2015

Inference for social network models from egocentrically-sampled data, with application to understanding persistent racial disparities in HIV prevalence in the US

Pavel N. Krivitsky

Martina Morris

Inference for social network models from egocentrically-sampled data, with application to understanding persistent racial disparities in HIV prevalence in the US

Abstract

Egocentric network sampling observes the network of interest from the point of view of a set of sampled actors, who provide information about themselves and anonymized information on their network neighbors. In survey research, this is often the most practical, and sometimes the only, way to observe certain classes of networks, with the sexual networks that underlie HIV transmission being the archetypal case. Although methods exist for recovering some descriptive network features, there is no rigorous and practical statistical foundation for estimation and inference for network models from such data. We identify a sub-class of exponential-family random graph models (ERGMs) amenable to being estimated from egocentrically sampled network data, and apply pseudo-maximum-likelihood estimation to do so and to rigorously quantify the uncertainty of the estimates. For ERGMs parametrized to be invariant to network size, we describe a computationally tractable approach to this problem. We use this methodology to help understand persistent racial disparities in HIV prevalence in the US.

NIASRA

NATIONAL INSTITUTE FOR APPLIED STATISTICS RESEARCH AUSTRALIA

National Institute for Applied Statistics Research Australia

University of Wollongong

Working Paper

05-15

Inference for Social Network Models from Egocentrically-Sampled Data, with Application to Understanding Persistent Racial Disparities in HIV Prevalence in the US.

Pavel N. Krivitsky and Martina Morris

National Institute for Applied Statistics Research Australia, University of Wollongong, Wollongong NSW 2522. Phone +61 2 4221 5435, Fax +61 2 4221 4845. Email: anica@uow.edu.au

Inference for Social Network Models from Egocentrically-Sampled Data, with Application to Understanding Persistent Racial Disparities in HIV Prevalence in the US

Pavel N. Krivitsky[*†] **and Martina Morris**[*]

Abstract: Egocentric network sampling observes the network of interest from the point of view of a set of sampled actors, who provide information about themselves and anonymized information on their network neighbors. In survey research, this is often the most practical, and sometimes the only, way to observe certain classes of networks, with the sexual networks that underlie HIV transmission being the archetypal case. Although methods exist for recovering some descriptive network features, there is no rigorous and practical statistical foundation for estimation and inference for network models from such data. We identify a subclass of exponential-family random graph models (ERGMs) amenable to being estimated from egocentrically sampled network data, and apply pseudo-maximum-likelihood estimation to do so and to rigorously quantify the uncertainty of the estimates. For ERGMs parametrized to be invariant to network size, we describe a computationally tractable approach to this problem. We use this methodology to help understand persistent racial disparities in HIV prevalence in the US.

Keywords and phrases: social network, ERGM, random graph, egocentrically-sampled data, pseudo maximum likelihood, pseudo likelihood.

1. Introduction

There is growing interest in the statistical modeling of network data across a wide range of fields: from the study of political coalitions in the social

[*]The authors wish to thank Professors Mark S. Handcock, Raymond Chambers, David Steel, and Robert Clark, and members of the University of Washington Network Modeling Group, particularly Professor Steven M. Goodreau, for helpful discussions and comments on this manuscript; and the Statnet Team for their software. Computations were performed on a cluster partially funded by an NICHD research infrastructure grant R24HD042828, to the Center for Studies in Demography and Ecology at the University of Washington; and both authors were supported, in part, by NIH award R01HD068395.

[†]Supported, in part, by ONR award N000140811015.

sciences, to protein-protein interaction networks in genetics and the spread of infectious diseases in epidemiology. In some cases, it is possible to observe the complete network of interest, but in others the network must be sampled. Estimating network models from sampled data raises some unique issues. While progress has been made in developing the general framework for statistical inference (Handcock and Gile, 2010), there is a need for feasible methods that can be used with common network sampling designs in different fields. In this work we present a framework for inference from egocentrically sampled network data, which often contain very limited information about network structure: for those individuals in the sample, only information about their immediate partners in the network is observed, and even that information is often limited to non-identifying demographics. The work was motivated by a specific question in the field of HIV epidemiology—Does network structure help explain the persistent racial disparities in HIV prevalence in the United States?—but it has the potential for wide application given the simplicity of collecting egocentrically sampled network data in the population sciences.

The HIV epidemic in the US is now in its third decade. While the rate of transmission has dropped, the racial disparities in HIV prevalence have become entrenched. An African American today is 10 times more likely than a white American to be living with HIV/AIDS. The disparity begins early in life (Morris et al., 2006), and persists through to old age (NCHHSTP, 2013), and is evident among all risk groups: heterosexuals, men who have sex with men (MSM), and injection drug users.

The disproportionate risks faced by heterosexual African-American women are especially steep. In 2010, the most recent year for which statistics are available (NCHHSTP, 2012), there were an estimated 5,300 heterosexually acquired new infections among African-American women. By comparison, there were 2,700 heterosexually acquired infections among African-American men, 1,300 among white women, and 620 among white men. While per-capita infection rates cannot be constructed for heterosexuals (because the denominators are not known), the annual rates of heterosexually acquired infections for the demographic subgroups are roughly 33, 19, 7 and 1 per 100,000 persons for African-American women and men and White women and men, respectively. Similar disparities are found among other sexually transmitted infections, both bacterial and viral (Morris et al., 2006). The magnitude varies by pathogen and changes over time, but the disparities have been remarkably persistent. For the older reportable STIs, like gonorrhea and syphilis, they stretch back to the earliest reports in the 1960s, and reach per-capita rate ratios of 50–100. (NCDC, 1967)

The determinants of these disparities remain elusive. Empirical studies repeatedly find that they cannot be explained by systematic differences in individual behavior, such as higher numbers of partners or rates of injection drug use, or lower condom use (Hallfors et al., 2007, for example). Nor have race-linked biological differences been identified that could explain disparities across this wide range of pathogens. What all of these infections do share is an underlying transmission network. The structure of a transmission network can channel the spread of infection in the same way that a transportation network can channel the flow of traffic, producing emergent patterns that reflect the connectivity of the system, rather than the behavior of any particular element.

A growing body of work is therefore focused on the role that network structure may play in explaining these disparities. Descriptive analyses and simulation studies (Laumann et al., 1992, 1994; Morris, 1993; Morris and Kretzschmar, 1997) have focused attention on two structural features: homophily and concurrency. Homophily is the strong propensity for within-group partner selection. It is a common pattern for many social attributes, though not all. (For example, most sexual partnerships are cross-sex rather than same-sex.) When present, homophily leads to clustered, segregated networks. Concurrency is non-monogamy—having partners that overlap in time. While there is a very strong norm of monogamy in sexual partnerships, deviations from the norm occur. When present, concurrency increases network connectivity by allowing for the emergence of stable network connected components larger than dyads (pairs of individuals).

The hypothesis is that these two network properties together can produce the sustained HIV/STI prevalence differentials we observe: differences in concurrency between groups are the mechanism that generates the prevalence disparity, while homophily is the mechanism that sustains it. To test this hypothesis, we need to assess the strength and significance of observed concurrency differentials and homophily by race, and to evaluate whether the observed network mechanisms predict differentials in network exposure by race and sex that are consistent with the differentials in observed HIV prevalence.

Generative models for social networks, like exponential-family random graph models (ERGMs), let us test for these effects, and we can simulate from them to predict network exposure; but these models must first be fit to available data that can support broad population-level inference. Our main statistical challenge is, therefore, to fit generative network models (ERGMs in particular) to egocentrically sampled data, and to obtain rigorous measures of uncertainty of these fits; and to do so in a computationally feasible

manner, for when the population of interest is very large or its size is unknown. We elaborate on these models and these data in turn.

1.1. Exponential-family random graph models

Exponential-family random graph models (ERGMs) are a popular and, importantly for us, parsimonious, class of generative models for graphs in general and for social networks in particular. (Frank and Strauss, 1986; Wasserman and Pattison, 1996; Hunter and Handcock, 2006) An ERGM expresses the probability of an observed graph $\boldsymbol{y}$ as an exponential family:

$$\Pr_{\boldsymbol{g}}(\boldsymbol{Y} = \boldsymbol{y}; \boldsymbol{x}, \boldsymbol{\theta}) \equiv \exp\{\boldsymbol{\theta}^\top \boldsymbol{g}(\boldsymbol{y}, \boldsymbol{x})\}/\kappa_{\boldsymbol{g}}(\boldsymbol{\theta}, \boldsymbol{x}),\ \boldsymbol{y} \in \mathcal{Y}. \tag{1.1}$$

It is specified by the sample space $\mathcal{Y}$ of possible networks (configurations of relationships) and a sufficient statistic vector $\boldsymbol{g}(\boldsymbol{y}, \boldsymbol{x})$, which is a function of the whole network $\boldsymbol{y}$ and possible covariates $\boldsymbol{x}$, and whose elements are selected to represent features of the network that are of substantive interest or believed relevant to the generative process of the network (e.g., count of monogamous actors to represent monogamy and count of ties within an exogenously defined group to represent homophily); and it is parametrized by its vector of natural parameters $\boldsymbol{\theta}$. The normalizing constant $\kappa_{\boldsymbol{g}}(\boldsymbol{\theta}, \boldsymbol{x}) \equiv \sum_{\boldsymbol{y}' \in \mathcal{Y}} \exp\{\boldsymbol{\theta}^\top \boldsymbol{g}(\boldsymbol{y}', \boldsymbol{x})\}$ is usually intractable when the choice of $\boldsymbol{g}(\boldsymbol{y}, \boldsymbol{x})$ induces dependence among the relationship states.

Analogously to $\Pr_{\boldsymbol{g}}(\cdot; \boldsymbol{x}, \boldsymbol{\theta})$, we define $\mathrm{E}_{\boldsymbol{g}}(\cdot; \boldsymbol{x}, \boldsymbol{\theta})$ and $\mathrm{var}_{\boldsymbol{g}}(\cdot; \boldsymbol{x}, \boldsymbol{\theta})$, as, respectively, the expectation and the variance under this ERGM process; and let $\boldsymbol{\mu}_{\boldsymbol{g}}(\boldsymbol{\theta}, \boldsymbol{x}) \equiv \mathrm{E}_{\boldsymbol{g}}\{\boldsymbol{g}(\boldsymbol{Y}, \boldsymbol{x}); \boldsymbol{x}, \boldsymbol{\theta}\}$, the smooth and invertible (Brown, 1986, Thm. 3.6, for example) mapping from the *natural* to the *mean-value* parameters of this model. Call its inverse $\boldsymbol{\theta}_{\boldsymbol{g}}(\boldsymbol{\mu}, \boldsymbol{x}) \equiv (\boldsymbol{\theta}$ s.t. $\boldsymbol{\mu}_{\boldsymbol{g}}(\boldsymbol{\theta}; \boldsymbol{x}) = \boldsymbol{\mu})$.

Estimating $\boldsymbol{\theta}$ facilitates inference about the social forces that shape the network as well as principled simulation of complete networks whose features are similar, on average, to those of the network observed. In the case of sampled network data in particular, it would allow recovering possible full networks from which the sample may have been drawn. Therefore, $\boldsymbol{\theta}$ is our target of inference.

1.2. Egocentrically sampled data

Network data are distinguished by having two units of analysis: the actors and the links between the actors. This gives rise to a range of sampling designs that can be classified into two groups: link tracing designs (e.g.,

snowball and respondent driven sampling) and egocentric designs. Much of the recent literature has focused on developing model- or design-based inference for link tracing designs. (Thompson and Frank, 2000; Salganik and Heckathorn, 2004; Volz and Heckathorn, 2008; Snijders, 2010; Handcock and Gile, 2010; Tomas and Gile, 2011; Illenberger and Fltter, 2012; Pattison et al., 2013) This work focuses on the egocentric designs that are more commonly used in the social sciences, but are less well developed statistically.

Egocentric network sampling comprises a range of designs developed specifically for the collection of network data in social science survey research. The design is (ideally) based on a probability sample of respondents ("egos") who, via interview, are asked to nominate a list of persons ("alters") with whom they have a specific type of relationship ("tie"), and then asked to provide information on the characteristics of the alters and/or the ties. The alters are typically not directly observed. Depending on the study design, alters may or may not be uniquely identifiable, and respondents may or may not be asked to provide information on one or more ties among alters (the "alter" matrices). Alters could, in theory, also be present in the data as an ego or as an alter of a different ego; the likelihood of this depends on the sampling fraction.

In this work, we focus on the minimal egocentric network study design, in which alters cannot be uniquely identified and alter matrices are not collected. (See Smith (2012) and Gjoka, Smith, and Butts (2014a) for considerations of when they are.) The minimal design is more common, and the data are more widely available, for three reasons.

The first is confidentiality, a key consideration with respect to alter identification. If the relationship of interest is sensitive, requiring full identification of the alters is likely to reduce respondent disclosure, and knowledge of alter–alter ties by the respondent may be unreliable. In addition, Institutional Review Boards often forbid the collection of identifiable data about the alters, as the alters have not given informed consent for their personal information to be collected. The minimal egocentric design allows for representative data to be collected in such contexts, with less intrusion and full consent. In public health research on HIV and other STIs, for example, egocentric study designs make it possible to conduct empirical research on how individual sexual behavior influences the population structure of infection transmission networks. There is a growing international archive of public data from such studies, with comparable surveys now available from over 50 different countries as far back as the early 1990s (MEASURE DHS, 2000–2014; Tanfer, 1991; Laumann et al., 1992; Udry, 2003; NSFG, 2002, 2006–2011, for example).

The second is time, a key consideration with respect to alter matrix collection. The number of potential ties grows with the number of actors in the network nominated, quickly making data collection burdensome for the respondent, and difficult to justify in large scale surveys that must serve multiple needs. As a result, even the less sensitive forms of social network data tend to be collected using the minimal egocentric design. Perhaps the best known example is the friendship network data collected annually by the General Social Survey since 1985 (Burt, 1984), which was used in the landmark study of the decline in American friendship and social support networks *Bowling Alone* (Putnam, 2000).

The third is compatibility with established methods for survey sampling with population-based inference. While in theory the "seed nodes" for a link-traced sample could be chosen at random from the population of interest, the real strength of these designs is the ability to sample from hidden or inaccessible populations, where no sampling frame is available, and this is the application context in which they are most often used. Egocentric designs, by contrast, sample egos using standard sampling methods, and the sampling of links is implemented through the survey instrument. As a result, these methods are easily integrated into population-based surveys, and, as we show below, inherit many of the inferential benefits.

Despite the widespread availability of egocentrically sampled network data, statistical methods for analyzing them are still relatively undeveloped. Early work focused on limited descriptive methods for analyzing "mixing matrices", cross-tabulations of ego-alter dyads by actor attributes (Marsden, 1981; Morris, 1991), or bivariate associations between ego attributes and alter summary statistics (Marsden, 1987; Admiraal, 2009). More recent work has focused on the key topic of recovering whole network attributes from egocentric data (Gjoka, Smith, and Butts, 2014b, for example).

Handcock and Gile (2010) established a general framework for model-based inference for networks based on sampled data that allows for egocentrically sampled data as a special case: when only dyads incident on those in the sample are observed, and Koskinen, Robins, and Pattison (2010) developed a similar approach in a Bayesian framework. Unfortunately, the likelihood approach is infeasible for our problem for three reasons. Firstly, the approach requires fitting an ERGM to a network of the size equal to that of the population from which the egos were sampled, which is, often, on the order of millions, and possibly unknown. Secondly, Handcock and Gile's development was for a case where each of the alters nominated could be uniquely identified: that one could identify when one ego nominates another ego and when two egos nominate the same alter. For most existing

egocentrically sampled data (including all of the studies cited above), alters nominated by distinct egos cannot be matched. Although a likelihood can be derived for this case as well, it requires integration over the space of networks that produce *exactly* the observed dataset—a more complex constraint. Thirdly, if the data come from a complex (even just weighted) design, ignorability of the sampling process might not hold, requiring nested integration over the sampling process as well.

Krivitsky, Handcock, and Morris (2011) described how the sufficient statistic needed to fit certain ERGMs may be derived from egocentrically sampled data and used to simulate networks consistent with egocentric observations. This approach has been used in applied contexts (Morris et al., 2009; Goodreau et al., 2010; Smith, 2012). What remains lacking, however, is a general, rigorous framework for ERGM inference for such data, and we turn to the pseudo-MLE (PMLE)[1] (Binder, 1983; Pfeffermann, 1993, for example) approach to develop one.

Outline

The rest of the article proceeds as follows. In Section 2, we describe the notation and the sampling framework for the egocentrically sampled network data, and in Section 3, we specify an ERGM subfamily amenable to being fit to such data. The pseudo-MLE for $\boldsymbol{\theta}$ and its asymptotic properties are derived in Section 4, along with how its uncertainty may be quantified. An overview of implementation issues and of a validating simulation study are given in Sections 5 and 6, respectively, with the details left to the Appendices. Finally, in Section 7, we apply our developments to the question of the impact of network structure on persistent racial disparities in HIV prevalence in the US.

2. Notation and sampling

Let N be the population being studied: a very large, but finite, set of actors whose relations are of interest, and let $\boldsymbol{x}_i$ be a vector of attributes (e.g., age, sex, race) of an actor $i \in N$, with $\boldsymbol{x}_N$ (or just $\boldsymbol{x}$, when there is no ambiguity) being the attributes of actors in N. Let $\mathbb{Y}(N) \equiv \{\{i, j\} : (i, j) \in N \times N \wedge i \neq j\}$ (distinct unordered pairs of actors) be the set of *dyads* (potential ties)

[1]This is not to be confused with the maximum pseudolikelihood estimation (MPLE) of Strauss and Ikeda (1990), the technique for approximating the MLE for an intractable likelihood for fully observed networks. We do not make direct use of it in this work.

in an undirected network of these actors. Then, let $\mathcal{Y}(N, \boldsymbol{x}) \subseteq 2^{\mathbb{Y}(N)}$ (set of subsets of potential ties) be the set of networks (sets of ties) of interest. $\mathcal{Y}(\cdot, \cdot)$ may incorporate exogenous constraints, which we discuss in Section 3.2. For a network $\boldsymbol{y} \in \mathcal{Y}(N, \boldsymbol{x})$, let $y_{i,j} \equiv y_{j,i}$ be an indicator function of whether a tie between i and j is present in $\boldsymbol{y}$ and $\boldsymbol{y}_i = \{j \in N : y_{i,j} = 1\}$, the set of i's network neighbors.

Throughout, $\boldsymbol{y}$ will refer to what we will call the *population network*: a fixed but unknown network of relationships of interest.

2.1. Egocentric data

Now, let $\boldsymbol{e}_i$ be the "egocentric" view of network $\boldsymbol{y}$ from the point of view of actor i ("ego"). It comprises $\boldsymbol{e}_i^{\mathrm{e}} \equiv \boldsymbol{x}_i$: i's own attributes, and $\boldsymbol{e}_i^{\mathrm{a}} \equiv (\boldsymbol{x}_j)_{j \in \boldsymbol{y}_i}$: an unordered list (technically, a multiset) of attribute vectors of i's immediate neighbors ("alters"), but *not* their identities (indices in N). For convenience, we refer to the kth attribute/covariate observed on ego i and its alters as $e_{i,k}^{\mathrm{e}} \equiv x_{i,k}$ and $\boldsymbol{e}_{i,k}^{\mathrm{a}} \equiv (x_{j,k})_{j \in \boldsymbol{y}_i}$.

Then, $[\boldsymbol{e}_i]_{i \in N}$ ($\boldsymbol{e}_N$ for short) represents the *egocentric census*, the information retained by the minimal egocentric sampling design discussed in Section 1.2. The information about $\boldsymbol{y}$ contained in an *egocentric sample* of actors $S \subseteq N$ can then be represented as $\boldsymbol{e}_S \equiv [\boldsymbol{e}_i]_{i \in S}$.

2.2. Sampling design considerations

In the following developments, we will assume that egocentric observations are sampled using a conventional sampling design, with N as the sampling frame, though as we discuss in Section 5, this is not critical in practice. The proposed methods can be applied to more complex—stratified, for example—designs, but here, we focus on simple probability designs, and designs that can be approximated with simple probability designs. Specifically, let inclusion probabilities $\pi_i \equiv \Pr(i \in S)$, for $i \in N$, and assume that a weight $w_i \propto \pi_i^{-1}$ is observed for each ego $i \in S$, but only up to proportion: $\sum_{i \in N} w_i$ is not known. In our application, in particular, $\boldsymbol{w}_S$ incorporate both stratification for oversampling and post-stratification to account for missing reports, making inclusion probabilities π_i difficult to obtain.

Analogously to the ERGM process, we will use $\mathrm{E}_S(\cdot)$ and $\mathrm{var}_S(\cdot)$ to refer to the expectation and the variance under the sampling process.

TABLE 1

Examples of egocentric statistics for undirected networks. $x_{i,k}$ may be a dummy variable indicating i's membership in a particular exogenously defined group. $h_k(\boldsymbol{e}_i)$ that sum over ties are halved because each tie is observed egocentrically twice: once at each end.

Statistic	$g_k(\boldsymbol{y}, \boldsymbol{x})$	$h_k(\boldsymbol{e}_i)$
General sum over ties	$\sum_{(i,j)\in\boldsymbol{y}} f_k(\boldsymbol{x}_i, \boldsymbol{x}_j)$	$\frac{1}{2}\sum_{\boldsymbol{z}\in\boldsymbol{e}_i^{\mathrm{a}}} f_k\left(\boldsymbol{e}_i^{\mathrm{e}}, \boldsymbol{z}\right)$
Number of ties in the network	$\lvert\boldsymbol{y}\rvert \equiv \sum_{(i,j)\in\boldsymbol{y}} 1$	$\frac{1}{2}\lvert\boldsymbol{e}_i^{\mathrm{a}}\rvert$
weighted by actor covariate $x_{i,k}$	$\sum_{(i,j)\in\boldsymbol{y}}(x_{i,k}+x_{j,k})$	$\frac{1}{2}\left(e_{i,k}^{\mathrm{e}}\lvert\boldsymbol{e}_i^{\mathrm{a}}\rvert + \sum_{z\in\boldsymbol{e}_{i,k}^{\mathrm{a}}} z\right)$
weighted by difference in $x_{i,k}$	$\sum_{(i,j)\in\boldsymbol{y}}\lvert x_{i,k}-x_{j,k}\rvert$	$\frac{1}{2}\sum_{z\in\boldsymbol{e}_{i,k}^{\mathrm{a}}}\lvert e_{i,k}^{\mathrm{e}}-z\rvert$
within groups identified by $x_{i,k}$	$\sum_{(i,j)\in\boldsymbol{y}} 1_{x_{i,k}=x_{j,k}}$	$\frac{1}{2}\sum_{z\in\boldsymbol{e}_{i,k}^{\mathrm{a}}} 1_{e_{i,k}^{\mathrm{e}}=z}$
General sum over actors	$\sum_{i\in N} f_k\left\{\boldsymbol{x}_i, (\boldsymbol{x}_j)_{j\in\boldsymbol{y}_i}\right\}$	$f_k\left(\boldsymbol{e}_i^{\mathrm{e}}, \boldsymbol{e}_i^{\mathrm{a}}\right)$
Number of actors with d neighbors	$\sum_{i\in N} 1_{\lvert\boldsymbol{y}_i\rvert=d}$	$1_{\lvert\boldsymbol{e}_i^{\mathrm{a}}\rvert=d}$
weighted by actor covariate $x_{i,k}$	$\sum_{i\in N} x_{i,k} 1_{\lvert\boldsymbol{y}_i\rvert=d}$	$x_{i,k} 1_{\lvert\boldsymbol{e}_i^{\mathrm{a}}\rvert=d}$

3. Egocentric ERGMs

Even if the whole population is observed (i.e., $S = N$, a census), not every ERGM can be fit to such data, and we turn to the notion of sufficiency to identify those that can be. Define an ERGM of the form (1.1) to be *egocentric* if both its sufficient statistic and its sample space constraints (if any) can be recovered from an egocentric census. We discuss them in turn.

3.1. Egocentric statistics

We call a network statistic $g_k(\cdot,\cdot)$ *egocentric* if it can be expressed as

$$g_k(\boldsymbol{y}, \boldsymbol{x}) \equiv \sum_{i\in N} h_k(\boldsymbol{e}_i), \tag{3.1}$$

for some function $h_k(\cdot)$ of egocentric information associated with a single actor. The space of egocentric statistics includes *dyadic-independent* (Hunter et al., 2008b) statistics that can be expressed in the general form of $g_k(\boldsymbol{y}, \boldsymbol{x}) = \sum_{(i,j)\in\boldsymbol{y}} f_k(\boldsymbol{x}_i, \boldsymbol{x}_j)$ for some symmetric function $f_k(\cdot,\cdot)$ of two actors' attributes; and some *dyadic-dependent* statistics that can be expressed as $g_k(\boldsymbol{y}, \boldsymbol{x}) = \sum_{i\in N} f_k\{\boldsymbol{x}_i, (\boldsymbol{x}_j)_{j\in\boldsymbol{y}_i}\}$ for some function $f_k(\cdot,\cdots)$ of the attributes of an actor and their network neighbors. Table 1 gives their representations in terms of of $h_k(\cdot)$, along with some examples. Egocentric statistics induce at most Markov graph dependence (Frank and Strauss, 1986) and are *local* by the definition of Krivitsky et al. (2011).

Statistics that are not egocentric include statistics for triadic closure, degree assortativity (e.g., whether high-degree actors tend to link with other high-degree actors), and 4-cycles. Other statistics that are not egocentric

include the average number of neighbors of an actor—$g_k(\boldsymbol{y}, \boldsymbol{x}) = 2|\boldsymbol{y}|/|N|$—because the corresponding $h_k(\boldsymbol{e}_i) = 2 \times \frac{1}{2}|\boldsymbol{e}_i^{\mathrm{a}}|/|N|$ depends on the network size, which is information not contained in $\boldsymbol{e}_i$. (That is, an individual cannot see exactly how big the network of interest is.) The latter are thus not *local* by the definition of Krivitsky et al. (2011). (This does not mean that the mean degree itself cannot be estimated from egocentric data, only that our inferential results may not apply.)

3.2. Egocentric sample space constraints

We call the sample space $\mathcal{Y}(\cdot, \cdot)$ of an ERGM *egocentric* if it can be expressed as

$$\mathcal{Y}(N, \boldsymbol{x}) \equiv \left\{ \boldsymbol{y} \in 2^{\mathbb{Y}(N)} : \prod_{i \in N} \mathcal{H}(\boldsymbol{e}_i) \neq 0 \right\},$$

for some indicator function $\mathcal{H}(\cdot)$ that depends only on egocentric information associated with a single actor. For example, $\mathcal{H}(\boldsymbol{e}_i) = 1_{|\boldsymbol{e}_i^{\mathrm{a}}| \leq d}$ would constrain $\boldsymbol{y} \in \mathcal{Y}(N, \boldsymbol{x})$ so that no actor has more than d ties; and, given a binary actor attribute $x_{i,k}$ (e.g., sex), $\mathcal{H}(\boldsymbol{e}_i) = \prod_{z \in \boldsymbol{e}_{i,k}^{\mathrm{a}}} 1_{e_{i,k}^{\mathrm{e}} \neq z}$ would force all of the ties to be between groups defined by $x_{i,k}$, modeling a bipartite network (if, say, the focus were on heterosexual partnerships).

For the remainder of this paper, we will fix $\mathcal{H}(\boldsymbol{e}_i) = 1$ so that $\mathcal{Y}(N, \boldsymbol{x}) = 2^{\mathbb{Y}(N)}$: our data include same-sex ties, and statistics $\boldsymbol{g}(\cdot, \cdot)$ with free parameters can be used to model the above-described features more flexibly. Also, hard constraints are less well understood, and techniques such as network size adjustment needed for the computational approach described in Section 5 have not been developed for even the simpler ones.

4. Inference

Our inferential goal is to fit ERGMs to unobserved networks based on egocentric samples from them: to recover the parameters that would have been estimated had an ERGM been fit to fully observed $\boldsymbol{y}$. Because $\boldsymbol{y}$ and $\boldsymbol{x}$ are fixed, we will drop them from $\boldsymbol{g}(\boldsymbol{y}, \boldsymbol{x})$ (i.e., $\boldsymbol{g}$) and others from now on, unless it is to emphasize the dependence.

Most treatments of ERGM estimation treat $\boldsymbol{\theta}$ as a parameter of a superpopulation process of which $\boldsymbol{y}$ is a single realization; and the maximum likelihood estimator (MLE) $\hat{\boldsymbol{\theta}}$ is obtained by solving the score equation

$$\mathrm{sc}(\hat{\boldsymbol{\theta}}) \equiv \boldsymbol{g}(\boldsymbol{y}) - \boldsymbol{\mu}_{\boldsymbol{g}}(\hat{\boldsymbol{\theta}}) = \boldsymbol{0}, \tag{4.1}$$

which has a unique solution $\hat{\boldsymbol{\theta}} = \boldsymbol{\theta}_{\boldsymbol{g}}\{\boldsymbol{g}(\boldsymbol{y})\}$. When the likelihood contains an intractable normalizing constant $\kappa_{\boldsymbol{g}}(\cdot)$ (which also makes $\boldsymbol{\mu}_{\boldsymbol{g}}(\cdot)$ and $\boldsymbol{\theta}_{\boldsymbol{g}}(\cdot)$ intractable), Monte-Carlo Maximum Likelihood Estimation (MCMLE) techniques of Geyer and Thompson (1992), as applied to ERGMs by Hunter and Handcock (2006), can be used. The variance of $\hat{\boldsymbol{\theta}}$ is then typically estimated by the inverse of the simulated negative Hessian of the log-likelihood.

In contrast, we treat $\boldsymbol{\theta}$ as a finite population parameter, defined implicitly for the unobserved population network $\boldsymbol{y}$ as the solution to (4.1). The inverse-negative-Hessian is not the correct variance for this estimation problem: whereas it reflects, loosely, the uncertainty in estimates due to the stochasticity of the generative process for the network, we treat the network as a fixed, unknown, finite population, so it is not a source of uncertainty in the first place. Rather, uncertainty comes from having to estimate $\boldsymbol{g}$ from an egocentric sample $\boldsymbol{e}_S$. Indeed, if $S = N$, (3.1) gives $\boldsymbol{g}$ exactly so $\mathrm{var}_S(\hat{\boldsymbol{\theta}}) = \mathbf{0}$. (We do address the superpopulation case in the Discussion.)

4.1. Pseudo maximum likelihood estimation

Following Binder (1983), substituting (3.1) into (4.1) gives a score equation of the form of Binder's eq. 2.6. Binder's Assumptions (a) (open parameter space), (c) (smoothness), and (d) (continuity of variance) are guaranteed by finite exponential family proprieties of ERGMs.

Assumption (b) calls for an asymptotically normal estimator of the population total $\boldsymbol{g}$ and a consistent estimator of its variance. For our design, we use the inverse-probability weighted estimator (*Hájek estimator*) scaled to the population size. (Hájek, 1971) With $\boldsymbol{y}$, $\boldsymbol{x}$, and therefore $\boldsymbol{g}$ being fixed, and letting $w_\cdot \equiv \sum_{i=1}^{|S|} w_i$,

$$\bar{\boldsymbol{g}}(\boldsymbol{e}_S) \equiv \textstyle\sum_{i\in S} w_i \boldsymbol{h}(\boldsymbol{e}_i)/w_\cdot \tag{4.2}$$

is a design-consistent—if slightly biased—estimator of $\bar{\boldsymbol{g}} \equiv \boldsymbol{g}/|N|$, the population mean contribution of each actor to the sufficient statistic. (Fuller, 2011, p. 61) Scaling it to the population size, $\tilde{\boldsymbol{g}}(\boldsymbol{e}_S) \equiv |N|\bar{\boldsymbol{g}}(\boldsymbol{e}_S)$ is then a design-consistent estimator for the population network statistic $\boldsymbol{g}$. Provided the joint distribution of $(w_i, w_i\boldsymbol{h}(\boldsymbol{e}_i))$ under the sampling process in Section 2.2 is not degenerate and the fourth moments of w_i and $\boldsymbol{h}(\boldsymbol{e}_i)$ are finite, Fuller (2011, Thm. 1.3.8, pp. 58–61) gives

$$|S|^{\frac{1}{2}}(\tilde{\boldsymbol{g}}(\boldsymbol{e}_S) - \boldsymbol{g}) = |N||S|^{\frac{1}{2}}(\bar{\boldsymbol{g}}(\boldsymbol{e}_S) - \bar{\boldsymbol{g}}) \xrightarrow{d} \mathrm{MVN}_p(\mathbf{0}, |N|^2\boldsymbol{\Sigma}_{\mathrm{H}}),$$

where

$$\boldsymbol{\Sigma}_{\mathrm{H}} \equiv \mu_w^{-2} \left(\bar{\boldsymbol{g}}\bar{\boldsymbol{g}}^\top \Sigma_{w,w} - \bar{\boldsymbol{g}}\boldsymbol{\Sigma}_{w,w\boldsymbol{h}} - \boldsymbol{\Sigma}_{w\boldsymbol{h},w}\bar{\boldsymbol{g}}^\top + \boldsymbol{\Sigma}_{w\boldsymbol{h},w\boldsymbol{h}} \right), \tag{4.3}$$

with $\mu_w \equiv \mathrm{E}_S(w_i)$, the expected sampling weight, and

$$\begin{bmatrix} \Sigma_{w,w} & \boldsymbol{\Sigma}_{w,w\boldsymbol{h}} \\ \boldsymbol{\Sigma}_{w\boldsymbol{h},w} & \boldsymbol{\Sigma}_{w\boldsymbol{h},w\boldsymbol{h}} \end{bmatrix} \equiv \boldsymbol{\Sigma}_{[w,w\boldsymbol{h}]} \equiv \mathrm{var}_S \left(\begin{bmatrix} w_i \\ w_i\boldsymbol{h}(\boldsymbol{e}_i) \end{bmatrix} \right).$$

Then, the PMLE $\tilde{\boldsymbol{\theta}} = \boldsymbol{\theta}_{\boldsymbol{g}}\{\tilde{\boldsymbol{g}}(\boldsymbol{e}_S)\}$ solving $\widetilde{\mathrm{sc}}(\tilde{\boldsymbol{\theta}}) = \tilde{\boldsymbol{g}}(\boldsymbol{e}_S) - \boldsymbol{\mu}_{\boldsymbol{g}}(\tilde{\boldsymbol{\theta}}) = \mathbf{0}$ is a consistent, asymptotically normal estimator of $\boldsymbol{\theta}$ (Binder, 1983):

$$|S|^{\frac{1}{2}}(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}) \xrightarrow{d} \mathrm{MVN}_p \left(\mathbf{0}, \{\nabla_{\boldsymbol{\theta}}\boldsymbol{\mu}_{\boldsymbol{g}}(\boldsymbol{\theta})\}^{-1}|N|^2\boldsymbol{\Sigma}_{\mathrm{H}}[\{\nabla_{\boldsymbol{\theta}}\boldsymbol{\mu}_{\boldsymbol{g}}(\boldsymbol{\theta})\}^{-1}]^\top \right). \tag{4.4}$$

Notably, $\boldsymbol{\theta}_{\boldsymbol{g}}\{\tilde{\boldsymbol{g}}(\boldsymbol{e}_S)\}$ is defined for every $\tilde{\boldsymbol{g}}(\boldsymbol{e}_S)$ in the convex hull of $\boldsymbol{g}(\mathcal{Y}; \boldsymbol{x})$ (the set of sufficient statistics attainable in the model's sample space), so $\tilde{\boldsymbol{\theta}}$ is defined even when, say, $\tilde{\boldsymbol{g}}(\boldsymbol{e}_S)$ estimates a fractional number for a network statistic that is a count (like $|\boldsymbol{y}|$), and MCMLE can be used in this situation without modification. (Hummel, Hunter, and Handcock, 2012)

4.2. Estimating the variance of the PMLE

We briefly turn to the question of how each component of the expression for the asymptotic variance in (4.4) can be estimated in practice. $\boldsymbol{\Sigma}_{[w,w\boldsymbol{h}]}$ can be estimated directly with the sample variance–covariance matrix of observed w_i and $w_i\boldsymbol{h}(\boldsymbol{e}_i)$; $\bar{\boldsymbol{g}}$ with $\bar{\boldsymbol{g}}(\boldsymbol{e}_S)$; μ_w with $\bar{w} \equiv w_./|S|$; and substituting these into (4.3) gives $\tilde{\boldsymbol{\Sigma}}_{\mathrm{H}}$, an estimator for $\boldsymbol{\Sigma}_{\mathrm{H}}$. $\nabla_{\boldsymbol{\theta}}\boldsymbol{\mu}_{\boldsymbol{g}}(\boldsymbol{\theta})$ can be approximated by $\nabla_{\boldsymbol{\theta}}\boldsymbol{\mu}_{\boldsymbol{g}}(\tilde{\boldsymbol{\theta}})$, which can be estimated as a byproduct of the likelihood maximization using MCMLE (e.g., Hunter and Handcock, 2006, eq. 3.5). In particular, for a minimal exponential family, $\nabla_{\boldsymbol{\theta}}\boldsymbol{\mu}_{\boldsymbol{g}}(\boldsymbol{\theta}) = -\nabla_{\boldsymbol{\theta}}\,\mathrm{sc}(\boldsymbol{\theta}) = \mathrm{E}_{\boldsymbol{g}}\{\mathrm{sc}(\boldsymbol{\theta})\,\mathrm{sc}(\boldsymbol{\theta})^\top; \boldsymbol{\theta}\} = \mathrm{var}_{\boldsymbol{g}}\{\boldsymbol{g}(\boldsymbol{Y}); \boldsymbol{\theta}\}$, so $\nabla_{\boldsymbol{\theta}}\boldsymbol{\mu}_{\boldsymbol{g}}(\boldsymbol{\theta})$ can be approximated by the sample variance–covariance matrix of $\boldsymbol{g}(\boldsymbol{Y})$ simulated at $\tilde{\boldsymbol{\theta}}$. That is,

$$\mathrm{var}_S(\tilde{\boldsymbol{\theta}}) \approx [\widetilde{\mathrm{var}}_{\boldsymbol{g}}\{\boldsymbol{g}(\boldsymbol{Y}); \tilde{\boldsymbol{\theta}}\}]^{-1}(|N|^2\tilde{\boldsymbol{\Sigma}}_{\mathrm{H}}/|S|)[\widetilde{\mathrm{var}}_{\boldsymbol{g}}\{\boldsymbol{g}(\boldsymbol{Y}); \tilde{\boldsymbol{\theta}}\}]^{-1}, \tag{4.5}$$

an estimator of the form of Binder (1983, eq. 3.4).

5. Implementation

Section 4 leads to the following procedure:

1. Estimate the sufficient statistic of the ERGM with $\tilde{\boldsymbol{g}}(\boldsymbol{e}_S)$.
2. Obtain $\tilde{\boldsymbol{\theta}}$, using MCMLE to solve $\widetilde{\mathrm{sc}}(\tilde{\boldsymbol{\theta}}) = \mathbf{0}$.
3. As a byproduct of Step 2, obtain $\widetilde{\mathrm{var}}_{\boldsymbol{g}}\{\boldsymbol{g}(\boldsymbol{Y}); \tilde{\boldsymbol{\theta}}\}$.
4. Estimate $\mathrm{var}_S(\tilde{\boldsymbol{\theta}})$ as described in Section 4.2.

For the simulation study and the analysis that follow, we use, mainly, the R (R Core Team, 2013) package `ergm` (Hunter et al., 2008b; Handcock et al., 2014) for fitting and simulating from ERGMs. The extensions to fit ERGMs to egocentrically sampled data have been implemented in a new R package, `ergm.ego`, under development for public release. We also use the R package `sna` (Butts, 2008) to calculate network connected component sizes.

Some additional implementation challenges arise as well.

Reconstructing $\boldsymbol{x}_N$ from sampled data Formally, our procedure depends on $\boldsymbol{x}$ being observed completely (i.e., a census), or, at least, its distribution being known to a very high degree of accuracy. Step 2's MCMLE, in particular, requires sampling over the space of possible population networks, conditional on all actor attributes, and its implementation requires constructing a network having actor attributes $\boldsymbol{x}_N$, which is unobserved. While this may seem like a major obstacle, in practice it is not: for $i \in S$, $\boldsymbol{x}_i$ are observed directly, and for the remainder, only a distribution of $\boldsymbol{x}$ is needed: actors having the same $\boldsymbol{x}_i$ are interchangeable.

Therefore, in the analyses performed here, we use the design-based estimator of the finite-population distribution of $\boldsymbol{x}_N$: we replicate each $\boldsymbol{x}_i$ for $i \in S$ as close to $|N|w_i/w_\cdot$ times as possible. This has consequences, which we illustrate in the simulation study in Section 6 and Appendix B.

Scalable estimation The procedure also calls for fitting an ERGM to a network of size $|N|$, often a computationally infeasible task. For example, the "population" of the NHSLS study we consider below is all individuals aged 18 through 59 and living in the US at the time of the study (1992)—hundreds of millions. We work around this using the network-size-invariant parametrization of Krivitsky et al. (2011): by adding an offset term, some ERGMs can be adjusted so that fitting them to networks having similar structure and composition but different sizes produces the same parameter estimates. We thus construct a "scaled-down" pseudopopulation of interest, N', and fit the adjusted model to it, thus approximating the $\tilde{\boldsymbol{\theta}}$ that would have been obtained by fitting to the full N. This approach requires that the model be amenable to such an adjustment, and this can be tested by simulation. Further details are given in Appendix A.

Thus, using the network-size-invariant parametrization, and estimating

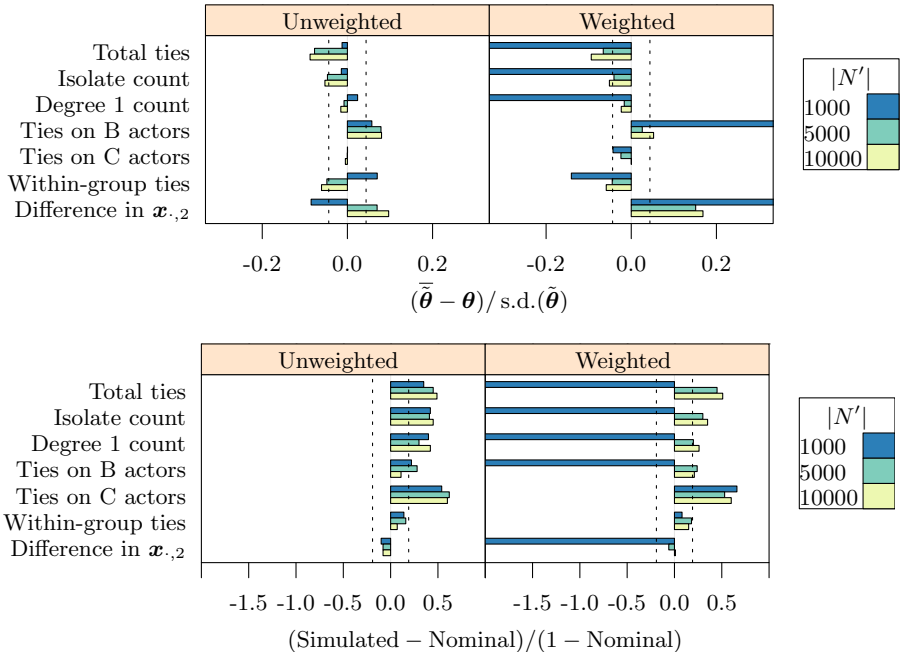


Fig 1: Simulated bias in the point estimates, relative to simulated standard deviation (top); and 95% confidence interval coverage, relative to the miss probability of 5% (bottom) for $|S| = 1{,}000$. Dashed lines are level 0.05 critical values: 95% of the results should fall between them, if the estimator is unbiased / has nominal coverage.

$\boldsymbol{x}_N$ from $\boldsymbol{x}_S$, the procedure does not require any information not in $\boldsymbol{e}_S$.

6. A simulation study

To evaluate the properties of our estimators we performed a simulation study, constructing a large population network with known ERGM parameters and simulating egocentric samples from it, using two sampling designs: unweighted and weighted. The sampling weights have a range similar to the NHSLS, oversample some groups of actors (A and C), and are correlated with a continuous covariate used in the model ($\boldsymbol{x}_{i,2}$). For each sample, we calculated point estimates and standard errors in order to assess their accuracy and the coverage of Wald confidence intervals. Details and full results are given in Appendix B.

Selected bias and coverage results for sample size $|S| = 1{,}000$ are shown

in Figure 1. The unweighted sampling estimates display some bias, though it does not appear to have a systematic pattern as a function of $|N'|$ or model term. None of the estimated biases are greater than 10% of the standard deviation of $\tilde{\boldsymbol{\theta}}$ under repeated sampling; that is, bias accounts for less than 1% of the mean squared error (MSE) of the estimator.

The weighted sampling estimators are, as one would expect, highly biased for smaller $|N'|$. For the largest $|N'|$, the bias tends to approach that of the unweighted, and the most biased parameter's (Difference in $\boldsymbol{x}_{\cdot,2}$) bias is less than 20% of its standard deviation ($\approx 4\%$ of MSE). A possible reason why it is the most biased is that egos with small $\boldsymbol{x}_{i,2}$ are by design severely undersampled, which means that there will exist many samples where the full range of $\boldsymbol{x}_{\cdot,2}$ is not represented. This is likely to be less problematic in real-world applications like the analysis in Section 7, where continuous covariates (like age) have an explicit range of interest.

Overall, we found the standard errors for both weightings to be conservative, overestimating the simulated standard deviation by between 1% and 20% (in a few cases). The resulting confidence interval coverage is consistent with these observations: for almost all terms, the intervals are somewhat conservative for both sampling designs (given sufficient $|N'|$).

We replicated the study for $|S| = 2{,}000$, and found that the biases decrease (both absolutely, and relative to their standard deviation, which is, itself, smaller), and the standard errors became more accurate as well. The coverage remains somewhat conservative. (See Appendix B.2.)

7. Understanding persistent racial disparities in HIV prevalence in the US

We now return to our motivating questions: 1) How strong is the race homophily in the the population? 2) Are there differences in the propensity towards monogamy and concurrency for the races and the sexes? And, 3) What impact do these network features have on overall network connectivity and differentials in network exposure by race and sex?

7.1. Data

The National Health and Social Life Survey (NHSLS) of 1992 (Laumann et al., 1992, 1994) was undertaken at the start of the AIDS epidemic in the US. The objectives of the study included obtaining the data on sexual behavior necessary to predict the long term trajectory of HIV and AIDS

prevalence, and to understand the disparities in HIV prevalence by race that had already begun to emerge. The survey collected, among other information, a representative egocentric sample of sexual partnerships of a stratified sample of residents of the US aged 18–59 (inclusive). A rich set of individual attributes, including respondent's age, sex, race/ethnicity were recorded, and respondents were asked for similar information about all their sexual partners from the past one year.

For this analysis, we focus on modeling the cross-sectional network of ongoing sexual partnerships of the residents of the US aged 18–59 (inclusive) in 1992, the age range for which this study was designed. Some reported partners were outside this age range, and a small number of respondents had turned 60 by the time they were interviewed. We exclude these partnerships and persons from the analysis. We also restrict analysis to cases with complete data on necessary attributes (i.e., race, sex and age). More appropriate handling of missing actor data in egocentrically sampled networks is subject for future research; for this analysis, we exclude egos who have missing attributes of interest for themselves or any of their alters. These exclusions lead to dropping 75 egos and 215 alters, leaving 3,357 egos and 2,555 alters in the sample for analysis. (We had expected egos with more partners to be excluded more often as a result, but we observed no such differential.)

The NHSLS study used a stratified multistage cluster sample, with oversampling of Black households. The public dataset includes weights that account for both stratification and attribute-based non-response, so we approximate the design by an independent weighted sample.

7.2. Methods and models

We divide the respondents into three racial/ethnic categories: White, Black, and Other. While the primary contrast of interest here is between Whites and Blacks, a significant fraction of egos reported other identifications for themselves and their alters. These cannot be dropped in a network analysis, as they can serve as connecting elements that influence the measures of interest.

Homophily is operationalized as an edge covariate, and is defined as concordance in actor attributes in a partnership, as reported by ego. We focus on homophily by sex and race in this analysis, allowing for differential homophily by race. Concurrency is operationalized at the actor level, and is defined as actor degree greater than 1. For modeling purposes, we fit a monogamy term to capture these effects, defined as actor degree equal to 1, again allowing for group-specific propensities for monogamy. This produces

more stable estimates, especially for smaller groups with very low rates of reported concurrency.

We fit a sequence of nested models to test the network hypothesis for the racial disparities in HIV (and other STIs). *Model 1* serves as a baseline, fitting the observed mean degree for each sex by race, as well as the prevalence of heterosexual mixing. It has terms for the main effects for each sex and each race, and a homophily term for sex. Since this is a largely heterosexual population, we expect the sex homophily term will be strongly negative, but same-sex partnerships are not precluded. This model assumes partners are selected at random with respect to race, and there is no propensity for monogamy in sexual partnerships. *Model 2* tests homophily by race by adding a term for each race to capture the prevalence of within-group mixing. We expect these terms to be large and positive. *Model 3* tests heterogeneities in the propensity for monogamy by adding a term for each sex by race to capture the prevalence of persons with exactly one partner. We expect these terms to be positive, given the strong norm of monogamy in sexual partnerships, but we also expect there to be significant differences by race and sex. Since the group-specific mean degrees have been fit by the baseline terms, lower coefficient values on the monogamy terms will imply higher prevalence of concurrent partners.

We evaluate the goodness of fit for each model by comparing the observed degree distribution to 100 realizations of complete networks from the specified model. (Hunter, Goodreau, and Handcock, 2008a) In principle, this approach allows us to evaluate the goodness of fit to any egocentric statistics. We choose the degree distribution because it is a primary determinant of network connectivity. A model that does not fit the degree distribution well is very unlikely to produce the unobserved network connectivity that we wish to infer.

We also use the simulated networks to evaluate the network hypothesis, comparing the overall network connectivity and group-specific network exposure differentials predicted by each model. The overall network connectivity is measured as the component size distribution. The propensity for monogamy in *Model 3* is expected to increase the number of components of size 2 (mutual monogamy) and decrease the number and size of the larger components. We measure network exposure at the actor level, using the probability of membership in components of size 3 or greater. This represents the risk of indirect exposure: an actor may have only one partner, so have little direct exposure, but by virtue of the network she may still be exposed to her partner's other partner(s) and beyond. Under the network hypothesis, only *Model 3* is expected to produce differentials in network risk

TABLE 2
Coefficients and standard errors for the three models. Coefficients reported are in the presence of an edge count offset of $-\log(44859) = -10.71$.

Model	*1* Main	*2* + Mix.	*3* + Monog.
Actor activity by sex			
Female	0.02 (0.10)	-0.99 (0.19)***	-1.88 (0.31)***
Male	0.46 (0.10)***	-0.55 (0.20)**	-1.18 (0.25)***
Same-sex partnership	-4.49 (0.21)***	-4.50 (0.20)***	-4.52 (0.21)***
Actor activity by race			
White		0 (baseline)	
Black	-0.09 (0.07)	-0.58 (0.29)*	-0.30 (0.38)
Other	-0.03 (0.07)	0.83 (0.33)*	0.93 (0.42)*
Race homophily by race			
Black		5.13 (0.35)***	5.15 (0.38)***
Other		2.06 (0.35)***	2.04 (0.35)***
White		2.25 (0.34)***	2.32 (0.36)***
Monogamy by sex and race			
Black Female			1.80 (0.47)***
Other Female			2.51 (0.67)***
White Female			2.25 (0.31)***
Black Male			0.99 (0.24)***
Other Male			1.40 (0.31)***
White Male			2.16 (0.25)***

Significance levels: $0.05 \geq {}^{*} > 0.01 \geq {}^{**} > 0.001 \geq {}^{***}$

exposure that are consistent with the observed disparities in HIV prevalence.

The population network would have had $|N| \approx 147$ million persons (Population Estimates Program, 2001), necessitating the scaled-down approach mentioned in Section 5. Based on reasoning detailed in Appendix C.1, we select, conservatively, $|N'| \approx 45{,}000 \approx 13.4 \times |S|$, or 44,859 after rounding the scaled sampling weights. This network size is also used in the simulation results we report. Notably, this may be overly conservative in practice: we obtained very similar results in a pilot analysis using $|N'| \approx 15{,}000$.

Verification of the assumption that our models are amenable to network-size-invariant parametrization is given in Appendix C.3.

7.3. Results

We report the model fits in Table 2. *Model 1* results are consistent with expectations. There is a significant and strong propensity for heterosexual ties, and a slightly higher mean degree for men than women. There are no significant differences in mean degree by race. In *Model 2*, the results are consistent with the network hypothesis: all of the race homophily terms are

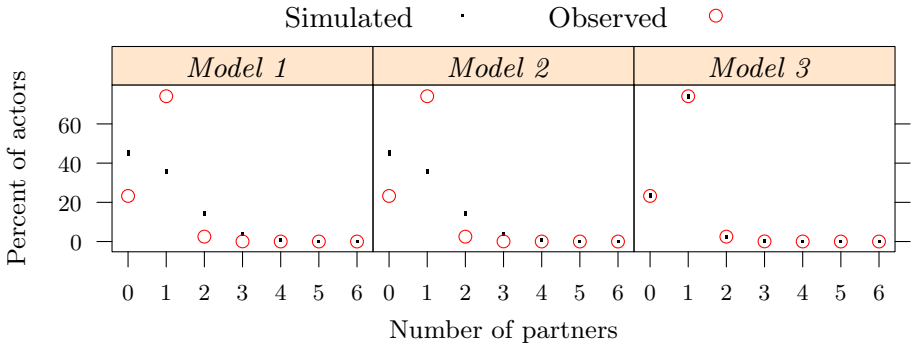
large and significant. In *Model 3*, the results are again consistent with the hypothesis. There is a strong propensity for monogamy in all groups, but the propensity is relatively lower among Black men and women. The difference between Whites and Blacks is significant for men (contrast diff. = 1.17, s.e. = 0.32, P-value < 0.001), but not for women (diff. = 0.45, s.e. = 0.51, P-value > 0.3). Women have higher rates of monogamy than men in all groups, especially among Blacks, but these differences are not statistically significant.

The goodness of fit for each model is shown in Figure 2a. The first two models do a very poor job fitting the observed degree distribution: both underestimate the fraction of persons with only one partner and overestimate both the fraction with no partner, and more than one partner. The data clearly indicate a strong propensity for monogamy, and *Model 3* captures this well. Because there are race and sex-specific monogamy terms in *Model 3*, it provides a good fit for all groups.

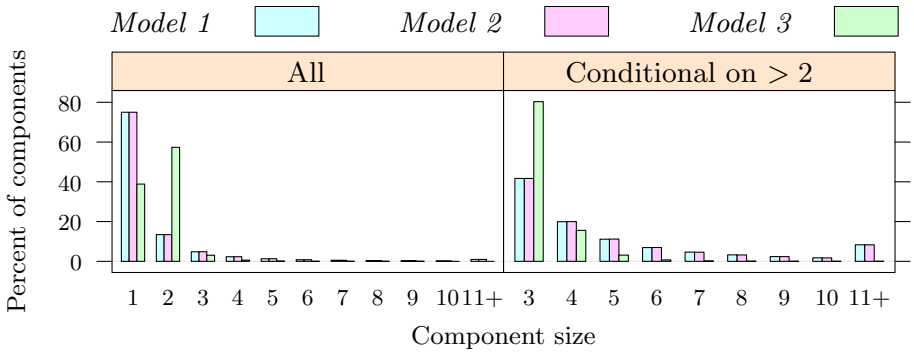
The overall network connectivity predicted by each model can be seen in Figure 2b. The plot shows the distribution of component sizes produced by each model. The first two models are, again, similar: both predict that about three-quarters of the components are size 1 (isolated actors), and the ties distributed to the remaining actors produce components with sizes that can reach 100 or more. By contrast, *Model 3* predicts that the modal component is size 2 (mutual monogamy), that only about 20% of the actors are isolates, and the maximum component size attained in the 100 simulated realizations has fallen to only 6. Monogamy thus has the expected effect: it dramatically reduces the connectivity in the overall network.

The differential network risk exposure by race and sex predicted by each model can be seen in Figure 2c. This plot shows the group-specific distributions of the probability of belonging to a component of size 3 or more for each model. In *Model 1*, overall network exposure probabilities are about 40%. There are some small race-specific differences, with lower probabilities of exposure predicted for Blacks than Whites. Since lower probabilities of exposure imply lower transmission risks, this pattern is the opposite of what would be expected given the HIV/STI prevalence disparities. The pattern is similar for *Model 2*, though the predicted differences by race for women have increased slightly, but are still in the wrong direction. Adding the differential monogamy terms in *Model 3*, however, reverses the predicted network exposure risk differentials for both sexes, producing a pattern that is consistent with the observed racial disparities in HIV/STI prevalence, and consistent with the network hypothesis.

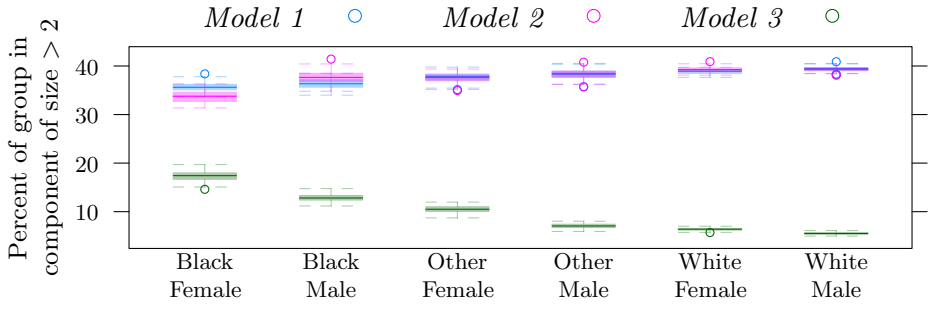
Note that within each racial group, women are more likely to be in com-



(a) Goodness of fit for degree distribution, for each model



(b) Distribution of connected component sizes, simulated from each model



(c) Network exposure, by race and sex, simulated from each model

Fig 2: Simulation results based on 100 realizations from each of the fitted models: (a) goodnesses of fit plot, comparing simulated degree frequencies (dot plot) to that observed in the data; (b) simulated network component size distributions, averaged over each simulation; and (c) simulated distribution of the proportions of individuals of each race and sex who are in components of size 3 or greater. (Because of the large $|N'|$ used in the simulation, there is little variability percent-wise between realizations in (a).)

ponents of size 3 or more than men. This is because 3 is by far the most common component size predicted among those not mutually monogamous ($\approx$ 80%, as seen in Figure 2b), and coupled with the higher rates of concurrency among men than women this means that these components typically comprise 1 man and 2 women. This a good example of the somewhat counterintuitive logic of network exposure in infectious diseases: your exposure is not just a function of your own behavior, but also a function of your partner's. In countries with generalized heterosexual HIV epidemics, such as those in sub-Saharan Africa, concurrency is similarly gendered, and women's HIV prevalence is typically much higher then men's (40% higher across this particular region (UNAIDS, 2014)). Our results suggest that the underlying transmission network structure may contribute to this disparity.

Model 3 is clearly the best fit, and it predicts network exposure risks that are consistent with the observed disparities in HIV prevalence by race and sex. It is, of course, not intended to be a complete specification of the network structure, as it excludes many factors that are known to influence sexual behavior and partner selection. For our purposes, the question is whether the results from *Model 3* are robust to the inclusion of additional factors. That would depend on whether the factors are correlated to race and sex, or interact with them in relevant ways. A good example to consider is age, as it influences both activity levels and partner selection patterns. We examined a model with several age-related terms. The results, given in Appendix C.2, are that age effects are generally significant, but that the key results reported for *Model 3* are robust.

8. Discussion

Theat stochastic dependence in networks can be complex, that information present in an egocentric sample is limited, and that such use of the data is often secondary, all make rigorous inference difficult. Using pseudo-maximum-likelihood estimation and exploiting exponential family proprieties, we have proposed a technique to conduct statistically valid ERGM inference nonetheless. This makes it possible to both estimate the parameters of a generative model for the observed structural properties of a network from a sample, and conduct principled simulation from a superpopulation of networks having properties similar to those observed; and by making use of a network-size-invariant parametrization of ERGMs, this can be done even when the target population is very large or even unknown. The result is a general statistical framework for leveraging whole network information from an efficient minimal sampling design.

At the same time, we made a number of approximations and assumptions, and our methodology has a number of limitations.

$\boldsymbol{x}$ is sampled Per Section 5, we had constructed $\boldsymbol{x}_{N'}$ by extrapolating from the sample, but we did not take this into account in our inference. Our simulation study suggests that the inference is still valid (conservative, in fact), but this issue can be addressed more rigorously. Recall that we only require the joint distribution of $\boldsymbol{x}_{N\setminus S}$, not their individual values. Fortunately, the distribution for demographic attributes such as sex, age, ethnicity, and geographic location is often known to a very high degree of precision—from a national census, for example; and it could be used to construct an $\boldsymbol{x}_{N'}$ with virtually no sampling variation. In fact, the weights in the NHSLS data in Section 7 had been calculated through post-stratification to reflect the population, so, in our analysis, this had already been done for us.

Alternatively, uncertainty from $\boldsymbol{x}$ being sampled may be incorporated into the inferential procedure: in particular, Fellows and Handcock (2012) propose an exponential-family model for jointly modeling actor attributes and ties. Provided the sufficient statistic associated with actor attributes could be recovered from egocentric data, our results should be applicable.

Stratified and cluster sampling We approximated the sampling design of the NHSLS study with that of Section 2.2: a simple probability sample. A more accurate estimate of $\mathrm{var}_S(\tilde{\boldsymbol{\theta}})$ could be obtained by substituting an estimate for $\mathrm{var}_S\{\bar{\boldsymbol{g}}(\boldsymbol{e}_S)\}$ into (4.5) that better reflects the design than $\tilde{\boldsymbol{\Sigma}}_{\mathrm{H}}/|S|$ does. And, for small N, finite-population correction can be used.

Bias Our simulation study in Section 6 and Appendix B shows our estimators to be slightly biased. There are four likely sources of bias: i) sampling variation of $\boldsymbol{x}$; ii) biasedness of the Hájek estimator (4.2) for $\bar{\boldsymbol{g}}$; iii) nonlinearity of the mapping $\boldsymbol{\theta}_{\boldsymbol{g}}(\cdot)$ and Jensen's Inequality (analogous to that in logistic regression (Firth, 1993)); and iv) scaled-down approximation, particularly for weighted samples. Notably, biases (i)–(iii) decrease in sample size $|S|$, while bias (iv) decreases in $|N'|$.

We had attempted to reduce (ii) using jackknife to little noticeable improvement in the simulation studies, suggesting that it is not a major source in this case. Judging by the small difference between the biases from the higher values of $|N'|$ considered (found in Appendix B.2), (iv) likely becomes negligible reasonably quickly: the estimates converge to a nonzero bias. We believe this remaining bias to be primarily due to (i) and (iii).

Unfortunately, while a technique like nonparametric bootstrap jointly resampling $\boldsymbol{x}_i$ and $\boldsymbol{e}_i$ can be used for bias reduction and uncertainty estimation

alike, this is likely to be computationally prohibitive: whereas every resample of bootstrap or jackknife for (ii) requires merely recalculating a weighted average, culminating in a single ERGM fit using debiased $\bar{\boldsymbol{g}}(\boldsymbol{e}_S)$, for (i) and (iii), every resample requires refitting an ERGM to a large network. At the same time, it may be possible to reduce (iii) using the penalized likelihood approach of Firth (1993). All this is subject for ongoing work.

Measurement error In our application, we had assumed that the responses were accurate: that, for example, the male respondents did not overreport their partnerships and female respondents did not underreport. But, it is worth noting that there is almost perfect correspondence between the weighted total number of ongoing heterosexual partnerships reported by women and that reported by men (1366 and 1388, respectively), which suggests some internal validity to the reports.

Directed networks Our development was aimed at undirected relations on unipartite networks, but the general inferential technique should be applicable to directed relations, provided each ego's in-ties, as well as out-ties, are observed, and to bipartite networks. Krivitsky and Kolaczyk (2015) network size adjustment of ERGMs for mutuality could be used for the scaled-down inference for the former, and a similar approach could be developed for the latter.

Higher-order terms We had defined $\boldsymbol{e}_i$ to contain no information about alters' connections. Though less common, data of this type are sometimes available in an egocentric design, through solicitation of alter–alter ties. Other variations and extensions to egocentric studies include studies that collect egocentric-like data but allow some individuals appearing twice in the data to be matched; couple studies (where the two individuals with a link are recruited together and are each asked about their alters); and a one-wave snowball sample. In each case, $\boldsymbol{e}_i$ would contain some additional information, and the set of statistics expressible in the form of (3.1) would expand accordingly, with much of the inferential argument applying directly. The scaled-down inference would then require a network-size-invariant parametrization for higher-order (e.g., triadic) effects, which might not exist, but if the population network is not overly large, it might not be necessary at all.

Inference for a superpopulation Lastly, in our framework, the population network $\boldsymbol{y}$ is fixed and unknown and $\boldsymbol{\theta}_{\boldsymbol{g}}\{\boldsymbol{g}(\boldsymbol{y})\}$ is a finite population property to be estimated. In some applications, it may be more meaningful to

view $\boldsymbol{Y}$ as being drawn from a superpopulation (e.g., ERGM) parametrized by $\boldsymbol{\theta}$, and then observed egocentrically. Although deriving rigorous asymptotics of this generative process may not be feasible, the variance of the estimator is straightforward: whatever the generative process for $\boldsymbol{Y}$, $\tilde{\boldsymbol{g}}\{\boldsymbol{e}_S(\boldsymbol{Y})\}$ remains an asymptotically unbiased estimator of $\boldsymbol{g}(\boldsymbol{Y})$ for any given $\boldsymbol{Y}$, under repeated egocentric sampling from $\boldsymbol{Y}$. Then, Law of Total Variance gives

$$\begin{aligned}\operatorname{var}_{S\circ\boldsymbol{g}}\left[\tilde{\boldsymbol{g}}\{\boldsymbol{e}_S(\boldsymbol{Y})\};\boldsymbol{\theta}\right] &= \operatorname{E}_{\boldsymbol{g}}\left(\operatorname{var}_S[\tilde{\boldsymbol{g}}\{\boldsymbol{e}_S(\boldsymbol{Y})\}|\boldsymbol{Y}];\boldsymbol{\theta}\right) + \operatorname{var}_{\boldsymbol{g}}\left(\operatorname{E}_S[\tilde{\boldsymbol{g}}\{\boldsymbol{e}_S(\boldsymbol{Y})\}|\boldsymbol{Y}];\boldsymbol{\theta}\right)\\ &\approx \operatorname{E}_{\boldsymbol{g}}\left(\operatorname{var}_S[\tilde{\boldsymbol{g}}\{\boldsymbol{e}_S(\boldsymbol{Y})\}|\boldsymbol{Y}];\boldsymbol{\theta}\right) + \operatorname{var}_{\boldsymbol{g}}\left\{\boldsymbol{g}(\boldsymbol{Y});\boldsymbol{\theta}\right\},\end{aligned}$$

which, for an ERGM superpopulation and $S \perp \boldsymbol{Y}$, reduces to

$$\begin{aligned}\operatorname{var}_{S\circ\boldsymbol{g}}(\tilde{\boldsymbol{\theta}}) &\approx \tilde{\boldsymbol{V}}^{-1}(|N|^2\tilde{\boldsymbol{\Sigma}}_{\mathrm{H}}/|S| + \tilde{\boldsymbol{V}})\tilde{\boldsymbol{V}}^{-1}\\ &\approx \tilde{\boldsymbol{V}}^{-1}(|N|^2\tilde{\boldsymbol{\Sigma}}_{\mathrm{H}}/|S|)\tilde{\boldsymbol{V}}^{-1} + \tilde{\boldsymbol{V}}^{-1},\end{aligned}$$

for $\tilde{\boldsymbol{V}} = \widetilde{\operatorname{var}}_{\boldsymbol{g}}\{\boldsymbol{g}(\boldsymbol{Y});\tilde{\boldsymbol{\theta}}\}$, and if $\operatorname{E}_{\boldsymbol{g}}\left(\operatorname{var}_S[\tilde{\boldsymbol{g}}\{\boldsymbol{e}_S(\boldsymbol{y})\}|\boldsymbol{Y}];\boldsymbol{\theta}\right)$ is approximated by $|N|^2\tilde{\boldsymbol{\Sigma}}_{\mathrm{H}}/|S|$. However, while the variance of the parameter estimates may be estimated thus, the normality of $\boldsymbol{g}(\boldsymbol{Y})$ under the superpopulation is not guaranteed. If the superpopulation process is an ERGM, it can be tested as a side-product of the estimation.

References

Admiraal, R. (2009). Dynamic Network Models based on Revealed Preference for Observed Relations and Egocentric Data. Ph.D. thesis, University of Washington, Seattle, WA.

Binder, D. A. (1983). On the Variances of Asymptotically Normal Estimators from Complex Surveys. *Int. Stat. Rev.* **51** 279–292.

Brown, L. D. (1986). *Fundamentals of Statistical Exponential Families with Applications in Statistical Decision Theory. Lecture Notes—Monograph Series* **9**. Institute of Mathematical Statistics, Hayward, California.

Burt, R. S. (1984). Network Items and the General Social Survey. *Soc. Networks* **6** 293–339.

Butts, C. T. (2008). Social Network Analysis with `sna`. *J. Stat. Softw.* **24** 1–51.

Fellows, I. and Handcock, M. S. (2012). Exponential-Family Random Network Models. *arXiv preprint arXiv:1208.0121* .

Firth, D. (1993). Bias Reduction of Maximum Likelihood Estimates. *Biometrika* **80** 27–38.

Frank, O. and Strauss, D. (1986). Markov Graphs. *J. Am. Stat. Assoc.* **81** 832–842.

Fuller, W. A. (2011). *Sampling Statistics. Wiley Series in Survey Methodology* **560**. John Wiley & Sons.

Geyer, C. J. and Thompson, E. A. (1992). Constrained Monte Carlo Maximum Likelihood for Dependent Data (with discussion). *J. R. Stat. Soc. Ser. B* **54** 657–699.

Gjoka, M., Smith, E., and Butts, C. (2014a). Estimating Clique Composition and Size Distributions from Sampled Network Data. In *Sixth IEEE International Workshop on Network Science for Communication Networks*.

Gjoka, M., Smith, E., and Butts, C. T. (2014b). Design-based Estimators for Attribute-Labeled, Low-Semidiameter Subgraphs. In *34th Sunbelt Network Conference*. INSNA, St. Pete Beach.

Goodreau, S., Cassels, S., Kasprzyk, D., Montao, D., Greek, A., and Morris, M. (2010). Concurrent Partnerships, Acute Infection and HIV Epidemic Dynamics Among Young Adults in Zimbabwe. *AIDS Behav.* pages 1–11.

Hájek, J. (1971). Comment on An Essay on the Logical Foundations of Survey Sampling by Basu, Debabrata. In *Foundations of Statistical Inference: Proceedings of the Symposium on the Foundations of Statistical Inference* (V. P. Godambe and D. A. Sprott, eds.). René Descartes Foundation, Holt McDougal, Department of Statistics, University of Waterloo, Ont., Canada, March 31 to April 9, 1970.

Hallfors, D. D., Iritani, B. J., Miller, W. C., and Bauer, D. J. (2007). Sexual and Drug Behavior Patterns and HIV and STD Racial Disparities: The Need for New Directions. *Am. J. Public Health* **97** 125–132.

Handcock, M. S. and Gile, K. J. (2010). Modeling Social Networks from Sampled Data. *Ann. Appl. Stat.* **4** 5–25.

Handcock, M. S., Hunter, D. R., Butts, C. T., Goodreau, S. M., Krivitsky, P. N., and Morris, M. (2014). *ergm: Fit, Simulate and Diagnose Exponential-Family Models for Networks*. The Statnet Project (http://www.statnet.org). R package version 3.1.2.

Hummel, R. M., Hunter, D. R., and Handcock, M. S. (2012). Improving Simulation-Based Algorithms for Fitting ERGMs. *J. Comput. Graph. Stat.* **21** 920–939.

Hunter, D. R., Goodreau, S. M., and Handcock, M. S. (2008a).

Goodness of Fit for Social Network Models. *J. Am. Stat. Assoc.* **103** 248–258.

Hunter, D. R. and Handcock, M. S. (2006). Inference in Curved Exponential Family Models for Networks. *J. Comput. Graph. Stat.* **15** 565–583.

Hunter, D. R., Handcock, M. S., Butts, C. T., Goodreau, S. M., and Morris, M. (2008b). `ergm`: A Package to Fit, Simulate and Diagnose Exponential-Family Models for Networks. *J. Stat. Softw.* **24** 1–29.

Illenberger, J. and Fltter, G. (2012). Estimating Network Properties from Snowball Sampled Data. *Soc. Networks* **34** 701–711.

Koskinen, J. H., Robins, G. L., and Pattison, P. E. (2010). Analysing Exponential Random Graph (p-star) Models with Missing Data Using Bayesian Data Augmentation. *Stat. Methodol.* **7** 366 – 384.

Krivitsky, P. N., Handcock, M. S., and Morris, M. (2011). Adjusting for Network Size and Composition Effects in Exponential-Family Random Graph Models. *Stat. Methodol.* **8** 319–339.

Krivitsky, P. N. and Kolaczyk, E. D. (2015). On the Question of Effective Sample Size in Network Modeling: An Asymptotic Inquiry. *Stat. Sci.* to appear.

Laumann, E. O., Gagnon, J. H., Michael, R. T., and Michaels, S. (1992). National Health and Social Life Survey. Chicago, IL, USA: University of Chicago and National Opinion Research Center [producer], 1995. Ann Arbor, MI, USA: Inter-university Consortium for Political and Social Research [distributor], 2008-04-17. Computer file.

Laumann, E. O., Gagnon, J. H., Michael, R. T., and Michaels, S. (1994). *The Social Organization of Sexuality.* University of Chicago Press, Chicago.

Marsden, P. V. (1981). Models and Methods for Characterizing the Structural Parameters of Groups. *Soc. Networks* **3** 1–27.

Marsden, P. V. (1987). Core Discussion Networks of Americans. *Am. Sociol. Rev.* **52** 122–131.

MEASURE DHS (2000–2014). *Demographic and Health Surveys.* ICF International.

Morris, M. (1991). A Log-Linear Modeling Framework for Selective Mixing. *Math. Biosci.* **107** 349–77.

Morris, M. (1993). Epidemiology and Social Networks: Modeling Structured Diffusion. *Socio. Meth. Res.* **22** 99–126.

Morris, M., Handcock, M. S., Miller, W. C., Ford, C. A., Schmitz, J. L., Hobbs, M. M., Cohen, M. S., Harris, K. M., and Udry, J. R. (2006). Prevalence of HIV Infection among Young Adults in the U.S.: Results from the ADD Health Study. *Am. J. Public Health* **96** 1091–

1097.

Morris, M. and Kretzschmar, M. (1997). Concurrent Partnerships and the Spread Of HIV. *AIDS* **11** 641–648.

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

National Center for HIV/AIDS, Viral Hepatitis, STD, and TB Prevention (NCHHSTP) (2012). HIV Surveillance Supplemental Report: Estimated HIV incidence in the United States, 2007–2010. Tech. Rep. 17(4), Centers for Disease Control and Prevention.

National Center for HIV/AIDS, Viral Hepatitis, STD, and TB Prevention (NCHHSTP) (2013). HIV Surveillaince Supplemental Report: Diagnoses of HIV Infection among Adults Aged 50 Years and Older in the United States and Dependent Areas, 2007–2010. Tech. Rep. 18(3), Centers for Disease Control and Prevention.

National Communicable Disease Center (NCDC) (1967). Morbidity and Mortality Weekly Report: Reported Incidence of Notifiable Diseases in the United States, 1966. Tech. Rep. 15(53), U.S. Department of Health, Education, and Welfare, Atlanta, GA.

National Survey of Family Growth Staff (2002, 2006–2011). National Survey of Family Growth (NSFG). Tech. rep., Division of Vital Statistics, National Center for Health Statistics.

Pattison, P. E., Robins, G. L., Snijders, T. A., and Wang, P. (2013). Conditional Estimation of Exponential Random Graph Models from Snowball Sampling Designs. *J. Math. Psychol.* **57** 284–296.

Pfeffermann, D. (1993). The Role of Sampling Weights when Modeling Survey Data. *Int. Stat. Rev.* **61** 317–337.

Plummer, M., Best, N., Cowles, K., and Vines, K. (2006). `CODA`: Convergence Diagnosis and Output Analysis for MCMC. *R News* **6** 7–11.

Population Estimates Program (2001). Resident Population Estimates of the United States by Age and Sex: April 1, 1990 to July 1, 1999, with Short-Term Projection to November 1, 2000. Population Division, U.S. Census Bureau. Online. Retrieved June 9, 2009.

Putnam, R. D. (2000). *Bowling Alone: The Collapse and Revival of American Community*. Simon & Schuster, New York.

R Core Team (2013). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Salganik, M. J. and Heckathorn, D. D. (2004). Sampling and Estimation in Hidden Populations Using Respondent-Driven Sampling. *Sociol.*

Methodol. **34** 193–239.

Smith, J. A. (2012). Macrostructure from Microstructure: Generating Whole Systems from Ego Networks. *Sociol. Methodol.* **42** 155–205.

Snijders, T. A. (2010). Conditional Marginalization for Exponential Random Graph Models. *J. Math. Sociol.* **34** 239–252.

Strauss, D. and Ikeda, M. (1990). Pseudolikelihood Estimation for Social Networks. *J. Am. Stat. Assoc.* **85** 204–212.

Tanfer, K. (1991). National Survey of Women. In *AIDS/STD Data Archive* (E. A. McKean, K. L. Muller, and E. L. Lang, eds.), 17–19. Sociometrics Corporation, Los Altos, CA.

Thompson, S. K. and Frank, O. (2000). Model-Based Estimation with Link-Tracing Sampling Designs. *Survey Methodol.* **26** 87–98.

Tomas, A. and Gile, K. J. (2011). The Effect of Differential Recruitment, Non-Response and Non-Recruitment on Estimators for Respondent-Driven Sampling. *Electron. J. Stat.* **5** 899–934.

Udry, J. R. (2003). The National Longitudinal Study of Adolescent Health (Add Health), Waves I & II, 1994–1996; Wave III, 2001–2002. Tech. rep., Carolina Population Center, University of North Carolina at Chapel Hill.

UNAIDS (2014). HIV Estimates with Uncertainty Bounds 1990–2013. Tech. rep., United Nations.

Volz, E. and Heckathorn, D. D. (2008). Probability Based Estimation Theory for Respondent Driven Sampling. *J. Off. Stat.* **24** 79–97.

Wasserman, S. S. and Pattison, P. (1996). Logit Models and Logistic Regressions for Social Networks: I. An Introduction to Markov Graphs and p^*. *Psychometrika* **61** 401–425.

Appendix A: Computationally efficient approximation using network size adjustment

Here, we give details for the computational approximation mentioned Section 5 and derive its properties.

Krivitsky, Handcock, and Morris (2011) suggested an approach for a network-size-invariant parametrization for some ERGMs for undirected graphs, where a network of size $|N|$ is modeled with an offset term, i.e.,

$$\Pr_{\boldsymbol{g}}(\boldsymbol{Y} = \boldsymbol{y}; \boldsymbol{x}_N, \boldsymbol{\theta}) \equiv \frac{\exp\{-\log(|N|)|\boldsymbol{y}| + \boldsymbol{\theta}^\top \boldsymbol{g}(\boldsymbol{y}, \boldsymbol{x}_N)\}}{\kappa_{\boldsymbol{g}}(\boldsymbol{\theta}, \boldsymbol{x}_N)}, \boldsymbol{y} \in 2^{\mathbb{Y}(N)}. \quad \text{(A.1)}$$

This adjustment works, particularly, for network processes that fulfill certain heuristics: locality, in that as the network size changes, an individual actor's

egocentric view of the network does not, on average, change; and stable degree distribution and per-capita mixing, in that the distribution of the number of ties an actor has (and the distribution of attributes of those to whom the actor has ties) remain stable as network size changes, provided the composition is preserved. For network processes and ERGM terms fulfilling this, networks having similar structure and composition but different sizes produced the same parameter estimates after the network size adjustment. They demonstrated this rigorously for dyadic-independent ERGM terms and by simulation for degree distribution terms. (Hunter et al., 2008b)

This finding suggests a straightforward computational shortcut: instead of constructing the full population network over actors N, one can construct a "scaled-down" version $N' \subseteq N$ having the same composition (distribution of $\boldsymbol{x}$) and large enough for the estimates to have asymptoted. Fitting (A.1) with N replaced by N' and $\tilde{\boldsymbol{g}}(e_S)$ replaced by $\tilde{\boldsymbol{g}}(e_S) \times |N'|/|N|$ would then yield $\tilde{\boldsymbol{\theta}}^{N'} \approx \tilde{\boldsymbol{\theta}}^{N}$.

A.1. Requirements for the adjustment

By design, the network composition is fixed, with only size changing, so for the adjustment to work in our case, the ERGM must be *local* (which, in this case, holds by construction as described in Section 3.1) and degree distribution and per-capita mixing must be stable. In the context of egocentric ERGMs, this can be operationalized as the distribution of individual measurements $\boldsymbol{h}(\boldsymbol{e}_i)$ being unaffected by the network size. This is a property of the model, not of the data, and from the perspective of the model, this requires that,

$$\lim_{|N'|\to\infty} |N'|^{-1} \boldsymbol{\mu}_{\boldsymbol{g}}^{N'}(\boldsymbol{\theta}) = \boldsymbol{\lambda}_{\boldsymbol{g}}(\boldsymbol{\theta}), \tag{A.2}$$

with $\boldsymbol{\lambda}_{\boldsymbol{g}}(\boldsymbol{\theta})$ being the asymptotic per-capita expected value of $\boldsymbol{h}$, if the distribution of $\boldsymbol{x}_{N'}$ does not change. (Intuitively, consider a sequence of actor attribute sets $\boldsymbol{x}_{N'_1}, \boldsymbol{x}_{N'_2}, \ldots$ such that $\boldsymbol{x}_{N'_i}$ is $\boldsymbol{x}_{N'_1}$ replicated i times.)

Verifying the property (A.2) requires deriving a closed form for $\boldsymbol{\mu}_{\boldsymbol{g}}(\cdot)$—at least asymptotically. Krivitsky et al. (2011, Sec. 4.3) showed this property for some dyadic-independent ERGM terms, but for dyadic-dependent ERGMs, it may not be possible to do so. In practice, this property only needs to hold in the neighborhood of the estimate $\tilde{\boldsymbol{\theta}}$, so it can be can be checked by simulating from the fitted (A.3) at a variety of network sizes with the same distributions of $\boldsymbol{x}_{N'}$, to confirm that $|N'|^{-1}\boldsymbol{\mu}_{\boldsymbol{g}}^{N'}(\boldsymbol{\theta})$ does not vary substantially in $|N'|$ for $|N'|$ large enough.

A.2. Point estimation

A model fit to network of size $|N'|$ approximating the coefficients of a model fit to network of size $|N|$ has the form

$$\Pr_{\boldsymbol{g}}(\boldsymbol{Y}=\boldsymbol{y};\boldsymbol{x}_{N'},\boldsymbol{\theta}) \equiv \frac{\exp\{-\log(|N'|/|N|)|\boldsymbol{y}|+\boldsymbol{\theta}^{\top}\boldsymbol{g}(\boldsymbol{y},\boldsymbol{x}_{N'})\}}{\kappa_{\boldsymbol{g}}(\boldsymbol{\theta},\boldsymbol{x}_{N'})},\ \boldsymbol{y}\in 2^{\mathbb{Y}(N')}, \tag{A.3}$$

for $\boldsymbol{g}(\boldsymbol{y},\boldsymbol{x}_{N'})$ estimated by $\tilde{\boldsymbol{g}}^{N'}(e_S) \equiv \tilde{\boldsymbol{g}}(e_S)\times|N'|/|N|$. Intuitively, the smaller a fraction of $|N|$ that $|N'|$ is, the more positive the offset coefficient on $|\boldsymbol{y}|$ is, forcing $\boldsymbol{\theta}$ to adjust to produce the more sparse network that N would induce. (More concretely, if, for some k, $g_k(\boldsymbol{y})=|\boldsymbol{y}|$, its PMLE coefficient would be shifted by $\log(|N'|/|N|)$). It is still a regular exponential family, so the PMLE can be found by solving

$$\widetilde{\mathrm{sc}}^{N'}(\tilde{\boldsymbol{\theta}}^{N'}) = \tilde{\boldsymbol{g}}^{N'}(\boldsymbol{e}_S) - \boldsymbol{\mu}_{\boldsymbol{g}}^{N'}(\tilde{\boldsymbol{\theta}}^{N'}) = \boldsymbol{0},$$

where $\boldsymbol{\mu}_{\boldsymbol{g}}^{N'}(\boldsymbol{\theta})$ is the expected value of $\boldsymbol{g}(\boldsymbol{Y},\boldsymbol{x}_{N'})$ under (A.3).

A.3. Evaluation of uncertainty

In a network process fulfilling the heuristics of Krivitsky et al. (2011), the distribution of individual measurements $\boldsymbol{h}(\boldsymbol{e}_i)$ should not be affected by the network size: the view of each individual in the network should not, for a sufficiently large network, be affected by how large the network is. Therefore, $\boldsymbol{\Sigma}_{[w,w\boldsymbol{h}]}$ in (4.4) does not depend on $|N|$, and, if (A.2) holds, the per-capita network statistics of interest should converge.

Then, provided $\boldsymbol{\lambda}_{\boldsymbol{g}}(\boldsymbol{\theta})$ itself is differentiable,

$$\lim_{|N'|\to\infty}|N'|^{-1}\nabla_{\boldsymbol{\theta}}\boldsymbol{\mu}_{\boldsymbol{g}}(\boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}}\lim_{|N'|\to\infty}|N'|^{-1}\boldsymbol{\mu}_{\boldsymbol{g}}(\boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}}\boldsymbol{\lambda}_{\boldsymbol{g}}(\boldsymbol{\theta}).$$

Then, $|N|^{-1}\nabla_{\boldsymbol{\theta}}\boldsymbol{\mu}_{\boldsymbol{g}}\{\boldsymbol{\theta}_{\boldsymbol{g}}(\boldsymbol{\mu})\}$ in (4.4) can be approximated by $|N'|^{-1}\nabla_{\boldsymbol{\theta}}\boldsymbol{\mu}_{\boldsymbol{g}}^{N'}(\tilde{\boldsymbol{\theta}})$, estimated as in (4.5). This means that the asymptotic variance from (4.4) with N' in place of N,

$$\begin{aligned}
\lim_{|N'|\to\infty}\mathrm{var}_S(\tilde{\boldsymbol{\theta}}^{N'}) &= \lim_{|N'|\to\infty}\{\nabla_{\boldsymbol{\theta}}\boldsymbol{\mu}_{\boldsymbol{g}}^{N'}(\boldsymbol{\theta})\}^{-1}|N'|^2\boldsymbol{\Sigma}_{\mathrm{H}}/|S|[\{\nabla_{\boldsymbol{\theta}}^{N'}\boldsymbol{\mu}_{\boldsymbol{g}}(\boldsymbol{\theta})\}^{-1}]^{\top}\\
&= \lim_{|N'|\to\infty}\{|N'|^{-1}\nabla_{\boldsymbol{\theta}}\boldsymbol{\mu}_{\boldsymbol{g}}^{N'}(\boldsymbol{\theta})\}^{-1}\boldsymbol{\Sigma}_{\mathrm{H}}/|S|[\{|N'|^{-1}\nabla_{\boldsymbol{\theta}}\boldsymbol{\mu}_{\boldsymbol{g}}^{N'}(\boldsymbol{\theta})\}^{-1}]^{\top}\\
&= \{\nabla_{\boldsymbol{\theta}}\boldsymbol{\lambda}_{\boldsymbol{g}}(\boldsymbol{\theta})\}^{-1}\boldsymbol{\Sigma}_{\mathrm{H}}/|S|[\{\nabla_{\boldsymbol{\theta}}\boldsymbol{\lambda}_{\boldsymbol{g}}(\boldsymbol{\theta})\}^{-1}]^{\top},
\end{aligned}$$

so the variance of the estimator ceases to depend on $|N'|$ for $|N'|$ sufficiently large. This is a logical and welcome result: the variance of the estimator depends primarily on the sample size, not the population size.

A.4. Scaled estimation procedure

This leads to the following estimation procedure:

1. Construct a pseudopopulation N' that is a "scaled-down" N: i.e., the distribution of $\boldsymbol{x}_{N'}$ must be the same as $\boldsymbol{x}_N$.
2. Estimate the scaled sufficient statistic of the ERGM with $\tilde{\boldsymbol{g}}(\boldsymbol{e}_S) \times |N'|/|N|$.
3. Obtain $\tilde{\boldsymbol{\theta}}$, using MCMLE to solve $\widetilde{\mathrm{sc}}^{N'}(\tilde{\boldsymbol{\theta}}) = \mathbf{0}$.
4. As a byproduct of Step 3, obtain $\widetilde{\mathrm{var}}_{\boldsymbol{g}}\{\boldsymbol{g}(\boldsymbol{Y}, \boldsymbol{x}_{N'}); \tilde{\boldsymbol{\theta}}\}$.
5. Estimate $\tilde{\boldsymbol{\Sigma}}_{w\boldsymbol{h}/w.}$ as described in Section 4.2.
6. Estimate $\mathrm{var}_S(\tilde{\boldsymbol{\theta}})$ with (4.5), using N' in place of N and $\boldsymbol{x}_{N'}$ in place of $\boldsymbol{x}$.
7. Simulate from the model fit for a variety of $|N'|$ to test property (A.2).

Appendix B: Simulation study details and results

In this appendix, we give more details on the simulation study and the results.

B.1. Study design

B.1.1. Simulated population network

The population network $\boldsymbol{y}$ of size $|N| = 100{,}000$ was constructed to have the following distribution of actor attributes:

$x_{i,1}$ categorical attribute with the following composition: "A" (25%), "B" (50%), and "C" (25%); and

$x_{i,2}$ quantitative attribute, generated from $\mathrm{N}(0, 1)$ distribution.

Simulated annealing was used to find a configuration of ties such that the network statistics of interest—listed in Table 3—were as close as possible to their target values, and the difference between the generated network's statistics and the target values are shown in the same table. This network serves as the population network $\boldsymbol{y}$ in this study. (In each case, the difference is negligible, compared to the magnitude of the statistic.) An ERGM was fit to the resulting network, producing $\boldsymbol{\theta}$.

B.1.2. Sampling design

We considered two sample sizes, both taken without replacement: $|S| = 1{,}000$, for a sampling fraction of 1%, and $|S| = 2{,}000$, for a sampling fraction

TABLE 3
Population network features

Feature	$\boldsymbol{g}(\boldsymbol{y})$	Target	Deviation[1]	$\boldsymbol{\theta}$
Total ties	$\lvert\boldsymbol{y}\rvert$	$\frac{3}{4}\lvert N\rvert = 75$k	-1.00	-10.394
Isolate count	$\sum_{i\in N} 1_{\boldsymbol{y}_i=\emptyset}$	$\frac{1}{5}\lvert N\rvert = 20$k	-1.00	1.180
Degree 1 count	$\sum_{i\in N} 1_{\lvert\boldsymbol{y}_i\rvert=1}$	$\frac{1}{2}\lvert N\rvert = 50$k	-1.00	1.555
Ties on B actors	$\sum_{(i,j)\in\boldsymbol{y}}(1_{x_{i,l}=\mathrm{B}} + 1_{x_{j,l}=\mathrm{B}})$	$1\lvert N\rvert = 100$k	0.00	0.246
Ties on C actors	$\sum_{(i,j)\in\boldsymbol{y}}(1_{x_{i,l}=\mathrm{C}} + 1_{x_{j,l}=\mathrm{C}})$	$\frac{1}{4}\lvert N\rvert = 25$k	0.00	0.000
Within-group ties	$\sum_{(i,j)\in\boldsymbol{y}} 1_{x_{i,l}=x_{j,l}}$	$\frac{1}{2}\lvert N\rvert = 50$k	0.00	1.004
Difference in $\boldsymbol{x}_{\cdot,2}$	$\sum_{(i,j)\in\boldsymbol{y}}\lvert x_{i,2} - x_{j,2}\rvert$	$\frac{1}{2}\lvert N\rvert = 50$k	$+0.06$	-0.916

[1] — Here, "Deviation" refers to the difference between the statistic of the network generated and the target value.

of 2%; and we considered two sampling designs: a simple random sample and a design with sampling weights that mimic sources of sampling weights that arise in applications, including oversampling of smaller subpopulations and a response rate that varied with the continuous covariate.

For each of the four combinations of sample size and weighting scheme, we drew 2,000 egocentric samples. For each of these 8,000 samples, we used the scaled estimation procedure described in Section A.4 with $|N'| \approx 1 \times |S| \approx$ 1,000 and 2,000, $5 \times |S| \approx$ 5,000 and 10,000, and $10 \times |S| \approx$ 10,000 and 20,000 to estimate $\boldsymbol{\theta}$ and evaluate uncertainty.

B.1.3. Sampling weights

We considered two sources of unequal sampling probabilities:

- Small subpopulations can be oversampled to facilitate separate inference about them. In our simulation, we reproduce this scenario by oversampling A and C actors by a factor of 2.
- Sampling weights are also used to control for nonresponse. We emulate this by setting the response rate of actor i to be proportional to $e^{x_{i,2}}$, though for the sake of simplicity, we assume that actors are drawn from the population until the target sample size is reached.

This leads to the following sampling probabilities

$$\pi_i \propto \exp\{1_{x_{i,1}\in\{\mathrm{A,C}\}} \log(2) + x_{i,2}\}, \quad i \in N$$

and $w_i \propto 1/\pi_i$.

B.2. Results

We summarize the biases in the point estimate for Figure 3a and compare the standard deviation of the sampling distributions of the parameter estimates to the standard errors produced by the procedure in Figure 3b. Deviations from nominal coverage are visualized in Figure 4. Numerical summaries can be found in Appendix B.3.

In the following section, we focus on the raw observations and patterns, and we discuss likely reasons and sources of bias in Section 8 in the main paper.

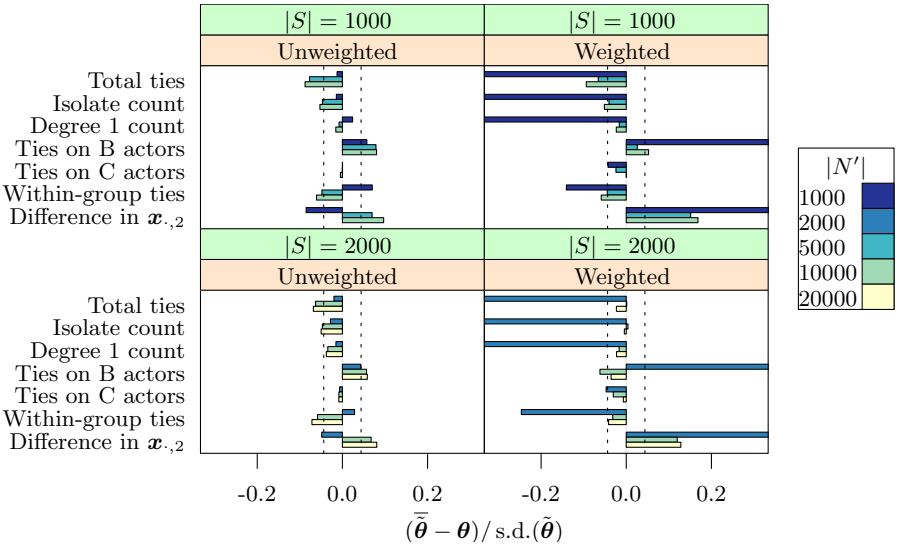
The unweighted sampling estimates display some bias, though it does not appear to have a systematic pattern as a function of $|N'|$ or in the model term. For $|S| = 1{,}000$, none of the estimated biases are greater than 10% of the standard deviation under repeated sampling, which is to say that bias accounts for less than 1% of the mean squared error (MSE) of the estimator. For $|S| = 2{,}000$ they are even smaller relative to their standard deviation (which is, itself, about $\sqrt{2}$ times smaller).

The weighted sampling estimators are, as one would expect, highly biased for smaller $|N'|$.For the largest $|N'|$, the bias of the most biased parameter estimate (Difference in $\boldsymbol{x}_{\cdot,2}$) is less than 20% of the standard deviation under repeated sampling (i.e., about 4% of the total MSE), even for $|S| = 1{,}000$. A possible reason why this particular estimate is the most biased is that egos with small $\boldsymbol{x}_{i,2}$ are (by design) severely undersampled, which means that there will exist many samples where the full range of $\boldsymbol{x}_{\cdot,2}$ is not represented. This is likely to be less problematic in real-world applications like the analysis in Section 7, where continuous covariates (like age) have an explicit range of interest. As expected, estimators under $|S| = 2{,}000$ exhibit uniformly smaller bias, even as a fraction of the smaller standard deviation.

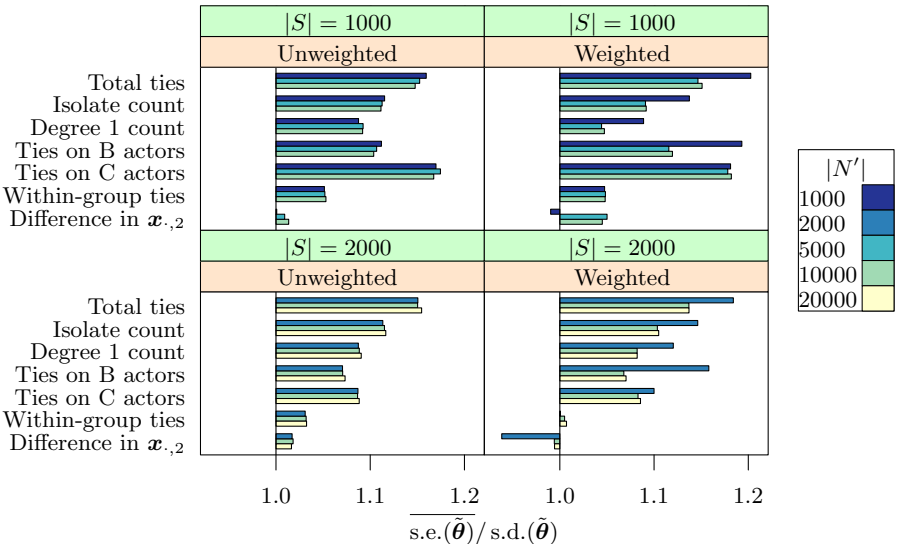
Overall, the standard errors under unweighted sampling appear to be conservative, overestimating the simulated standard deviation under repeated sampling by between 1% and 20% in some cases, and there is evidence of them becoming more accurate as the sample size increases.

This positive bias in standard errors may be a consequence of estimating the distribution of $\boldsymbol{x}_N$ from $\boldsymbol{x}_S$: somewhat counterintuitively, it may reduce the actual variance of $\tilde{\boldsymbol{\theta}}$ in the presence of homophily, because those samples that happen to contain, say, an excess of members of Group B will also contain an excess of ties incident on members of Group B, which is consistent with ERGM behavior for changing composition (Krivitsky et al., 2011).

The resulting Wald confidence interval coverage, summarized in Figure 4, is consistent with the above observations: in almost all terms, the inter-



(a) Bias in point estimates



(b) Bias in standard errors

Fig 3: Simulated bias in the point estimates and standard errors: the point estimates, normalized by s.d.($\tilde{\boldsymbol{\theta}}$). Dashed lines are positioned at $\pm 1.96/\sqrt{2000}$, so about 95% of the simulated biases should fall into these intervals if their true mean is 0. Some of the biases are truncated to preserve detail.

vals are somewhat conservative for both unweighted and weighted sampling (given sufficient $|N'|$), likely a consequence of the variance being overestimated. The coverage does appear to improve with the sample size.

We also find that for $|S| = 1{,}000$, while most parameter estimates' sampling distributions were statistically indistinguishable from normal (based on 2,000 simulated realizations each), the parameters corresponding to total number of ties and number of ties incident on actors in Group B show slight deviations from normality in both unweighted and weighted simulations. (Shapiro–Wilk P-val. < 0.01 for each.) The former term's parameter estimates exhibit negative skewness while the latter term's exhibit positive skewness. This may be because their corresponding statistics are fairly strongly negatively correlated with each other (because, with B being the largest group, and there being positive within-group homophily, 82% of the ties in $\boldsymbol{y}$ involve an actor in Group B). This strong correlation may be slowing down the rate at which their joint distribution asymptotes, further exacerbated by actors in Group B being undersampled. For $|S| = 2{,}000$, none are significantly non-normal.

These results are encouraging, in that even with a fairly moderate sample size, and in the presence of fairly heavy weighting, the confidence intervals are reasonable, provided $|N'|$ is sufficiently large. In particular, additional error due to the distribution of $\boldsymbol{x}_N$ being inferred from the sample does not appear to invalidate them.

B.3. Simulation study summary tables

The following tables give numerical summaries of the simulation studies. $\overline{\tilde{\boldsymbol{\theta}}} - \boldsymbol{\theta}$ is the bias of the point estimates and s.d.$(\tilde{\boldsymbol{\theta}})$ is their simulated standard deviation, both obtained based on 2,000 replications of egocentric sampling and estimation; and $\overline{\text{s.e.}(\tilde{\boldsymbol{\theta}})}$ is the mean of the standard errors calculated from (4.5) for each replication. Coverages for 90%, 95%, and 99% Wald confidence intervals are also given.

B.3.1. $|S| = 1{,}000$, $|N'| = 1{,}000$

Summaries:

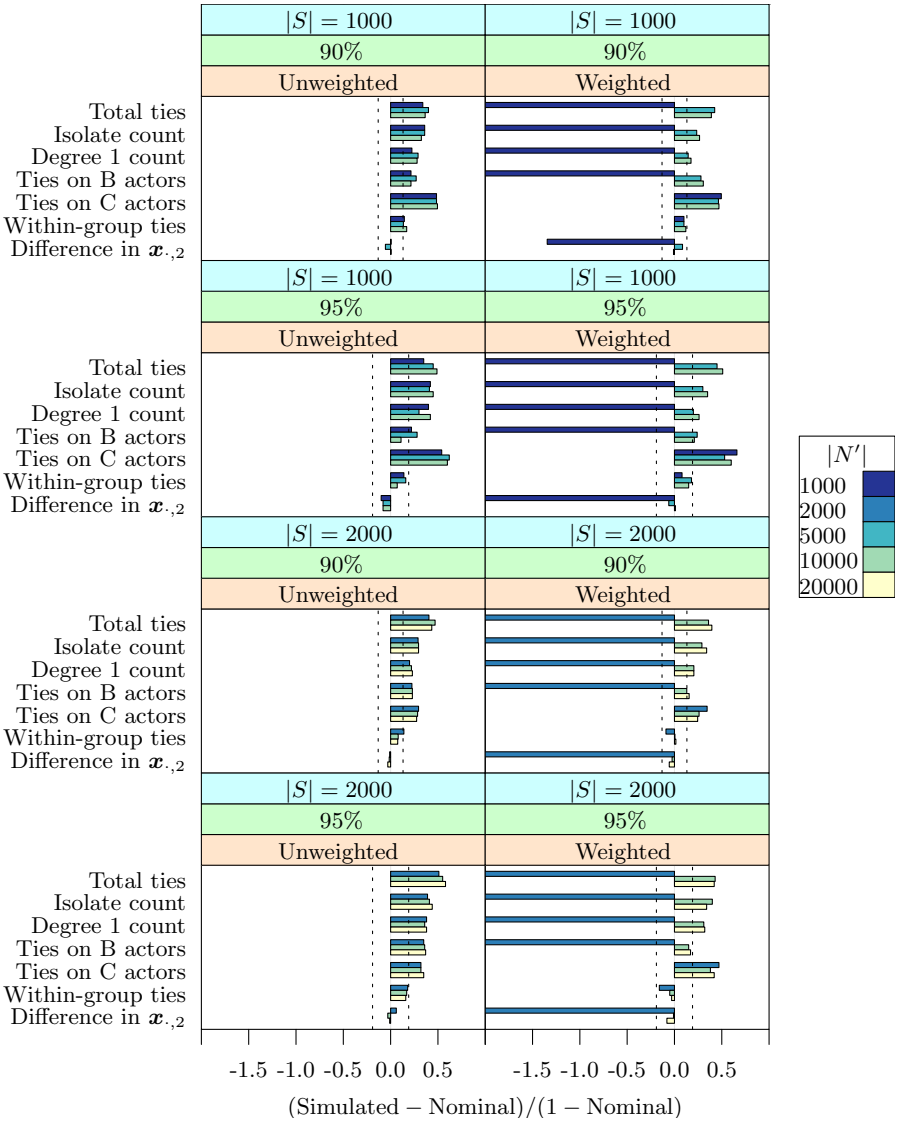


Fig 4: Coverage simulation results: the difference between the simulated and the nominal coverage is given, relative to the nominal probability of a miss. (Note that this quantity cannot be greater than 1.) Dashed lines are positioned at $\pm 1.96\sqrt{\mathrm{CL}(1-\mathrm{CL})/2000}/(1-\mathrm{CL})$, so about 95% of the simulated coverages should fall into these intervals if they do, on average, equal to nominal. Some of the coverages are truncated to preserve detail.

		Unweighted			Weighted		
	$\boldsymbol{\theta}$	$\bar{\tilde{\boldsymbol{\theta}}} - \boldsymbol{\theta}$	s.d.($\tilde{\boldsymbol{\theta}}$)	s.e.($\bar{\tilde{\boldsymbol{\theta}}}$)	$\bar{\tilde{\boldsymbol{\theta}}} - \boldsymbol{\theta}$	s.d.($\tilde{\boldsymbol{\theta}}$)	s.e.($\bar{\tilde{\boldsymbol{\theta}}}$)
Total ties	−10.394	−0.002	0.168	0.195	−0.789	0.273	0.328
Isolate count	1.180	−0.003	0.183	0.204	−0.569	0.280	0.319
Degree 1 count	1.555	0.003	0.118	0.129	−0.258	0.156	0.170
Ties on B actors	0.246	0.004	0.077	0.085	0.398	0.117	0.140
Ties on C actors	0.000	0.000	0.082	0.096	−0.004	0.101	0.119
Within-group ties	1.004	0.004	0.063	0.066	−0.009	0.065	0.069
Difference in $\boldsymbol{x}_{\cdot,2}$	−0.916	−0.004	0.051	0.051	0.052	0.060	0.059

Coverage:

	Unweighted			Weighted		
	90	95	99	90	95	99
Total ties	93.4	96.8	99.6	10.1	22.0	63.6
Isolate count	93.6	97.1	99.6	43.1	60.0	87.2
Degree 1 count	92.2	97.0	99.4	55.3	68.0	88.6
Ties on B actors	92.2	96.1	99.1	1.7	5.7	29.8
Ties on C actors	94.8	97.7	99.7	95.0	98.3	99.8
Within-group ties	91.5	95.7	99.2	91.0	95.4	99.2
Difference in $\boldsymbol{x}_{\cdot,2}$	90.0	94.5	98.9	76.5	84.7	94.3

B.3.2. $|S|$ = 1,000, $|N'|$ = 5,000

Summaries:

		Unweighted			Weighted		
	$\boldsymbol{\theta}$	$\bar{\tilde{\boldsymbol{\theta}}} - \boldsymbol{\theta}$	s.d.($\tilde{\boldsymbol{\theta}}$)	s.e.($\bar{\tilde{\boldsymbol{\theta}}}$)	$\bar{\tilde{\boldsymbol{\theta}}} - \boldsymbol{\theta}$	s.d.($\tilde{\boldsymbol{\theta}}$)	s.e.($\bar{\tilde{\boldsymbol{\theta}}}$)
Total ties	−10.394	−0.013	0.169	0.195	−0.012	0.183	0.210
Isolate count	1.180	−0.009	0.183	0.204	−0.009	0.215	0.234
Degree 1 count	1.555	−0.001	0.118	0.129	−0.002	0.143	0.149
Ties on B actors	0.246	0.006	0.077	0.085	0.002	0.076	0.085
Ties on C actors	0.000	0.000	0.082	0.096	−0.002	0.076	0.090
Within-group ties	1.004	−0.003	0.063	0.066	−0.003	0.065	0.068
Difference in $\boldsymbol{x}_{\cdot,2}$	−0.916	0.003	0.050	0.050	0.008	0.056	0.059

Coverage:

	Unweighted			Weighted		
	90	95	99	90	95	99
Total ties	94.0	97.2	99.5	94.2	97.2	99.2
Isolate count	93.6	97.0	99.6	92.3	96.5	99.3
Degree 1 count	92.9	96.5	99.4	91.5	96.0	99.2
Ties on B actors	92.7	96.4	99.3	92.8	96.2	98.9
Ties on C actors	94.8	98.1	99.8	94.7	97.7	99.6
Within-group ties	91.3	95.8	99.2	91.0	95.9	99.4
Difference in $\boldsymbol{x}_{\cdot,2}$	89.5	94.6	98.8	90.8	94.7	99.2

B.3.3. $|S|$ = 1,000, $|N'|$ = 10,000

Summaries:

		Unweighted			Weighted		
	$\boldsymbol{\theta}$	$\bar{\tilde{\boldsymbol{\theta}}} - \boldsymbol{\theta}$	s.d.($\tilde{\boldsymbol{\theta}}$)	s.e.($\tilde{\boldsymbol{\theta}}$)	$\bar{\tilde{\boldsymbol{\theta}}} - \boldsymbol{\theta}$	s.d.($\tilde{\boldsymbol{\theta}}$)	s.e.($\tilde{\boldsymbol{\theta}}$)
Total ties	−10.394	−0.015	0.169	0.194	−0.017	0.183	0.210
Isolate count	1.180	−0.010	0.183	0.204	−0.011	0.215	0.235
Degree 1 count	1.555	−0.002	0.118	0.129	−0.003	0.143	0.149
Ties on B actors	0.246	0.006	0.077	0.084	0.004	0.076	0.085
Ties on C actors	0.000	0.000	0.082	0.096	0.000	0.076	0.090
Within-group ties	1.004	−0.004	0.063	0.066	−0.004	0.065	0.068
Difference in $\boldsymbol{x}_{\cdot,2}$	−0.916	0.005	0.050	0.050	0.009	0.056	0.058

Coverage:

	Unweighted			Weighted		
	90	95	99	90	95	99
Total ties	93.7	97.5	99.5	93.9	97.5	99.2
Isolate count	93.2	97.2	99.5	92.7	96.8	99.4
Degree 1 count	92.8	97.1	99.4	91.8	96.3	99.3
Ties on B actors	92.2	95.5	99.2	93.0	96.0	99.1
Ties on C actors	95.0	98.0	99.8	94.7	98.0	99.8
Within-group ties	91.7	95.3	99.2	91.2	95.8	99.2
Difference in $\boldsymbol{x}_{\cdot,2}$	90.0	94.6	98.6	89.9	95.0	99.1

B.3.4. $|S|$ = 2,000, $|N'|$ = 2,000

Summaries:

		Unweighted			Weighted		
	$\boldsymbol{\theta}$	$\bar{\tilde{\boldsymbol{\theta}}} - \boldsymbol{\theta}$	s.d.($\tilde{\boldsymbol{\theta}}$)	s.e.($\bar{\tilde{\boldsymbol{\theta}}}$)	$\bar{\tilde{\boldsymbol{\theta}}} - \boldsymbol{\theta}$	s.d.($\tilde{\boldsymbol{\theta}}$)	s.e.($\bar{\tilde{\boldsymbol{\theta}}}$)
Total ties	−10.394	−0.002	0.116	0.133	−0.773	0.188	0.223
Isolate count	1.180	−0.004	0.126	0.141	−0.555	0.190	0.218
Degree 1 count	1.555	−0.001	0.082	0.089	−0.258	0.104	0.117
Ties on B actors	0.246	0.002	0.054	0.057	0.390	0.082	0.095
Ties on C actors	0.000	0.000	0.060	0.065	−0.003	0.073	0.081
Within-group ties	1.004	0.001	0.043	0.045	−0.011	0.046	0.046
Difference in $\boldsymbol{x}_{\cdot,2}$	−0.916	−0.002	0.034	0.034	0.052	0.043	0.040

Coverage:

	Unweighted			Weighted		
	90	95	99	90	95	99
Total ties	94.0	97.5	99.6	0.2	0.8	6.6
Isolate count	92.9	97.0	99.5	11.3	22.3	52.8
Degree 1 count	92.0	96.9	99.6	26.3	39.0	66.3
Ties on B actors	92.2	96.8	99.3	0.0	0.0	0.4
Ties on C actors	93.0	96.6	99.4	93.5	97.4	99.7
Within-group ties	91.4	95.9	99.4	89.1	94.2	99.0
Difference in $\boldsymbol{x}_{\cdot,2}$	89.9	95.3	99.2	61.8	73.1	87.8

B.3.5. $|S|$ = 2,000, $|N'|$ = 10,000

Summaries:

		Unweighted			Weighted		
	$\boldsymbol{\theta}$	$\bar{\tilde{\boldsymbol{\theta}}} - \boldsymbol{\theta}$	s.d.($\tilde{\boldsymbol{\theta}}$)	s.e.($\bar{\tilde{\boldsymbol{\theta}}}$)	$\bar{\tilde{\boldsymbol{\theta}}} - \boldsymbol{\theta}$	s.d.($\tilde{\boldsymbol{\theta}}$)	s.e.($\bar{\tilde{\boldsymbol{\theta}}}$)
Total ties	−10.394	−0.007	0.116	0.133	0.000	0.126	0.143
Isolate count	1.180	−0.006	0.126	0.141	0.001	0.146	0.161
Degree 1 count	1.555	−0.003	0.082	0.089	−0.002	0.095	0.103
Ties on B actors	0.246	0.003	0.054	0.057	−0.003	0.054	0.057
Ties on C actors	0.000	0.000	0.060	0.065	−0.002	0.056	0.060
Within-group ties	1.004	−0.003	0.043	0.045	−0.001	0.046	0.046
Difference in $\boldsymbol{x}_{\cdot,2}$	−0.916	0.002	0.034	0.034	0.005	0.040	0.040

Coverage:

	Unweighted			Weighted		
	90	95	99	90	95	99
Total ties	94.7	97.8	99.6	93.6	97.2	99.4
Isolate count	93.0	97.0	99.6	92.9	97.0	99.5
Degree 1 count	92.2	96.8	99.6	92.0	96.5	99.3
Ties on B actors	92.3	96.8	99.4	91.3	95.8	99.1
Ties on C actors	92.8	96.6	99.5	92.6	96.9	99.6
Within-group ties	90.8	95.9	99.4	90.0	94.8	99.0
Difference in $\boldsymbol{x}_{\cdot,2}$	89.8	94.8	99.2	89.8	95.0	99.1

B.3.6. $|S| = 2{,}000$, $|N'| = 20{,}000$

Summaries:

		Unweighted			Weighted		
	$\boldsymbol{\theta}$	$\bar{\tilde{\boldsymbol{\theta}}} - \boldsymbol{\theta}$	s.d.($\tilde{\boldsymbol{\theta}}$)	s.e.($\bar{\tilde{\boldsymbol{\theta}}}$)	$\bar{\tilde{\boldsymbol{\theta}}} - \boldsymbol{\theta}$	s.d.($\tilde{\boldsymbol{\theta}}$)	s.e.($\bar{\tilde{\boldsymbol{\theta}}}$)
Total ties	−10.394	−0.008	0.115	0.133	−0.003	0.126	0.143
Isolate count	1.180	−0.006	0.126	0.141	−0.001	0.146	0.161
Degree 1 count	1.555	−0.003	0.082	0.089	−0.002	0.095	0.103
Ties on B actors	0.246	0.003	0.054	0.058	−0.002	0.053	0.057
Ties on C actors	0.000	−0.001	0.060	0.065	0.000	0.056	0.060
Within-group ties	1.004	−0.003	0.043	0.045	−0.002	0.046	0.046
Difference in $\boldsymbol{x}_{\cdot,2}$	−0.916	0.003	0.034	0.034	0.005	0.040	0.040

Coverage:

	Unweighted			Weighted		
	90	95	99	90	95	99
Total ties	94.3	97.9	99.6	94.0	97.1	99.4
Isolate count	93.0	97.2	99.7	93.4	96.7	99.4
Degree 1 count	92.3	96.9	99.6	92.0	96.6	99.3
Ties on B actors	92.3	96.9	99.2	91.5	95.9	99.1
Ties on C actors	92.8	96.8	99.4	92.5	97.1	99.6
Within-group ties	90.8	95.8	99.5	90.1	94.8	99.0
Difference in $\boldsymbol{x}_{\cdot,2}$	89.7	95.0	99.1	89.5	94.6	98.9

Appendix C: Auxiliary results for the application

In this appendix, we report auxiliary results to our analysis in Section 7. In particular, we discuss the heuristic for our choice of $|N'|$, confirm that our results remain after controlling for age, and verify assumption (A.2) as it pertains to the models fit.

C.1. Selecting $|N'|$ for the analysis

Our choice of $|N'|$ is driven by the data including sampling weights: inference requires that the N' be as representative of N as possible, and, as we show in the Appendix B, insufficient $|N'|$ can significantly bias estimation. We cannot compare the two situations directly, but, heuristically, the weights are somewhat more varied in the NHSLS data than in the weighted simulation study: in the simulation study's samples of 2,000, $w_{\max}/w_{\min}$ averaged 9.1 for $|S| = 1{,}000$ and 9.9 for $|S| = 2{,}000$, and in the NHSLS study (after excluding respondents with missing data), this ratio is somewhat higher, 15.3 (albeit at a greater sample size). Another metric is the amount by which the variance of the sample mean would be inflated due to unequal weighting, relative to an SRS, which equals to $|N|^{-1}\sum_{i=1}^{|N|} w_i^2/\bar{w}^2$. This is 1.18 for the simulation study and 1.34 for the NHSLS data. $|N'| \approx 5|S|$ appear to be adequate for the simulated data, though $|N'| \approx 10|S|$ produces a noticeable improvement, so we select, conservatively, $|N'| \approx 45{,}000 \approx 13.4 \times |S|$. The respondent with the smallest sampling weight represents about $w_i/w_\cdot = 7.28 \times 10^{-5}$ of the population, so she is represented in N' about three times.

C.2. Controlling for age

In this section, we report a model fit that, in addition to representing mixing by race and monogamy, also incorporates age effects.

In this, we follow the analysis of these data by Krivitsky et al. (2011), modeling the effects of age semiparametrically. As predictors, we consider the age of the actor, the square root of age, the age difference and squared difference in a potential partnership, and the difference and the squared difference of the square roots of ages. To improve numeric conditioning of the model, we perform an affine transformation on the ages, shifting and scaling them into a $[-1/2, +1/2]$ interval: $x'_{i,\text{age}} = (x_{i,\text{age}} - 18)/(60 - 18) - 1/2$. This change merely scales the coefficient and changes the baseline coefficients (number of ties, by sex), without changing the family of distributions being modeled. For the square root of age effects, the corresponding transformation is

$$x'_{i,\sqrt{\text{age}}} = \sqrt{\frac{x_{i,\text{age}} - 18}{60 - 18}} - \frac{1}{2}.$$

The use of the square root and linear effect, rather than linear and quadratic, is motivated by the notion that the effect of a one-year difference will be greater for younger actors than older: going from 20 to 21 is likely to have a greater effect than going from 50 to 51.

TABLE 4

Coefficients and significance for the Model 3 *(main, mixing, and monogamy effects) and a model that also incorporates age effects. Coefficients reported are in the presence of an edge count offset of* $-\log(44859) = -10.71$.

	Main + Mix. + Monog.	+ Age
Actor activity by sex		
Female	$-1.88\ (0.31)^{***}$	$-1.78\ (0.40)^{***}$
Male	$-1.18\ (0.25)^{***}$	$-1.08\ (0.35)^{**}$
Same-sex partnership	$-4.52\ (0.21)^{***}$	$-4.13\ (0.21)^{***}$
Actor activity by race		
White	0 (baseline)	
Black	$-0.30\ (0.38)$	$-0.35\ (0.36)$
Other	$0.93\ (0.42)^{*}$	$0.87\ (0.43)^{*}$
Race homophily by race		
Black	$5.15\ (0.38)^{***}$	$5.16\ (0.38)^{***}$
Other	$2.04\ (0.35)^{***}$	$2.09\ (0.39)^{***}$
White	$2.32\ (0.36)^{***}$	$2.31\ (0.39)^{***}$
Monogamy by sex and race		
Black female	$1.80\ (0.47)^{***}$	$1.94\ (0.50)^{***}$
Other female	$2.51\ (0.67)^{***}$	$2.56\ (0.69)^{***}$
White female	$2.25\ (0.31)^{***}$	$2.36\ (0.32)^{***}$
Black male	$0.99\ (0.24)^{***}$	$1.11\ (0.28)^{***}$
Other male	$1.40\ (0.31)^{***}$	$1.54\ (0.33)^{***}$
White male	$2.16\ (0.25)^{***}$	$2.30\ (0.25)^{***}$
Age effects		
$\sqrt{\text{age}}$ effect		$3.29\ (1.35)^{*}$
age effect		$-2.73\ (1.15)^{*}$
Age difference effects		
Difference in $\sqrt{\text{age}}$		$-8.07\ (2.20)^{***}$
Difference in age		$-6.03\ (1.92)^{**}$
Squared difference in $\sqrt{\text{age}}$		$3.22\ (3.73)$
Squared difference in age		$2.32\ (2.80)$
Older-male-younger-female		$0.93\ (0.05)^{***}$

Significance levels: $0.05 \geq {}^{*} > 0.01 \geq {}^{**} > 0.001 \geq {}^{***}$

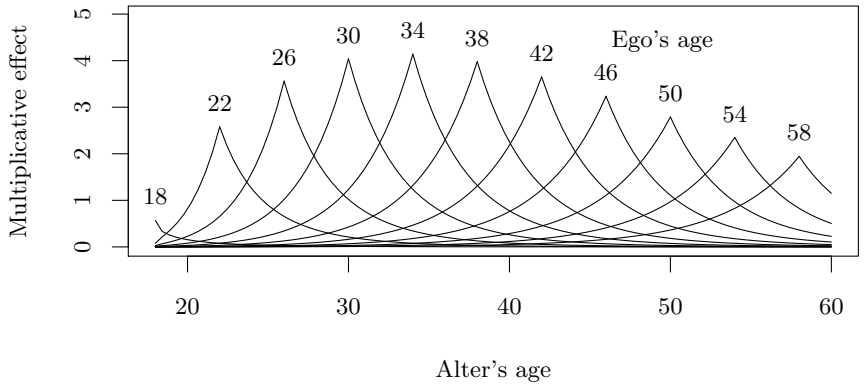


Fig 5: Estimated multiplicative effects of the age of ego and age of alter (ignoring age–sex interaction) on the odds of a tie. Note that the peak of each ego age curve represents the overall propensity for egos of that age to have ties.

The results for *Model 3* with age, along with the estimates for *Model 3* itself for comparison, are given in Table 4. The most important aspect of this result is that the coefficients estimated for the Monogamy model have not changed qualitatively after age effects are controlled for: our results in Section 7 are robust to age effects. (We also performed the degree distribution and component size simulations for the age model. Those were not qualitatively different either.)

Age effects themselves can be interpreted as well. We provide a visualization of the overall estimated age effect in Figure 5. Age difference effects are particularly significant. There is also strong evidence for the tendency of the male to be older than the female in a heterosexual partnership.

C.3. Simulation results to verify Assumption (A.2)

We test the assumption (A.2) by simulating from the most complex model fit to obtain 10,000 realizations (with some serial dependence) of $\boldsymbol{g}(\boldsymbol{Y}, \boldsymbol{x}_{N''})/|N''|$ with $|N''| \approx 90{,}000$ with offsets adjusted appropriately. If the assumption is violated, we would expect them to be different, on average, from the observed $\bar{\boldsymbol{g}}(\boldsymbol{e}_S)$.

We report the simulation results in Table 5. The differences between the observed values and those simulated for $|N'| \approx 90,000$ are statistically significant in a few cases—as they would inevitably be, given a sufficient simulation size, but they are not practically so: the statistics with the greatest

TABLE 5

Difference between the observed per-capita statistics (denoted $\bar{\boldsymbol{g}}(\boldsymbol{e}_S)$) and the per-capita moments of the sufficient statistics simulated from a network with $|N''| \approx 90{,}000$ using coefficients obtained with $|N'| \approx 45{,}000$ (denoted $\boldsymbol{\mu}_{\boldsymbol{g}}^{N''}(\tilde{\boldsymbol{\theta}}^{N'})/|N''|$). The differences have been scaled by 10^4 for readability, and the simulation's standard errors are adjusted for autocorrelation. Effective Sample Sizes (ESS) are also given. R (R Core Team, 2013) package `coda` *(Plummer et al., 2006) was used to evaluate the latter.*

Term	Observed $\bar{\boldsymbol{g}}(\boldsymbol{e}_S)$	Simulated $\left\{\frac{\boldsymbol{\mu}_{\boldsymbol{g}}^{N''}(\tilde{\boldsymbol{\theta}}^{N'})}{\lvert N''\rvert} - \bar{\boldsymbol{g}}(\boldsymbol{e}_S)\right\} \times 10^4$ (ESS, s.e.)	$\frac{\text{Diff.}}{\bar{\boldsymbol{g}}(\boldsymbol{e}_S)}$
Actor activity by sex			
Female	0.396	0.085 (1366, 0.229)	0.002%
Male	0.399	0.033 (1429, 0.228)	0.001%
Same-sex partnership	0.005	−0.056 (4336, 0.034)	−0.120%
Actor activity by race (White as baseline)			
Black	0.087	−0.287 (564, 0.306)	−0.033%
Other	0.102	0.242 (777, 0.231)	0.024%
Race homophily by race			
Black	0.040	−0.113 (508, 0.166)	−0.028%
Other	0.038	0.136 (411, 0.184)	0.036%
White	0.288	0.025 (1421, 0.186)	0.001%
Monogamy by sex and race			
Black Female	0.042	−0.148 (735, 0.136)	−0.035%
Other Female	0.052	0.046 (1025, 0.107)	0.009%
White Female	0.284	0.358 (1749, 0.177)*	0.013%
Black Male	0.031	−0.268 (776, 0.132)*	−0.086%
Other Male	0.041	−0.189 (971, 0.122)	−0.046%
White Male	0.290	−0.053 (1979, 0.168)	−0.002%
Age effects			
$\sqrt{\text{age}}$ effect	0.104	0.142 (1285, 0.120)	†
age effect	−0.046	0.069 (1294, 0.125)	†
Age difference effects			
Difference in $\sqrt{\text{age}}$	0.028	0.130 (1135, 0.049)**	0.046%
Difference in age	0.034	0.113 (1114, 0.058)	0.034%
Squared difference in $\sqrt{\text{age}}$	0.004	0.023 (1850, 0.013)	0.055%
Squared difference in age	0.006	0.011 (1787, 0.017)	0.019%
Older-male-younger-female	0.242	0.713 (452, 0.504)	0.029%

Significance levels: $0.05 \geq {}^\star > 0.01 \geq {}^{\star\star} > 0.001 \geq {}^{\star\star\star}$

† — Percent differences are not meaningful for statistics that are not counts or sums of nonnegative quantities.

relative difference between $|N'| \approx 45,000$ and $|N'| \approx 90,000$ are ones with the smallest counts and effective numbers of observations, so one might expect them to asymptote more slowly; even among them, the greatest one has 0.120% difference.