) * MM_{(k,l)}(g)(1 - P), \qquad (7)
\end{aligned}
$$

where $P(h(e_1, e_2; g) \notin \mathscr{G})$ is the probability of creating a multi-edge by switching edges $e_1 \in T_1(g)$ and $e_2 \in T_2(g)$, P is the average probability of $P(h(e_1, e_2; g) \notin \mathscr{G})$ over all $g \in C_g$, and I_α is the indicator function for an event α. P can be calculated by studying an arbitrary edge switch in $g \in C_g$ that could lead to a valid $h \in C_h$. Let A, B, C, and D be nodes in g with the following covariate patterns: $m_A^1 = i, m_B^2 = j, m_C^1 = k$, and $m_D^2 = l$; A and C are of type 1 and B and D are of type 2. Let h be the resulting graph after replacing edges (A, B) and (C, D) with (A, D) and (C, B) where $(A, B) \in g$ and $(C, D) \in g$. A multi-edge occurs in h when $(A, D) \in g$ or $(C, B) \in g$; an illustration is shown in figure 7. The probability of a multi-edge can be express as the following:

$$P = P((A,D) \in g \text{ or } (C,B) \in g | (A,B),(C,D) \in g) \quad (8)$$

$$\approx P_1 + P_2 - P_1 * P_2, \quad (9)$$

where $P_1 = P((A,D) \in g | (A,B),(C,D) \in g)$ and $P_2 = P((C,B) \in g | (A,B),(C,D) \in g)$.

To calculate an expression for P_1 and P_2, we need to introduce some notation. Let S_1 be the set of edges between node A and nodes with covariate pattern l. Let $p(S_1, D | (C,D) \in g)$ be the probability that an edge $e \in S_1$ connects to D given $(C,D) \in g$. An approximate expression for P_1 is the following:

$$P_1 \approx p(S_1, D | (C,D) \in g) * |S_1|. \quad (10)$$

Let $\gamma^1(i)$ denote the average degree of a type 1 node with covariate pattern i; if covariate patterns are defined by nodal degree then $\gamma^1(i) = i$. The size of S_1 approximately followings the distribution below:

$$|S_1| \sim Bin\left(\gamma^1(i) - 1, \frac{MM_{(i,l)}}{\sum_r MM_{(i,r)} - 1}\right). \quad (11)$$

Since one edge of node A is known to attach to node B and $m_B \neq m_D$, there are only $\gamma(i) - 1$ remaining edges that can attach to D. Let S_2 be the set of edges between node D and nodes with covariate pattern i. Similar to S_1, the size of S_2 approximately followings the distribution below:

$$|S_2| \sim Bin\left(\gamma^2(l) - 1, \frac{MM_{(i,l)}}{\sum_r MM_{(r,l)} - 1}\right). \quad (12)$$

Figure 8 depicts many of the quantities described above.

$p(S_1, D | (C,D) \in g)$ is equal to the probability that one of the edges in S_1 connects to one of the edges in S_2. Each edge in S_1 can only connect to one of a possible $MM_{(i,l)}$ edges associated with

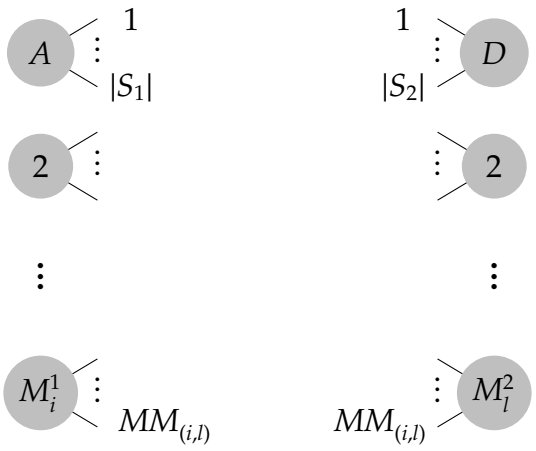

Figure 8: A depiction of the edges associated with type 1 nodes with covariate pattern i and edges associated with type 2 nodes with covariate pattern l. A multi-edge, as described in figure 7 can occur if one of the edges depicted connects node A and node D.

nodes of covariate pattern l. Since $|S_1|$ and $|S_2|$ are small compared to M_i^1 and M_l^2, each of the S_1 edges have approximately $|S_2|/MM_{(i,l)}$ probability of connecting with an edge in S_2. Therefore,

$$P_1 \approx E[|S_1| * |S_2|/MM_{(i,l)}] = E[|S_1|] * E[|S_2|] * 1/MM_{(i,l)} \tag{13}$$

$$= \frac{(\gamma^1(i)-1) * (\gamma^2(j)-1) * MM_{(i,j)}(g)}{(\sum_r MM_{(i,r)}(g)-1) * (\sum_r MM_{(r,j)}(g)-1)}. \tag{14}$$

Similarly,

$$P_2 = P((C,B) \in g | (A,B),(C,D) \in g) \approx \frac{(\gamma^1(k)-1) * (\gamma^2(l)-1) * MM_{(k,l)}(g)}{(\sum_r MM_{(k,r)}g-1) * (\sum_r MM_{(r,l)}g-1)}. \tag{15}$$

Substituting equations (14) and (15) in equation (9) and using that quantity in equation (7), we get $f(C_g, C_h) \approx MM_{(j,k)}(g) * MM_{(l,i)}(g) * (1 - P_1 - P_2 + P_1 * P_2)$.

8 Appendix B: Proof of Theorem

Theorem: *A matrix, DMM, is graphical by a bipartite undirected network if and only if the following four conditions are met :*

1. $D_i^1 := (\sum_j DMM_{(i,j)})/i \in \mathbb{Z}^+ \forall i$
2. $D_j^2 := (\sum_i DMM_{(i,j)})/j \in \mathbb{Z}^+ \forall j$
3. $DMM_{(i,j)} \leq D_i^1 * D_j^2$
4. $DMM_{(i,j)} \geq 0$.

Proof of Theorem: Given an undirected bipartite graph it is clear that the degree mixing matrix will satisfy the conditions in the Theorem. Thus, we need only show that a matrix which satisfies the four criteria is graphical via a bipartite graph. This will be shown by constructing a realization of the matrix. We begin by generating an empty network with $\sum_i D_i^1$ and $\sum_i D_i^2$ nodes of type 1 and type 2, respectively, where D_i^1 and D_i^2 of type 1 and type 2 will have degree i. Conditions (1) and (2) guarantee that D_i^1 and D_i^2 are non-negative integers.

The next step is to add edges between type 1 nodes of degree i to type 2 nodes of degree j, for each $i \in \{1, \cdots, r\}$ and $j \in \{1, \cdots, s\}$. We will use a similar approach to connect type 1 nodes with degree i to type 2 nodes of degree j as presented in the proof for Theorem 1 in Goyal et al. (2013). The approach starts by defining the components of $\vec{\alpha_i^{(1)}}$, $\alpha_{i_k}^{(1)}$, as the available edges left to be connected for the k^{th} type 1 node with degree i. The components of $\vec{\beta_i^{(1)}}$ are defined so that $\beta_{i_k}^{(1)} \in \{\lfloor \frac{DMM_{(i,j)}}{D_i^1} \rfloor, \lceil \frac{DDM_{(i,j)}}{D_i^1} \rceil\}$, $\sum_k \beta_{i_k}^{(1)} = DMM_{(i,j)}$, and $\beta_{i_1}^{(1)} \geq \beta_{i_2}^{(1)} \geq \cdots \geq \beta_{i_{D_i^1}}^{(1)}$. Let $\vec{\alpha_j^{(2)}}$ and $\vec{\beta_j^{(2)}}$ be defined similarly for type 2 nodes with degree j. Without loss of generality assume that $\vec{\alpha_i^{(1)}}$ and $\vec{\alpha_j^{(2)}}$ are in decreasing order. The construction procedure will connect the first type 1 node of degree i to the first $\beta_{i_1}^{(1)}$ type 2 nodes in $\vec{\alpha_j^{(2)}}$. Next, we connect the second type 1 node of degree i to the next $\beta_{i_2}^{(1)}$ nodes in $\vec{\alpha_j^{(2)}}$, and repeat this process for all D_i^1 degree i nodes. There are only three issues that could arise.

Issue 1: $\beta_{i_k}^{(1)} > D_j^2$.

Issue 1 occurs when a single node of type 1, k, of degree i must connect to $\beta_{i_k}^{(1)}$ type 2 nodes of degree j, but $\beta_{i_k}^{(1)}$ is greater than the number of type 2 nodes of degree j, D_j^2. Thus, node k must form two edges with the same node of degree j. This cannot occur because $\beta_{i_k}^{(1)} \leq \lceil \frac{DDM_{(i,j)}}{D_i^1} \rceil \leq D_j^2$ by our condition (3).

Issue 2: $\alpha_{i_k}^{(1)} < \beta_{i_k}^{(1)}$.

Initially there is a total of $\sum_j DMM_{(i,j)}$ available edges of type 1 nodes of degree i to connect to type 2 nodes. At each step of connecting type 1 nodes of degree i to type 2 nodes of l, $DMM_{(i,l)}$ available edges are removed. Thus, at the step of connecting type 1 nodes of degree i to connect to type 2 nodes of degree j, there exists at least $DMM_{(i,j)}$ available edges. So, $\sum_k \alpha_{i_k}^{(1)} \geq DMM_{(i,j)} = \sum_k \beta_{i_k}^{(1)}$. Let p_1 and p_2 denote partitions of $\sum_k \alpha_{i_k}^{(1)}$ and $\sum_k \beta_{i_k}^{(1)}$ into D_i^1 values such that the values in the partitions are decreasing and as balanced as possible. These partitions have the property that $p_1 \geq p_2$ for each pairwise element. These particular partitions are exactly what the algorithm is generating with $\alpha_{i_k}^{(1)}$ and $\beta_{i_k}^{(1)}$. So, it can be concluded that $\alpha_{i_k}^{(1)} \geq \beta_{i_k}^{(1)}$.

Issue 3: $\alpha_{j_k}^{(2)} < \beta_{j_k}^{(2)}$.

Due to the symmetry of type 1 and type 2 nodes, the proof that $\alpha_{j_k}^{(2)} < \beta_{j_k}^{(2)}$ is not possible is identical to the proof for issue 2. ■