Title

Estimating contact network properties by integrating multiple data sources associated with infectious diseases

Journal

Statistics in Medicine, 42(20)

ISSN

0277-6715

Authors

Goyal, Ravi

Carnegie, Nicole

Slipher, Sally

et al.

Publication Date

2023-09-10

DOI

10.1002/sim.9816

Peer reviewed

HHS Public Access

Author manuscript

Stat Med. Author manuscript; available in PMC 2024 September 10.

Published in final edited form as:

Stat Med. 2023 September 10; 42(20): 3593–3615. doi:10.1002/sim.9816.

Estimating contact network properties by integrating multiple data sources associated with infectious diseases

Ravi Goyal,
Division of Infectious Diseases and Global Public, University of California San Diego

Nicole Carnegie,
The Public Health Company

Sally Slipher,
Department of Mathematical Sciences, Montana State University

Philip Turk,
Department of Data Science, University of Mississippi Medical Center

Susan J. Little,
Division of Infectious Diseases and Global Public, University of California San Diego

Victor De Gruttola
Department of Biostatistics, Harvard T.H. Chan School of Public Health

Abstract

To effectively mitigate the spread of communicable diseases, it is necessary to understand the interactions that enable disease transmission among individuals in a population; we refer to the set of these interactions as a *contact network*. The structure of the contact network can have profound effects on both the spread of infectious diseases and the effectiveness of control programs. Therefore, understanding the contact network permits more efficient use of resources. Measuring the structure of the network, however, is a challenging problem. We present a Bayesian approach to integrate multiple data sources associated with the transmission of infectious diseases to more precisely and accurately estimate important properties of the contact network. An important aspect of the approach is the use of the congruence class models for networks. We conduct simulation studies modeling pathogens resembling SARS-CoV-2 and HIV to assess the method; subsequently, we apply our approach to HIV data from the University of California San Diego Primary Infection Resource Consortium. Based on simulation studies, we demonstrate that the integration of epidemiological and viral genetic data with risk behavior survey data can lead to large decreases in mean squared error (MSE) in contact network estimates compared to estimates based strictly on risk behavior information. This decrease in MSE is present even in settings where

r1goyal@health.ucsd.edu .

Data

Software to execute our approach is available through a public GitHub repository (https://github.com/ravigoyalgit/Bayes_Net_Inf_CCM). Simulation results are available (https://www.dropbox.com/scl/fo/j370zm5r0whg3e52m3lkz/h?dl=0&rlkey=aeoflh5s27wjdmq6lx0p5k4qh). Information to request University of California San Diego Primary Infection Resource Consortium data can be found here: https://www.pirc.ucsd.edu/submit-proposal.

the risk behavior surveys contain measurement error. Through these simulations, we also highlight certain settings where the approach does not improve MSE.

Keywords

Bayesian inference; phylodynamics; contact network; epidemic model

1 Introduction

Effective control of the spread of communicable diseases requires knowledge about the interactions that enable disease transmissions among individuals in a population; we refer to the set of these interactions as a contact network. The structure of the contact network can have profound effects on both the spread of infectious disease and the effectiveness of control programs.[45, 49, 29, 42, 65] Therefore, understanding the contact network permits more efficient use of resources by improving our ability to predict potential impacts of interventions designed to control infectious diseases. Measuring the structure of the network, however, is a challenging problem, but–as we demonstrate–integration of multiple data sources can aid in this effort. Studies investigating complex disease dynamics typically collect a broad range of data, such as risk behavior information, infection and treatment times of infected individuals (which we refer to as epidemiological data), and viral genetic sequences. For example, the University of California San Diego Primary Infection Resource Consortium (PIRC)–an observational cohort of people living with HIV–collects such information.[36] Another example is the Botswana Combination Prevention Project, which was a cluster randomized trial to compare a combined HIV prevention intervention to standard of care in Botswana.[40, 67] Risk behavior, epidemiological, and viral sequence data sets are important and complementary in understanding disease transmission dynamics. Currently, there are limited methods for integrating infectious disease data into a single model to estimate contact network properties. To address this gap, this manuscript presents a Bayesian approach to integrate multiple data sources associated with transmission of infectious disease to estimate important properties of contact networks.

Network science research has identified a number of important properties for investigating disease dynamics, such as mixing patterns and clustering; among the most important properties is degree distribution.[48] Estimates of a population's degree distribution provide information on the mean number of interactions per individual. Communities with a higher mean tend to have more rapid spread of the disease. Furthermore, the degree distribution provides information on heterogeneity of interactions across the population; the amount of heterogeneity is also associated with the rate of disease spread. For example, communities with right-skewed degree distributions–those that include a small proportion of individuals with a large number of interactions–tend to have higher transmission rates.[44] Fortunately, effective strategies to reduce spread can be tailored to the precise nature of these properties. For example, a potentially effective strategy for populations with right-skewed degree distributions would be to focus provision of services to individuals with high number of connections.[51]

Integration of multiple data sources may be necessary to accurately estimate contact network properties. For example, surveys of risk behavior may not be sufficient due to the logistical and financial considerations making them challenging to administer to a large number of individuals. Furthermore, estimates derived from these surveys often have biases that arise from using non-representative samples and measurement error, in particular surveys that deal with sensitive or stigmatized behaviors, such as sexual relationships.[23, 35, 38, 64, 14] Although epidemiological and viral genetic data are less likely to be impacted by measurement error than are self-reported risk survey data, they are only relevant for inferring characteristics of contacts that result in actual disease transmissions; such characteristics can differ from those that apply to the broader contact network. As shown in Groendyke et al. [20], using epidemiological data as a single source of data can result in very imprecise estimates of network properties with confidence bands that cover most of the range of plausible values. Previous efforts to infer contact network structure for a population using only pathogen genetic data revealed that the shape of a phylogeny is dependent on the contact network structure, although that study evaluated highly idealized contact networks that differed in very significant ways from one another.[37] Frost and Volz [10] noted that contact networks may influence the grouping of subpopulations in a phylogeny. However, other studies have indicated that some properties of the underlying contact network, specifically clustering, have relatively little impact on the transmission tree in particular settings; [66, 3] therefore, information on the transmission tree only may not provide insight into the contact network.

Our proposed approach to integrate multiple data sources builds on an existing Bayesian framework that makes inferences about contact network structure solely from epidemiological data. [2, 19, 20] The most recent approaches in this framework use exponential random graph models (ERGMs) as the statistical model for the contact network. [20] The use of ERGMs provides a more flexible framework than did previous approaches by allowing the inclusion of covariate information on individuals. Furthermore, methods exist to estimate ERGM parameters from behavioral survey data. [32, 33] In order to integrate such estimates from behavioral surveys into the existing Bayesian framework, however, the model must use these estimates as prior information. We know of no available methods for doing so. Beyond the need for such methods, there are computational limitations of the multiple data source integration approach using ERGMs that restrict estimation to a limited set of network properties, specifically dyadic-independent properties. [20] This limitation precludes exploration of dyadic-dependent network properties, which include such important measures as degree distribution and the number of triangles, when integrating multiple data sources. Therefore, to appropriately integrate data and estimate network properties important for infectious disease transmission dynamics, we propose an approach that makes use of an alternative network model.

In this paper, we make use of a flexible network model, referred to as the congruence class model (CCM) for networks.[13] CCMs form a broad family of models that includes, as special cases, several common network models, such as the Erdős-Rényi (ER) model, the stochastic block (SB) model, and many ERGMs. A particular CCM is defined by (1) the set of networks properties (such as degree distribution) included in the model and (2) a

probability mass function (PMF) on the values of these network properties; additional details and illustrations are provided in Section 2.

Our proposed approach, using CCMs to integrate multiple data sources to estimate contact network properties, provides several important advantages; all but the last arise from the fact that there are minimal restrictions on the functional form of the PMF of network properties included in CCMs. First, CCMs make it possible for specification of the mean and uncertainty of network properties. For example, the PMF for network properties can reflect greater uncertainty when collecting risk behavior information from a small fraction of individuals in the population compared to collecting information from a larger fraction of the population. Second, by using the same PMF for properties estimated from the risk behavioral surveys in the network model, in our case a CCM, one can statistically compare the results that are derived from integrating the data to those based only on behavioral survey data. Third, we can investigate the impact of measurement error in the risk behavior responses on estimating network properties. Fourth, the parameterization of CCMs avoids the computationally expensive estimation procedure necessary for ERGMs to fit dyadic-dependent network properties. In the proposed approach using CCMs, one can investigate more complex network properties, such as degree distribution, which, to our knowledge, is not possible using currently available approaches. The use of CCM requires addressing additional complexities not necessary for ERGMs. This paper demonstrates these advantages as well as provides approaches to address the methodological complexities associated with CCMs.

The organization of this paper is as follows. The next section (Section 2) provides necessary background information. Section 3 presents the general approach to integrate multiple data sources for estimating the structure of the contact network. Sections 4 and 5 present simulation studies that investigate the potential of our approach to increase accuracy and precision of network property estimates by comparing estimates that integrate multiple data sources to those that rely only on behavior surveys. Sections 4 and 5 simulate spread of SARS-CoV-2 and HIV using epidemic processes appropriate for these pathogens. These simulation studies focus on degree distribution for two reasons: (1) it plays an important role in disease dynamics, and (2) modeling degree distribution is not possible using currently available approaches. Section 6 demonstrates the usefulness of our approach by application to the PIRC HIV data set. The paper concludes with a discussion (Section 7) that highlights complexities in real-world data sets and areas of further development to address them. Software to execute our approach is available through a public GitHub repository (https://github.com/ravigoyalgit/Bayes_Net_Inf_CCM).

2 Background

2.1 Epidemic model

The proposed approach requires specifying an epidemic process for the pathogen of interest. In this paper, we investigate two processes: SI^AI^LR and *SEIR*. For the SI^AI^LR epidemic model, we denote the stages as: Susceptible (S), Acute Infectious $\left(I^A\right)$, Long-term Infectious $\left(I^L\right)$, and Recovered (R). Individuals in Stage S are negative for the disease, but are

susceptible to acquiring it. If an individual acquires the disease, they move from Stage S to Stage I^A. Individuals in Stage I^A are acutely infected and are highly infectious. They transmit the infection to any of their contacts with probability β_A at each time unit. Individuals remain in Stage I^A for an exponentially distributed waiting time with scale κ_A. Then, individuals move to Stage I^L. During this phase of the infection they are still infectious, but transmit the infection to any of their contacts with probability β_L, which can differ from β_A. Individuals remain in stage I^L for an exponentially distributed waiting time with scale κ_L. Afterwards, individuals move to Stage R, which represents that the individual has recovered, been admitted to the hospital, been placed on treatment, or died. Epidemic models commonly use multiple infectiousness periods to investigate HIV,[9] including processes with two infectious periods.[46, 18, 65]

Another common epidemic process is an $SEIR$ process. The Stages S and R still refer to Susceptible and Recovered, respectively. Stages E (Exposed) and I (Infectious) represent the latent phase of the infection (pre-infectious period) and the infectious period, respectively. During Stage E, individuals are not infectious and therefore cannot transmit the pathogen to any of their contacts. The $SEIR$ process is a special case of the SI^AI^LR where $\beta_A = 0$ and is a common process for modeling influenza and SARS-CoV-2.[17] We denote the transmission probability during Stage I as β and the exponentially distributed waiting times in Stages E and I as κ_E and κ_I, respectively. For both the SI^AI^LR and $SEIR$ processes, we assume a fixed β_A and β_L across all contacts. Previous modeling studies have developed epidemic processes that varied the transmission probability for different partnership types (e.g., steady and casual).[21, 68, 5, 28] Similarly, COVID transmission models have different probabilities based on type of contact.[17] Including dyadic covariates that modify transmission probabilities would be a straight-forward extension to the current framework; for clarity of presentation we do not include dyadic covariates.

2.2 Terminology and notation

We denote a contact network as $g^c = (v_{g^c}, e_{g^c})$, where v_{g^c} and e_{g^c} are the set of individuals (vertices) and contacts (edges) within a population, respectively. Let n denote the number of individuals in v_{g^c} and let m denote the number of individuals infected during the epidemic. The transmission network induced by an epidemic, denoted as $g^p = (v_{g^p}, e_{g^p})$, is a directed network consisting of the m infected individuals. An edge from individual i to j, denoted as (i, j), is in g^p (denoted as $(i, j) \in g^p$) if and only if i infected j. As we assume transmission can only occur between two individuals with contact, so g^p is a subgraph of g^c. When discussing properties or methods that are applicable to networks in general, we use the notation g, without a superscript, to denote a network.

Let the degree of vertex i, denoted as $d_i(g)$, be the number of edges between that vertex and others. Let $d(g) = (d_1(g), \ldots, d_n(g))$ represent the vector of degrees of vertices in set v_g, commonly referred to as a degree sequence. The degree distribution, denoted $D(g)$, is a vector such that the j^{th} entry represents the number of vertices having degree $j - 1$, e.g., $D_j(g^c) = \sum_{i=1}^{n} I_{\{d_i(g^c) = j\}}$.

Let $\mathcal{G}_n$ denote the space of all networks with n vertices. Let ϕ denote an algebraic map from $\mathcal{G}_n$ to network summary statistics of interest (e.g., degree distribution or mixing patterns) and let $c_\phi(x) = \{g : \phi(g) = x, g \in \mathcal{G}_n\}$ denote the inverse image associated with ϕ. We refer to these inverse images as congruence classes;[13] they also have been referred to as fibers in algebraic statistics literature.[53] Let $|c_\phi(x)|$ denote the number of graphs for which network property ϕ equals x; this quantity has been referred to as a volume factor.[59]

In this paper, epidemiological data consist of transition times for the m individuals ultimately infected. For the SI^AI^LR epidemic process, we denote the start times of the infection, i.e., the times that individuals transition from Stage S to Stage I^A, as $T^A = (T_1^A, T_2^A, \ldots, T_m^A)$, where T_i^A represents the transition time for individual i. We denote the end time of the acute period, i.e., transitioning from Stage I^A to Stage I^L, as $T^L = (T_1^L, T_2^L, \ldots, T_m^L)$ and the time of recovery, treatment, hospitalization, or death, i.e., the times that individuals transition from Stage I^L to Stage R, as $T^R = (T_1^R, T_2^R, \ldots, T_m^R)$. For the $SEIR$ epidemic process, we use the notation of T^E and T^I for the start times of the exposed and infectious stages, respectively. We denote the collection of these transition times as T. For the SI^AI^LR epidemic process $T = \left\{T^A, T^L, T^R\right\}$, while for the $SEIR$ process $T = \left\{T^E, T^I, T^R\right\}$. Let H^t be the viral genetic sequences for all infected individuals at time t; h_i^t is the genetic sequence for individual i. Let $Dist(h_i^t, h_j^t)$ be a distance between sequences h_i^t and h_j^t; that is, a genetic distance between individuals i and j at time t.

2.3 Congruence class model

As noted above, a particular CCM is defined by (1) a network property or set of properties (such as degree distribution) and (2) a PMF on the congruence classes defined by values of the network property. We denote the PMF as $P_\phi(x \mid \theta)$, which specifies the total probability of all networks that are elements in $c_\phi(\phi(g))$ given a set of parameter values θ, i.e.,

$$P_\phi\left(x \mid \theta\right) = \sum_{g \in c_\phi(x)} P_{\mathcal{G}_n}\left(g \mid \theta\right), \tag{1}$$

where $P_{\mathcal{G}_n}(g \mid \theta)$ is the probability of graph g. CCMs assume that all networks within a congruence class have the same probability of being observed; this assumption is also made in commonly used network models including the ER, SB, and ERGMs. Therefore, the probability distribution on $\mathcal{G}_n$ for a CCM is the following:

$$P_{\mathcal{G}_n}\left(g \mid \theta\right) = \left(\frac{1}{|c_\phi(\phi(g))|}\right) P_\phi\left(\phi\left(g\right) \mid \theta\right). \tag{2}$$

The CCM does not impose any constraints on specifying the probability distribution associated with network properties included in a model, i.e., $P_\phi(x \mid \theta)$. To illustrate this

flexibility, we provide examples of CCMs in the following section. Furthermore, this flexibility enables investigators to integrate multiple data sources, which we describe in Section 3 and demonstrate in Sections 4–6.

2.4 Illustrations of Congruence Class Models

We present two sets of examples to illustrate CCMs. In the first set, the network property of interest is the number of edges; in the the second, it is degree distribution. For both illustrations, we investigate networks with $n = 4$ vertices. Investigating networks of small size allows for complete enumeration of networks ($|\mathcal{G}_4| = 64$), stratified by values associated with the two network properties of interest (i.e., number of edges and degree distribution).

2.4.1 Examples of CCMs: number of edges—Let ϕ_1 denote an algebraic map from a network g to its number of edges. Networks for size $n = 4$ have between 0 and 6 edges. Therefore, $\mathcal{G}_4$ contains 7 distinct congruence classes associated with the number of edges; each of the 64 possible networks resides in exactly one of these classes. We denote the congruence class that contains all of the networks with x edges as $c_{\phi_1}(x)$. We can calculate the size of each congruence class defined by number of edges using the following formula:[22]

$$|c_{\phi_1}(x)| = \left| \binom{\binom{n}{2}}{x} \right|; \tag{3}$$

these calculates are shown in Table 1.

Table 1 shows three different possible ways to assign probabilities to the congruence classes, i.e., $P_{\phi_1}(x \mid \theta)$. Each of these assignments results in the mean number of edges being equal to 3. In the first probability distribution, $P_{\phi_1}(x \mid \theta)$ follows the uniform distribution, i.e., a network has equal probability of being drawn from congruence class c_i as congruence class c_j (c_i and c_j are arbitrary classes). In the second, the probability distribution follows a binomial distribution with $\theta = 0.5$, i.e., $P_{\phi_1}(x \mid \theta) = \theta^x \cdot (1-\theta)^{6-x}$. In the third, it is a bi-modal distribution. For each of these three distributions, we calculate the probability of drawing a particular network within each congruence class, i.e., $P_{\mathcal{G}_n}(g \mid \theta)$; this calculation is based on Equation 2. These different choices for specification demonstrate the flexibility of CCMs over other network models. For example, the probability distribution for $P_{\phi_1}(x \mid \theta)$ and $P_{\mathcal{G}_n}(g \mid \theta)$ under the Erdős-Rényi (ER) model and ERGMs are identical to the binomial distribution shown in Table 1.

An issue that arises in some ERGM specifications is having a large amount of probability mass on extremal networks (e.g., the empty and complete networks) and a small amount of probability is on networks similar to the one observed; this issue is referred to as near-degeneracy. [1, 58] In the CCM framework, one specifies the probability distribution on congruence classes, $P_{\phi}(x \mid \theta)$; therefore, degeneracy is not an issue.

2.4.2 Examples of CCMs: degree distribution—

Let ϕ_2 denote the mapping from a network to its degree distribution, i.e., $\phi_2(g) = D(g)$. The space of networks of size $n = 4$ contains networks with 11 distinct degree distributions; these degree distributions are listed in Table 2 in Column 2. We denote the congruence class containing networks with degree distribution x as $c_{\phi_2}(x)$. We calculate the number of networks in each of the 11 congruence classes by enumerating all 64 networks in $\mathcal{G}_4$ and summarizing their degree distribution. For larger networks, graph enumeration methods exist for estimating the size of congruence classes. [39, 16]

Table 2 shows two different possibilities for assigning probabilities to the congruence classes, i.e., $P_{\phi_2}(x \mid \theta)$. The first probability distribution for $P_{\phi_2}(x \mid \theta)$ follows the uniform distribution. The second, follows a multinomial distribution with $\theta = [0.15, 0.35, 0.35, 0.15]$. For each of these two distributions on congruence classes, we calculate the probability of a network within each congruence class based on Equation 2. The multinomial distribution is used in the simulation studies and PIRC analysis to assign probabilities to the congruence classes. The paper focuses on integration of multiple data sources to estimate θ for the multinonial distribution.

3 Methods

This section provides technical details on our approach for integrating multiple data sources in order to estimate θ–the parameter (or vector of parameters) that defines the probability distribution on network properties of interest; in this manuscript, the network property of interest is the degree distribution for the contact network. In order to estimate θ, we use the Bayesian paradigm, which provides a natural approach to update prior beliefs regarding a parameter based on additional data. For the analysis, we use the behavioral survey data to develop a prior distribution for θ, and use epidemiological and viral genetic information to formulate the likelihood function. The specification of the Bayesian model is presented below.

3.1 Model

Below we present the proposed likelihood function as well as the prior and posterior distributions. In addition, the section discusses the connections among the data, network model (CCM), and epidemic process. The analysis assumes that transition times (T) and genetic sequences (H) are available for the population of interest; hence, the likelihood is $L(\theta; T, H)$. The prior distribution for θ is defined using the risk behavior survey data.

3.1.1 Likelihood function—

Though interest lies in the estimation of θ, we treat g^c and g^p as extra parameters to simplify the likelihood function.[2] The likelihood function $L(g^c, g^p, \theta; T, H)$ can be factored into five components; the first four follow the work by Groendyke et al. [20] but are expanded for the more general SI^AI^LR epidemic process, while the fifth formalizes the inclusion of genetic sequence data. The likelihood is based on the SI^AI^LR epidemic process and parameters described in the previous section. The first

component is the contribution of contacts over which infections were transmitted as shown below:

$$L_1 = \prod_{(j,k) \in g^p} \left[\beta_A \cdot \exp\left(-\beta_A\left(T_k^A - T_j^A\right)\right) \cdot I_{\left\{T_j^A < T_k^A < T_j^L\right\}} + \beta_L \cdot \exp\left(-\beta_A\left(T_j^L - T_j^A\right) - \beta_L\left(T_k^A - T_j^L\right)\right) \cdot I_{\left\{T_j^L < T_k^A < T_j^R\right\}} \right]. \tag{4}$$

The first two components are separated into two parts corresponding to the two disease stages (acute and long-term). We assume that the time to infection over an edge is exponentially distributed with rate β_A during the acute phase and β_L during the long-term phase. This yields the following likelihood for the second component:

$$L_2 = \prod_{1 \le j \le m, (j,k) \in g^c \setminus g^p} \left[\exp\left(-\beta_A \cdot \left(\min\left(T_j^L, T_k^A\right) - T_j^A\right)\right) \times \exp\left(-\beta_L \cdot \left(\max\left(\min\left(T_j^R, T_k^A\right), T_j^L\right) - T_j^L\right)\right) \times I_{\left\{T_k^A > T_j^A\right\}} + I_{\left\{T_k^A < T_j^A\right\}} \right]. \tag{5}$$

The third component is the contribution associated with the transitions from the acute to the long-term phase of the disease, while the fourth component is the transition from the long-term phase to recovery. We assume that the waiting time in each phase follows an exponential distribution with rate κ. Components 3 and 4 are shown below:

$$L_3 = \prod_{i=1}^{m} \kappa_A \exp\left(-\kappa_A\left(T_i^L - T_i^A\right)\right), \tag{6}$$

and:

$$L_4 = \prod_{i=1}^{m} \kappa_L \exp\left(-\kappa_L\left(T_i^R - T_i^L\right)\right). \tag{7}$$

The fifth component is the contribution of the genetic sequence data:

$$L_5 = \prod_{(j,k) \in g^p} p_s\left(h_k^{t_k} \mid g^p, h_j^{t_j}, T_j^A, T_k^A\right), \tag{8}$$

where t_j and t_k are the times that sequences are collected for individuals j and k, respectively, and $p_s\left(h_k^{t_k} \mid g^p, h_j^{t_j}, T_j^A, T_k^A\right)$ is the probability that the sequence for individual k at time of collection would be $h_k^{t_k}$, given that k was infected by individual j and the sequence for j is $h_j^{t_j}$. This probability distribution could be determined by a model of genetic evolution,

[69, 70, 31] or as a function of the distance between sequences.[62] These five components combined give the following likelihood function for the model:

$$L(g^c, g^p, \theta; T, H) = L_1 L_2 L_3 L_4 L_5 . \tag{9}$$

3.1.2 Prior distribution

The prior distribution for the parameter θ and expanded parameters g^c and g^p, denoted as $\pi_0(g^c, g^p, \theta)$, can be written as the following:

$$\pi_0(g^c, g^p, \theta) = \pi_0(g^c, g^p \mid \theta)\pi_0(\theta) \tag{10}$$

$$= \pi_0(g^p \mid g^c, \theta)\pi_0(g^c \mid \theta)\pi_0(\theta) . \tag{11}$$

Because $\pi_0(g^p \mid g^c, \theta) = \pi_0(g^p \mid g^c)$,

$$\pi_0(g^c, g^p, \theta) = \pi_0(g^p \mid g^c)\pi_0(g^c \mid \theta)\pi_0(\theta) . \tag{12}$$

We model $\pi_0(g^c \mid \theta)$ as a CCM; therefore, the functional form of $\pi_0(g^c \mid \theta)$ is shown in Equation 2.

As discussed above, the ability to integrate behavioral survey data with the other available data for estimating contact network properties arises from the flexibility afforded by CCMs in specifying $\pi_0(g^c \mid \theta)$. The prior distribution for θ, $\pi_0(\theta)$, is specified based on the risk behavior survey data. As in Britton and O'Neill [2] and Groendyke et al. [20], we use the uniform distribution for $\pi_0(g^p \mid g^c)$, which makes all g^p that are possible given a particular contact network g^c equally likely.

3.1.3 Posterior distribution—Together, the priors and likelihood yield the following posterior distribution for the parameters, θ, g^c and g^p:

$$\pi(\theta, g^c, g^p \mid T, H) \propto L(g^c, g^p, \theta; T, H)\pi_0(g^p \mid g^c)\pi_0(g^c \mid \theta)\pi_0(\theta) . \tag{13}$$

3.2 Estimation procedure

Sampling directly from the posterior of this model (Equation 13) is computationally intractable. We instead sample from the posterior using a Markov Chain Monte Carlo (MCMC) approach. Specifically, we use a Gibbs sampler to sample from the posterior. At

each iteration, the parameters g^c, g^p, and θ are sequentially updated resulting in a collection of contact networks, transmission trees, and parameter values that are consistent with the data; details are provided below.

3.2.1 Update contact network—To update g^c, the approach uses a nested MCMC procedure–within the Gibbs sampler–to construct a sequence of graphs, $g_1, \ldots, g_M$; the nested MCMC procedure is a Metropolis-Hastings (MH) algorithm. At each iteration, the MCMC algorithm selects an edge or potential edge between two vertices at random, denoted as e_{ij}, to toggle, i.e., the edge is removed if it is currently in the network and vice versa. This graph is referred to as a proposal network; the proposal at iteration t is denoted as gp_{t-1}. At the end of the iteration, either $g_t = gp_{t-1}$ or $g_t = g_{t-1}$. This selection is based on the following acceptance probability for the MH algorithm:

$$a_{ij} = \begin{cases} 1 & \text{if } (i,j) \in g^p \text{ or } (j,i) \in g^p \\ r_{ij} & \text{if both } i \text{ and } j \text{ remain susceptible} \\ \dfrac{\mu_{ij} r_{ij}}{(1 - r_{ij}) + \mu_{ij} r_{ij}} & \text{otherwise,} \end{cases} \tag{14}$$

where:

$$r_{ij} = \min\left(1, \frac{P_{\mathcal{G}_n}(gp_{t-1} \mid \theta)}{P_{\mathcal{G}_n}(g_{t-1} \mid \theta)}\right), \tag{15}$$

and:

$$\mu_{ij} = \begin{cases} \left(1 - F_A\left(\min\left(T_j^L, T_i^A\right) - T_j^A\right)\right) \times \left(1 - F_L\left(\max\left(\min\left(T_j^R, T_i^A\right), T_j^L\right) - T_j^L\right)\right) & \text{if } I_{\left\{I_j^A < I_i^A\right\}} \\ \left(1 - F_A\left(\min\left(T_i^L, T_j^A\right) - T_i^A\right)\right) \times \left(1 - F_L\left(\max\left(\min\left(T_i^R, T_j^A\right), T_i^L\right) - T_i^L\right)\right) & \text{if } I_{\left\{I_j^A > I_i^A\right\}}; \end{cases} \tag{16}$$

F_A and F_L are the cumulative probability functions for the exponential distribution with parameters β_A and β_L, respectively. Methods to calculate the ratio in r_{ij} are presented in Goyal et al. [13].

3.2.2 Update transmission network—To update g^p, we select an infector for each infected individual j sequentially. The candidate vertices for the individual that infected j, referred to as the parent of j, are vertices that are infectious when individual j became infected and have a contact with j; that is, vertex i is a candidate parent if $T_i^A < T_j^A < T_i^R$ and $(i,j) \in g^c$. Denote these candidate vertices by $i_1, \ldots, i_k$. To specify the probability that i_q is the parent of j, we define the following:

$$x_{(i_q,j)} = \begin{cases} \beta_A \exp\left(-\beta_A\left(T_j^A - T_{i_q}^A\right)\right) & \text{if } T_{i_q}^A < T_j^A < T_{i_q}^L \\ \beta_L \exp\left(-\beta_L\left(T_j^A - T_{i_q}^L\right)\right)\left(1 - F_A\left(T_{i_q}^L - T_{i_q}^A\right)\right) & \text{if } T_{i_q}^L < T_j^A < T_{i_q}^R, \end{cases} \tag{17}$$

and:

$$y_{(i_a,j)} = \begin{cases} \left(1 - F_A\left(T_j^A - T_{i_a}^A\right)\right) & \text{if } T_{i_a}^A < T_j^A < T_{i_a}^L \\ \left(1 - F_A\left(T_{i_a}^L - T_{i_a}^A\right)\right)\left(1 - F_L\left(T_j^A - T_{i_a}^L\right)\right) & \text{if } T_{i_a}^L < T_j^A < T_{i_a}^R. \end{cases} \tag{18}$$

The probability that i_q is the parent of j, given that one of the candidates is known to have infected j, is:

$$\frac{x_{(i_q,j)} p_s\left(h_j^{t_j} \mid g^p, h_{i_q}^{t_{i_q}}, T_{i_q}^A, T_j^A\right) \prod_{a \neq q} y_{(i_a,j)}}{\sum_{z \in \{1,\ldots,k\}} x_{(i_z,j)} p_s\left(h_j^{t_j} \mid g^p, h_{i_z}^{t_{i_z}}, T_{i_z}^A, T_j^A\right) \prod_{a \neq z} y_{(i_a,j)}} = \frac{x^*_{(i_q,j)} p_s\left(h_j^{t_j} \mid g^p, h_{i_q}^{t_{i_q}}, T_{i_q}^A, T_j^A\right)}{\sum_{z \in \{1,\ldots,k\}} x^*_{(i_z,j)} p_s\left(h_j^{t_j} \mid g^p, h_{i_z}^{t_{i_z}}, T_{i_z}^A, T_j^A\right)}, \tag{19}$$

where:

$$x^*_{(i_q,j)} = \begin{cases} \beta_A & \text{if } T_{i_q}^A < T_j^A < T_{i_q}^L \\ \beta_L & \text{if } T_{i_q}^L < T_j^A < T_{i_q}^R. \end{cases} \tag{20}$$

3.2.3 Update parameters for CCM—Based on the posterior distribution shown in Equation 13, the full conditional distribution for θ, $P(\theta \mid g^c, g^p, T, H)$, is proportional to $\pi_0(g^c, \theta)$, which can be further simplified as presented below:

$$P(\theta \mid g^c, g^p, T, H) \propto P_{\mathcal{G}_n}(g^c \mid \theta) * \pi_0(\theta) \tag{21}$$

$$= \left(\frac{1}{\left|c_\phi(\phi(g^c))\right|}\right) P_\phi(\phi(g^c) \mid \theta) * \pi_0\left(\theta\right) \tag{22}$$

$$\propto P_\phi(\phi(g^c) \mid \theta) * \pi_0(\theta). \tag{23}$$

In general, having $\pi_0(\theta)$ and $P_\phi(\phi(g^c) \mid \theta)$ be conjugate distributions will ease computational burden. Sections 4–6 provide examples on specifying these distributions.

4 Simulation study: SEIR epidemic process

This section describes simulation studies for assessing the method's performance at estimating the degree distribution of the contact network. For clarity of presentation, the analysis focuses on only one network property; but CCMs can be defined based on several properties. For example, Goyal et al. [13] and Goyal and De Gruttola [14] specified a CCM with both degree distribution and mixing patterns. As the focus is on investigating the value of integrating multiple data sets, the simulation studies assume that the epidemiological data and genetic data are completely and perfectly observed. Groendyke et al. [20] describes an approach to adjust for missing epidemiological data.

To generate the necessary data for the simulation study (behavior survey, epidemiological, and viral sequence data), we simulate a contact network, an epidemic that propagates over the contact network, a simple viral evolution process, and a behavioral survey sampling process. The epidemic process is parameterized to resemble the spread of SARS-CoV-2. The subsection below provide details of these components for the simulation studies; subsequent subsections provide details on the specification of the prior distribution, viral sequence evolution process, and results of the simulation studies.

4.1 Overview of data generating procedure

The following provides details for the four components that are necessary to generate data needed to illustrate our presented approach:

1. **Contact network model:** The study simulates a population of $n = 1000$ individuals and generates a contact network based on a CCM, where the degree distribution follows a Poisson distribution with parameter λ, representing mean degree. The simulations use λ values from the following set: {10, 20, 30, 40, 50}.

2. **Epidemic process:** The epidemic begins with a single infected individual and proceeds via the stochastic SEIR model detailed earlier. The parameters for the epidemiological model are based on estimates for SARS-CoV-2. Specifically, the parameters for the exponential distribution that governs the lengths of the latent and infectious periods are set such that the duration has a mean of 4 days ($k_E = 4$) and 14 days ($k_I = 14$), respectively.[34] Based on existing literature, $\beta = 0.0097$. [17] In this simulation study, we assume that the epidemic parameters (k_E and k_I) are known and constant throughout the simulation; however, there are recent methods using Bayesian inference to estimate epidemic parameters as well as changes in these parameters over time due to either external (e.g., government interventions) or internal (e.g., evolution of pathogen) causes.[57]

3. **Simple viral evolution process:** The genetic data for the pathogen consist of sequences of 1048 base pairs. The sequences change over time at a rate of 1 substitution per day. In addition, each transmission is associated with

10 substitutions. These viral evolution parameters are selected to illustrate the approach; HIV specific parameters are used in the next section.

4. **Simulate data:** Behavior survey data consist of k sampled individuals, where k can be 25, 50, 100 or 200 individuals (i.e., ranging from 2.5% to 20% of the population). For each sampled individual, the simulation records the number of contacts. We denote the degree for a sampled individual i as $d_i(g^c)$.

4.2 Prior Distribution

To specify $\pi_0(g^c \mid \theta)$, we postulate a CCM where the network property of interest is degree distribution; the PMF is shown below:

$$P_{\mathcal{G}_n}(g^c \mid \theta) = \left(\frac{1}{\left|c_\phi(\phi(g^c))\right|}\right) P_\phi(\phi(g^c) \mid \theta). \tag{24}$$

The investigator selects the probability distribution for $P_\phi(\phi(g^c) \mid \theta)$. Here, we assume $P_\phi(\phi(g^c) \mid \theta)$ follows a multinomial distribution with parameter θ; entry θ_i is the probability that an individual has degree i. Note that our choice of a multinomial distribution for the degree distribution differs from the distribution used to generate the simulation data, which is based on a Poisson distribution. We selected the multinomial distribution because it can represent a wide range of degree distributions for populations due to its large number of parameters allows. For example, it can represent populations with minimal heterogeneity in degree to populations with right-skewed degree distributions. In addition, specifying different distributions for the data generating process and inferential approach helps ensure the conclusions from the approach are not over-optimistic. Furthermore, the ability to specify different distributions illustrates the flexibility of CCMs.

As with $P_\phi(\phi(g^c) \mid \theta)$, the investigator selects the specification of $\pi_0(\theta)$. In these simulations, to specify the prior distribution $\pi_0(\theta)$, we assume that θ is drawn from a Dirichlet distribution with parameter vector α_0. Therefore,

$$\theta \sim Dirichlet(\alpha_0), \tag{25}$$

where the parameter vector α_0 is based on the sampled risk behavior survey data. To set α_0, we conduct a separate Bayesian inference analysis. Specifically, $Dirichlet(\alpha_0)$ is the posterior distribution where the behavior survey data is modeled using a multinomial distribution and the prior distribution is a Dirichlet with parameter α_0^{hyper}. We set α_0^{hyper} to result in a non-informative prior; specifically, α_0^{hyper} is equal to all $\frac{1}{n}$, where n is the number of individuals in the population. Therefore,

$$\alpha_0(j) = \frac{1}{n} + \sum_{i=1}^{k} I_{\{d_i(g^c)=j\}}. \tag{26}$$

The assumption that $P(\phi(g) \mid \theta)$ follows a multinomial distribution and $\pi_0(\theta)$ is a Dirichlet distribution eases the computational burden of updating CCMs parameters as these two distributions are conjugate distributions.

4.3 Viral sequence evolution process

For the viral genetic model, we assume that all sequences are collected simultaneously at the end of the simulation, denoted as time t. We assume that:

$$p_s\left(h_j^{t_j} \mid g^p, h_{i_q}^{t_{i_q}}, T_{i_q}^E, T_j^E\right) \propto Dist\left(h_{i_q}^t, h_j^t\right)^{-1}. \tag{27}$$

4.4 Results

For each potential pair of mean degree λ and number of behavior survey respondents (ranging from 2.5% to 20% of the population), we perform the Bayesian estimation procedure 250 times; therefore, there are a total of 5,000 simulations. We remove simulations with fewer than $m = 50$ infected individuals (5% of the population) at the end of the simulation, which we designate as too few to constitute an epidemic. As expected, the simulations with smaller λs are more likely to have fewer than 50 infected individuals, as the contact networks had fewer edges to transmit the virus; Table 3 shows the number of simulations (out of 250) with sufficient number of infected individuals, stratified by λ and the number of behavior survey respondents.

For each simulation, we construct a Markov chain of length 2, 000. For each of the 2, 000 iterations, g^c, g^p, and θ need to be updated based on the Gibbs sampler described in Section 3. While each of these updates are computationally intensive, the update to g^c results in the most computational burden. The process to update g^c requires a separate MCMC algorithm that is nested within the larger Gibbs sampler. The MCMC to update g^c is similar to sampling from a CCM or ERGM.[13, 27] For the analysis in the paper, we set the number of iterations for the nested MCMC as 10, 000. Therefore, for each of the 5, 000 simulations, 20 million networks were generated–resulting in 100 billion networks for the entire simulation study. During the MCMC procedure, some of the chains enter into a region of the network space with low probability mass (extremely high network density). Given the large size of the space of networks with n vertices ($\mathcal{G}_n$), we remove these chains as it will take a long time to find (or re-find) a network with non-negligible probability mass. The phenomenon where chains enter low probability regions has been observed in the literature.[43] Addressing this phenomenon may require advanced MCMC algorithms such as Hamilton Monte Carlo to address.[25] Table 3 shows the number of simulations contained in the analysis.

To assess when the MCMC chain reaches an approximately steady state, we review trace plots and calculate Geweke's convergence diagnostic.[11] In these simulations, we are estimating a vector of parameters for a multinomial distribution associated with the degree distribution for networks of size $n = 1{,}000$, i.e., there are separate parameters for degrees from 0 to 999. Therefore, we have potentially 5, 000, 000 chains to assess for this simulation study as we conduct 5, 000 simulations in this study. Based on visual review of a subset of trace plots, it appears that the MCMC procedure reaches convergence around the 1000^{th} iteration. Figure 1 shows trace plots from simulations with varying values for $\lambda(10, 20, 30, 40, 50)$ (columns) and sample size for the behavior surveys (25, 50, 100, 200) (rows). Each plot shows the number of nodes with degree equal to λ (y-axis) for 1000 MCMC iterations (x-axis). Each trace plot presented is from the simulation with the median mean squared error (MSE) value for the specific combination of λ and behavior survey sample size; we discuss the calculation for the MSE below. Hence for the results, the analysis uses the chain only between iterations 1001 – 2000; the first 1000 iterations are MCMC burn-in values. To investigate whether iterations 1001 – 2000 are samples from the stationary distribution, we use the Geweke z-score diagnostic to assess whether iterations 1001–2000 differ from iterations 2001–5000 for a subset of 50 simulations using the CODA library in R.[54, 55] Approximately 76.6% of chains show no or minimal variability (for example, no networks–across all simulations-contain a node with a degree of 999); for these chains it is not possible to calculate a Geweke z-score. The median absolute z-score is 1.50 (25% and 75% quantiles are 1.00 and 2.24) for the remaining chains. Therefore, the majority of chains were assessed to be in a stationary state for iterations 1001–2000; this estimate is conservative as we only include chains with sufficient variability to calculate a Geweke z-score. Based on the Geweke z-score, several chains are not in a stationary state for all iterations between 1001–2000. However, running these chains for longer should only increase the performance of our approach; therefore, we consider our estimates conservative. For the chains with variability, we have a median effective sample size of 257 (25% and 75% quantiles are 107 and 310) using the CODA library in R.[54, 55]

For each simulation included in the analysis, we calculate two MSEs–one based on the posterior distribution that integrates multiple data (referred to as $MSE_{posterior}$) and one based on the prior distribution developed from the behavior survey samples (referred to as MSE_{prior}). For each simulation, $MSE_{posterior}$ is evaluated by first calculating the bias squared $\left(b^2\right)$ for each degree, i.e., for a given degree, we calculate the mean number of nodes in g^c, across MCMC iterations from 1001 to 2000, minus the true number of nodes, and then square this difference. Next, we calculate the variance $\left(\sigma^2\right)$ for each degree across these MCMC iterations. Third, we calculate the MSE for each degree as $b^2 + \sigma^2$. Finally, we sum the MSE for each degree, across all degrees (0 to 999), to evaluate the $MSE_{posterior}$ for a simulation. To evaluate MSE_{prior} for each simulation, we draw 1000 samples from the prior Dirichlet distribution estimate and multiply the distribution by the population size; the 1000 samples correspond to the number of MCMC iterations used to estimate $MSE_{posterior}$. This procedure to calculate the MSEs provides a conservative estimate of the improvement of our approach, as the procedure to calculate $MSE_{posterior}$ include more variation than the calculation for MSE_{prior}.

Figure 2 provides boxplots for $MSE_{posterior}$ and MSE_{prior} for degree distribution estimates with (posterior) and without (prior) epidemiological and viral genetic data, respectively. The MSEs are stratified by λ and the number of individuals sampled for the behavior survey. For both $MSE_{posterior}$ and MSE_{prior}, we observe a decrease on average, as the number of sexual behavior survey samples increases from 25 to 200 individuals. For $\lambda = 20$ to 50, we observe that $MSE_{posterior}$ is smaller than MSE_{prior} on average. For $\lambda = 10$, only when the number of individuals sampled is small (25 individuals) does using data integration outperform using solely behavior survey data, on average; see the Discussion for thoughts regarding this result.

We calculate the mean percent improvement in MSE (referred to as $MSE_{improve}$) by comparing $MSE_{posterior}$ and MSE_{prior} :

$$MSE_{improve} = \frac{MSE_{prior} - MSE_{posterior}}{MSE_{prior}} \times 100\,. \tag{28}$$

Values for $MSE_{improve}$ are shown in Figure 3. For example, in simulations where $\lambda = 50$, the inclusion of genetic sequences and epidemiological information resulted in $MSE_{improve}$ estimates of 5.6% percent to 62.1% for the number of sexual behavior survey samples ranging from 200 down to 25 individuals, respectively. In general, we observe that, for a given λ, $MSE_{improve}$ decreases as the number of individuals sampled increases.

5 Simulation Study: SI^AI^LR epidemic process

In this section, we describe a simulation study based on an SI^AI^LR epidemic process to assess the method's performance at estimating the contact network degree distribution in the presence of measurement error. The simulation study is similar for the *SEIR* process (including the formulation of the posterior and prior distributions and likelihood function), except for the data generating procedure, which we describe below.

5.1 Overview of data generating procedure

The following provides details for the four components of the data generating procedure for the SI^AI^LR process:

1. **Contact network model:** A contact network consisting of 1000 individuals is generated based on a CCM where the degree distribution follows a negative binomial distribution fitted to the PIRC HIV data.
2. **Epidemic process:** The simulated epidemic begins with a single infected individual and proceeds via the stochastic SI^AI^LR model detailed earlier. The parameters for the epidemiological model are based on estimates for HIV. Specifically, we set the parameters for the exponential distributions that govern the lengths of the infectious acute and long-term periods such that the durations

have a mean of 3 months ($\kappa_A = 3$) and 42 months ($\kappa_L = 42$), respectively.[18] Based on existing literature, $\beta_A = 0.01505$ and $\beta_L = 0.008911817$.[18]

3. **Simple viral evolution process:** Genetic data for the pathogen consists of sequences of 1048 base pairs. These sequences change based on an annual per site mutation rate of 0.0012 substitutions.[46] In addition, each transmission is associated with 1 substitution to ensure unique sequences for each individual.

4. **Simulate data:** The simulation studies generate behavior survey data by sampling k individuals, where k can be 25, 50, 100 or 200 individuals (i.e., ranging from 2.5% to 20% of the population). For each sampled individual, we introduce measurement error by inflating or deflating their number of contacts by a given percentage, representing over- or under- reporting, respectively. Each simulation selects a measurement error percentage from the set: {−50%, −40%, …, +50%}, where negative percentages indicate individuals reported fewer numbers of partners than they actually had, and positive percentages indicate they reported greater numbers of partners.

5.2 Results

For each measurement error percentage and number of behavior survey respondents (ranging from 2.5% to 20% of the population), we perform the procedure 250 times, resulting in a total of 11, 000 simulations. We remove simulations in which fewer than 50 individuals (5% of the population) were infected at the end of the simulation. We apply an approach similar to that in the previous section to identify chains that were never able to find (or re-find) a network with non-negligible probability mass. Table 4 shows the number of simulations of the sets of 250 that had a sufficient amount of infected individuals and the number of simulations that are included in the analysis.

Figure 4 provides boxplots for $MSE_{posterior}$ and MSE_{prior} for degree distribution estimates with (posterior) and without (prior) the epidemiological and viral genetic data, respectively. The MSEs are stratified by the measurement error percentage and the number of behavior survey respondents. In the absence of measurement error (misreporting = 0%), we observe, on average, a decrease in MSE (from 70.1% to 32.9%) when integrating multiple data compared to using only risk behavior data. The number of sexual behavior survey samples ranged from 25 to 200. When the number of reported partners is higher than the actual number of sexual partners, we observe a consistent large decrease in MSE. Even when fewer partners are reported than the actual number, we nonetheless observe a decrease in MSE, on average, but to a smaller extent. Figure 5 shows values for the median $MSE_{improve}$ across the settings.

6 Investigation of PIRC Cohort

In this section, we apply our approach to the PIRC cohort to demonstrate using the approach on HIV real-world HIV data. PIRC is an observational cohort of antiretroviral naive people newly diagnosed with acute, early and established HIV, which started enrolling participants on July 1, 1996. As mentioned above, PIRC participants are asked to provide behavior

risk information, viral genetic sequences, and epidemiological data during longitudinal follow-up. Regarding risk behavior data, PIRC participants provide information on the number of sexual partners they had in the last 3 months. Napper et al. [47] found that a recall period of 3 months produced the most reliable data. In terms of epidemiological data, it is uncommon to know the exact date of infection (DoI) for individuals due to the delay in diagnosing HIV. In the United States, there is approximately a 3-year gap between infection and diagnosis.[6] The use of a CD4 depletion model is a common approach to estimate DoI, which uses an individual's date of diagnosis and first cell count after diagnosis (prior to treatment).[61] Using data from PIRC, Tang et al. found that the sensitivity and specificity of the CD4 depletion model can be low for recent infections.[63] At time of their enrollment, participants of the PIRC study are newly diagnosed with HIV infection and an estimated DoI (eDoI) is calculated for those with recent infection using virologic and serologic data. Data on PIRC participants also include dates of treatment initiation and achieving viral suppression.

In this analysis, we focus on estimating the degree distribution for the contact network among the PIRC participants. That is, we aim to estimate the total number of sexual partners per PIRC participant restricted to PIRC participants over the entire duration of follow-up. Therefore, our available risk behavior data of the reported number of sexual partners in the previous 3 months (which includes partners both enrolled and not enrolled in the PIRC cohort) can be viewed as providing noisy information (i.e., having measurement error) regarding the total number.

We limit the analysis to the 535 PIRC participants who had behavioral risk information, epidemiological data, and a viral sequence. For the analysis, T^A is set to the eDoI; the length of the acute phase was assumed to be 3 months, i.e., $T^L = T^A + 3$. We set T^R as the date of treatment initiation. For participants without a recorded ART initiation date, the treatment time is set as the date of achieving viral suppression. We use the responses from all 535 participants in the analysis to construct a prior distribution for parameters associated with degree distribution. For each pair of individuals with a sequence, we calculated the genetic distance. Pairwise distance was measured using the Tamura-Nei 93 algorithm.[62] In accordance with previous analyses, for pairwise distances less than 1.5%, we assume this indicates evidence of possible linkage.[50] For distances greater than 1.5%, we assumed that the distance is not informative of transmission, but does not preclude this possibility. The formulation of the posterior and prior distributions and likelihood function are the same as the simulation studies.

As the true number of sexual partners among PIRC participants is not available, the analysis investigates changes in the degree distribution between estimates derived from the survey data and those using the presented approach. Figure 6 shows density plots for the number of individuals with each degree from 0 to 20 for estimates from the survey solely (blue area) and estimates from our approach (red area). The estimated mean number of partners from the survey, 9.09 partners, decreases to 2.18 partners, on average, with the application of our Bayesian approach. Given the difference between the survey question and the quantity being estimated (that is, degree distribution only among PIRC participants), we would expect

to see this decrease. The magnitude of decrease provides insight into the potential size of the number of individuals missing from the contact network that includes both PIRC participants and their partners (irrespective of whether the partners are enrolled in PIRC); an important limitation is that the reported number of partners is limited to 3 months.

7 Discussion

In a world where diverse data sources are increasingly easy to obtain, there is a great need for principled methods to combine those data sources. In the infectious disease realm, we now have an array of data types that inform different aspects of the complex system underlying disease transmission. This creates an opportunity to learn about properties of that system that are otherwise difficult or impossible to measure accurately. In this manuscript, we present such a principled approach to integrate multiple data sources associated with infectious disease dynamics in order to estimate properties of the underlying contact network. An understanding of the structure of contacts makes it possible to develop more efficient disease prevention programs that use information about network structure. The core of the approach is the use of CCMs for analyses of networks. The ability of CCMs to represent a broad family of different models provides the flexibility necessary to integrate multiple data sources. Beyond CCMs, other network models also might potentially be used to integrate multiple data; these include SIENA,[60] latent space models,[24, 30] and hierarchical longitudinal models.[52] There exists promising research in developing data integration methods using these flexible models to investigate dyad-dependent network properties.

Our simulation results show that integration of routinely-collected data can lead to large increases in precision and accuracy of contact network estimates. The simulations using an SEIR model parameterized to resemble SARS-CoV-2 spread show that, in most settings, the MSE of our network property (degree distribution) estimates is lower when based on data integrated across multiple sources. Exceptions to this finding arose in settings where the contact network is sparse (small mean degree) and the percentage of the population with risk behavior data exceeds 5%. As we simulated the degree distribution based on a Poisson distribution, a small mean degree corresponds to small variance in the degrees of individuals. Therefore, in the settings where MSE did not improve under our approach, we have more survey data (prior information) and fewer entries of the degree distribution with non-zero values. In such settings, it seems that a straightforward estimation procedure may have benefits compared to our Bayesian framework. The benefits of data integration for estimating properties of the contact network are further demonstrated in the simulations using the SI^AI^LR model and epidemic process that resembles HIV. These simulations suggest a potentially substantial reduction in MSE in settings reporting bias is present. In general, the improvement from data integration appears to greater in setting with over-reporting compared to under-reporting of the number of partners.

We apply our approach to data from PIRC, a longstanding cohort of persons newly diagnosed with HIV. The demonstration of our approach that uses the PIRC cohort highlights several directions of further research. First, for the simulation studies and the analysis of PIRC data, we consider the possibility of measurement error, but not of biases

in the sampling of individuals who respond to the risk behavior survey. Second, we assume that the transmission probability is the same per sexual partner; previous research has shown the importance of considering differences in transmission probabilities among various types of partners (e.g., those in stable relationships, sex workers, etc.).[12] Third, our approach assumes that the contact network is static. Given the duration of HIV studies, including PIRC, this assumption will probably be violated. In fact, temporal patterns in contacts have been shown to impact the spread of disease [26], implying a need to extend the approach to dynamic networks. We note that our approach has sufficient flexibility to accommodate these improvements. For example, one advantage of the CCM framework is that known biases in selection can be incorporated directly in the formulation of $\pi_0(\theta)$. Furthermore, the CCM framework has already been expanded to dynamic networks.[15]

In addition to addressing the complexities in the PIRC data, there is a need to develop statistical methods in several areas. The first is assessing MCMC algorithms when modeling a large number of network model parameters, which is possible for CCMs and demonstrated in our simulation studies. There are challenges in using MCMC algorithms in high-dimensions as well as assessing their convergence. [7, 56] Furthermore, parameters for network models are highly correlated.[8] Another area of further research is on statistical approaches for addressing the range of missing data common in infectious disease data. For example, it is common for HIV sequences and epidemiological data to be missing for individuals.[4, 41] Within the presented framework, it possible to impute epidemiological times for a particular event (e.g., infection) based on an observed time for another event (e.g., diagnosis) within the Gibbs sampler.[20] Such imputation requires the ability to develop a probability distribution for the missing event time based on observed times for other events. For example, if diagnosis times are only available (along with biomedical information), one can develop a distribution for the diagnosis delay to estimate date of infection.[61] Finally, methodological research is necessary for integrating novel network data. For example, responses to the COVID-19 pandemic are making use of a range of new technologies and devices for collecting network data, such as those that arise from Bluetooth proximity data and geo-location information available from smart devices. This information can be used to inform contact structure. Integrating data from these sources with surveillance and other epidemiological data will be necessary to best inform design and evaluate interventions intended to mitigate the spread of infectious diseases. The presented approach provides a rigorous framework for developing such statistical methods.

Acknowledgments

This research is supported by grants from the National Institutes of Health (R37 AI-51164, R01 AI-147441, R01 MH-100974, P30 AI-036214, and R24 AI-106039) and by the James Pendleton Charitable Trust. Conflict of Interest: SJL has received funding from Gilead Sciences paid to her institution.

References

[1]. Blackburn B and Handcock MS (2022). Practical network modeling via tapered exponential-family random graph models. Journal of Computational and Graphical Statistics, pages 1–14.

[2]. Britton T and O'Neill PD (2002). Bayesian inference for stochastic epidemics in populations with random social structure. Scandinavian Journal of Statistics, 29:375–390.

[3]. Carnegie NB (2018). Effects of contact network structure on epidemic transmission trees: implications for data required to estimate network structure. Statistics in Medicine, 37(2):236–248. sim.7259. [PubMed: 28192859]

[4]. Carnegie NB, Wang R, Novitsky V, and DeGruttola VG (2013). Phylogenetic linkage among HIV-infected village residents in Botswana: Estimation of clustering rates in the presence of missing data. Harvard University Biostatistics Working Paper Series, 160:1–29.

[5]. Coffee M, Lurie MN, and Garnett GP (2007). Modelling the impact of migration on the hiv epidemic in south africa. Aids, 21(3):343–350. [PubMed: 17255741]

[6]. Dailey AF, Hoots BE, Hall HI, Song R, Hayes D, Fulton P Jr, Prejean J, Hernandez AL, Koenig LJ, and Valleroy LA (2017). Vital signs: human immunodeficiency virus testing and diagnosis delays—united states. Morbidity and Mortality Weekly Report, 66(47):1300. [PubMed: 29190267]

[7]. Durmus A (2016). High dimensional Markov chain Monte Carlo methods: theory, methods and applications. PhD thesis, Université Paris-Saclay (ComUE).

[8]. Duxbury SW (2021). Diagnosing multicollinearity in exponential random graph models. Sociological Methods & Research, 50(2):491–530.

[9]. Eaton JW, Johnson LF, Salomon JA, Bärnighausen T, Bendavid E, Bershteyn A, Bloom DE, Cambiano V, Fraser C, Hontelez JA, et al. (2012). Hiv treatment as prevention: systematic comparison of mathematical models of the potential impact of antiretroviral therapy on hiv incidence in south africa. PLoS medicine, 9(7):e1001245. [PubMed: 22802730]

[10]. Frost SD and Volz EM (2013). Modelling tree shape and structure in viral phylodynamics. Philosophical Transactions of the Royal Society B, 368:20120208.

[11]. Geweke J (1992). Evaluating the accuracy of sampling-based approaches to the calculations of posterior moments. Bayesian statistics, 4:641–649.

[12]. Gorbach PM and Holmes KK (2003). Transmission of stis/hiv at the partnership level: beyond individual-level analyses. Journal of Urban Health, 80(3):iii15–iii25. [PubMed: 14713668]

[13]. Goyal R, Blitzstein J, and De Gruttola V (2014). Sampling networks from their posterior predictive distribution. Network Science, 2(01):107–131. [PubMed: 25339990]

[14]. Goyal R and De Gruttola V (2018). Inference on network statistics by restricting to the network space: applications to sexual history data. Statistics in medicine, 37(2):218–235. [PubMed: 28745004]

[15]. Goyal R and De Gruttola V (2020). Dynamic network prediction. Network Science, 8(4):574–595. [PubMed: 36035743]

[16]. Goyal R and De Gruttola V (2022). A general computational approach for counting labeled graphs. Algorithms, 16(1): 16.

[17]. Goyal R, Hotchkiss J, Schooley RT, De Gruttola V, and Martin NK (2021a). Evaluation of severe acute respiratory syndrome coronavirus 2 transmission mitigation strategies on a university campus using an agent-based network model. Clinical Infectious Diseases.

[18]. Goyal R, Hu C, Klein PW, Hotchkiss J, Morris E, Mandsager P, Cohen SM, Luca D, Gao J, Jones A, et al. (2021b). Development of a mathematical model to estimate the cost-effectiveness of hrsa's ryan white hiv/aids program. JAIDS Journal of Acquired Immune Deficiency Syndromes, 86(2):164–173. [PubMed: 33109934]

[19]. Groendyke C, Welch D, and Hunter DR (2011). Bayesian inference for contact networks given epidemic data. Scandinavian Journal of Statistics, 38:600–616.

[20]. Groendyke C, Welch D, and Hunter DR (2012). A network-based analysis of the 1861 Hagelloch measles data. Biometrics, 68:755–765. [PubMed: 22364540]

[21]. Hansson D, Leung KY, Britton T, and Strömdahl S (2019). A dynamic network model to disentangle the roles of steady and casual partners for hiv transmission among msm. Epidemics, 27:66–76. [PubMed: 30738786]

[22]. Harary F and Palmer EM (2014). Graphical enumeration. Elsevier.

[23]. Helleringer S, Kohler H, Kalilani-Phiri L, Mkandawire J, and Armbruster B (2011). The reliability of sexual partnership histories: implications for the measurement of partnership concurrency during surveys. AIDS (London, England), 25(4):503. [PubMed: 21139490]

[24]. Hoff PD, Raftery AE, and Handcock MS (2002). Latent space approaches to social network analysis. Journal of the american Statistical association, 97(460):1090–1098.

[25]. Hoffman MD, Gelman A, et al. (2014). The no-u-turn sampler: adaptively setting path lengths in hamiltonian monte carlo. J. Mach. Learn. Res, 15(1):1593–1623.

[26]. Holme P and Saramäki J (2012). Temporal networks. Physics reports, 519(3):97–125.

[27]. Hunter DR, Handcock MS, Butts CT, Goodreau SM, and Morris M (2008). ergm: A package to fit, simulate and diagnose exponential-family models for networks. Journal of Statistical Software, 24(3):1–29. [PubMed: 18612375]

[28]. Jenness SM, Goodreau SM, Rosenberg E, Beylerian EN, Hoover KW, Smith DK, and Sullivan P (2016). Impact of the centers for disease control's hiv preexposure prophylaxis guidelines for men who have sex with men in the united states. The Journal of infectious diseases, 214(12):1800–1807. [PubMed: 27418048]

[29]. Keeling MJ (2005). The implications of network structure for epidemic dynamics. Theoretical Population Biology, 67:1–8. [PubMed: 15649519]

[30]. Kim B, Lee KH, Xue L, and Niu X (2018). A review of dynamic network models with latent variables. Statistics surveys, 12:105. [PubMed: 31428219]

[31]. Kosiol C, Holmes I, and Goldman N (2007). An empirical codon model for protein sequence evolution. Molecular biology and evolution, 24(7):1464–1479. [PubMed: 17400572]

[32]. Krivitsky PN and Morris M (2017). Inference for social network models from egocentrically sampled data, with application to understanding persistent racial disparities in hiv prevalence in the us. The annals of applied statistics, 11(1):427. [PubMed: 29276550]

[33]. Krivitsky PN, Morris M, and Bojanowski M (2022). Impact of survey design on estimation of exponentialfamily random graph models from egocentrically-sampled data. Social Networks, 69:22–34. [PubMed: 35400801]

[34]. Larremore DB, Fosdick BK, Bubar KM, Zhang S, Kissler SM, Metcalf CJE, Buckee CO, and Grad YH (2021). Estimating sars-cov-2 seroprevalence and epidemiological parameters with uncertainty from serological surveys. Elife, 10:e64206. [PubMed: 33666169]

[35]. Laumann EO (1994). The social organization of sexuality: Sexual practices in the United States. University of Chicago Press.

[36]. Le T, Wright EJ, Smith DM, He W, Catano G, Okulicz JF, Young JA, Clark RA, Richman DD, Little SJ, et al. (2013). Enhanced cd4+ t-cell recovery with earlier hiv-1 antiretroviral therapy. New England Journal of Medicine, 368(3):218–230. [PubMed: 23323898]

[37]. Leventhal GE, Kouyos R, Stadler T, von Wyl V, Yerly S, Böni J, Cellerai C, Klimkait T, Günthard HF, and Bonhoeffer S (2012). Inferring epidemic contact structure from phylogenetic trees. PLoS Computational Biology, 8:e1002413. [PubMed: 22412361]

[38]. Lewin B, Fugl-Meyer K, Helmius G, Lalos A, and Månsson S (1996). Sex in sweden. Stockholm: National Institute of Public Health.

[39]. Liebenau A and Wormald N (2017). Asymptotic enumeration of graphs by degree sequence, and the degree sequence of a random graph. arXiv preprint arXiv:1702.08373.

[40]. Magosi LE, Zhang Y, Golubchik T, DeGruttola V, Tchetgen ET, Novitsky V, Moore J, Bachanas P, Segolodi T, Lebelonyane R, et al. (2022). Deep-sequence phylogenetics to quantify patterns of hiv transmission in the context of a universal testing and treatment trial-bcpp/ya tsie trial. Elife, 11:e72657. [PubMed: 35229714]

[41]. Mazrouee S, Hallmark CJ, Mora R, Del Vecchio N, Carrasco Hernandez R, Carr M, McNeese M, Fujimoto K, and Wertheim JO (2022). Impact of molecular sequence data completeness on hiv cluster detection and a network science approach to enhance detection. Scientific Reports, 12(1):19230. [PubMed: 36357480]

[42]. Miller JC (2009). Spread of infectious disease through clustered populations. Journal of the Royal Society Interface, 6:1121–1134. [PubMed: 19324673]

[43]. Minsley BJ (2011). A trans-dimensional bayesian markov chain monte carlo algorithm for model assessment using frequency-domain electromagnetic data. Geophysical Journal International, 187(1):252–272.

[44]. Moreno Y, Pastor-Satorras R, and Vespignani A (2002). Epidemic outbreaks in complex heterogeneous networks. The European Physical Journal B-Condensed Matter and Complex Systems, 26(4):521–529.

[45]. Morris M and Kretzschmar M (1995). Concurrent partnerships and transmission dynamics in networks. Social Networks, 17:299–318.

[46]. Moshiri N, Ragonnet-Cronin M, Wertheim JO, and Mirarab S (2019). Favites: simultaneous simulation of transmission networks, phylogenetic trees and sequences. Bioinformatics, 35(11):1852–1861. [PubMed: 30395173]

[47]. Napper LE, Fisher DG, Reynolds GL, and Johnson ME (2010). Hiv risk behavior self-report reliability at different recall periods. AIDS and Behavior, 14(1):152–161. [PubMed: 19475504]

[48]. Newman ME (2010). Networks An Introduction. Oxford University Press, New York.

[49]. Newman MEJ (2002). Mixing patterns in networks. Physical Review E, 67:026126.

[50]. Oster AM, Wertheim JO, Hernandez AL, Ocfemia MCB, Saduvala N, and Hall HI (2015). Using molecular hiv surveillance data to understand transmission between subpopulations in the united states. Journal of acquired immune deficiency syndromes (1999), 70(4):444. [PubMed: 26302431]

[51]. Pastor-Satorras R and Vespignani A (2002). Immunization of complex networks. Physical review E, 65(3):036104.

[52]. Paul S and O'Malley AJ (2013). Hierarchical longitudinal models of relationships in social networks. Journal of the Royal Statistical Society: Series C (Applied Statistics), 62(5):705–722. [PubMed: 24729637]

[53]. Petrovic S (2017). A survey of discrete methods in (algebraic) statistics for networks. Algebraic and Geometric Methods in Discrete Mathematics, 685:260–281.

[54]. Plummer M, Best N, Cowles K, and Vines K (2006). Coda: Convergence diagnosis and output analysis for mcmc. R News, 6(1):7–11.

[55]. R Core Team (2021). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

[56]. Rajaratnam B and Sparks D (2015). Mcmc-based inference in the era of big data: A fundamental analysis of the convergence complexity of high-dimensional chains. arXiv preprint arXiv:1508.00947.

[57]. Ray D, Salvatore M, Bhattacharyya R, Wang L, Du J, Mohammed S, Purkayastha S, Halder A, Rix A, Barker D, et al. (2020). Predictions, role of interventions and effects of a historic national lockdown in india's response to the covid-19 pandemic: data science call to arms. Harvard data science review, 2020(Suppl 1).

[58]. Schweinberger M (2011). Instability, sensitivity, and degeneracy of discrete exponential families. Journal of the American Statistical Association, 106(496):1361–1370. [PubMed: 22844170]

[59]. Shalizi CR and Rinaldo A (2013). Consistency under sampling of exponential random graph models. Annals of Statistics, 41(2):508–535. [PubMed: 26166910]

[60]. Snijders TA, Van de Bunt GG, and Steglich CE (2010). Introduction to stochastic actor-based models for network dynamics. Social networks, 32(1):44–60.

[61]. Song R, Hall HI, Green TA, Szwarcwald CL, and Pantazis N (2017). Using cd4 data to estimate hiv incidence, prevalence, and percent of undiagnosed infections in the united states. JAIDS Journal of Acquired Immune Deficiency Syndromes, 74(1):3–9. [PubMed: 27509244]

[62]. Tamura K and Nei M (1993). Estimation of the number of nucleotide substitutions in the control region of mitochondrial dna in humans and chimpanzees. Molecular biology and evolution, 10(3):512–526. [PubMed: 8336541]

[63]. Tang M, Goyal R, Anderson C, Mehta S, and Little S (2022). Estimating incidence and assessing the impact of ending the hiv epidemic initiatives. Under Review.

[64]. Todd J, Cremin I, McGrath N, Bwanika J, Wringe A, Marston M, Kasamba I, Mushati P, Lutalo T, and Hosegood V (2009). Reported number of sexual partners: comparison of data from four african longitudinal studies. Sexually transmitted infections, 85 (Suppl 1):i72–i80. [PubMed: 19307344]

[65]. Wang R, Goyal R, Lei Q, Essex M, and DeGruttola V (2013). Sample size considerations in the design of cluster randomized trials of combination HIV prevention. Harvard University Biostatistics Working Paper Series, 161:1–29.

[66]. Welch D (2011). Is network clustering detectable in transmission trees? Viruses, 3:659–676. [PubMed: 21731813]

[67]. Wirth KE, Gaolathe T, Holme MP, Mmalane M, Kadima E, Chakalisa U, Manyake K, Mbikiwa AM, Simon SV, Letlhogile R, et al. (2020). Population uptake of hiv testing, treatment, viral suppression, and male circumcision following a community-based intervention in botswana (ya tsie/bcpp): a cluster-randomised trial. The Lancet HIV, 7(6):e422–e433. [PubMed: 32504575]

[68]. Xiridou M, Geskus R, De Wit J, Coutinho R, and Kretzschmar M (2003). The contribution of steady and casual partnerships to the incidence of hiv infection among homosexual men in amsterdam. Aids, 17(7):1029–1038. [PubMed: 12700453]

[69]. Yang Z (1994). Maximum likelihood phylogenetic estimation from dna sequences with variable rates over sites: approximate methods. Journal of Molecular evolution, 39(3):306–314. [PubMed: 7932792]

[70]. Zaheri M, Dib L, and Salamin N (2014). A generalized mechanistic codon model. Molecular Biology and Evolution, 31(9):2528–2541. [PubMed: 24958740]

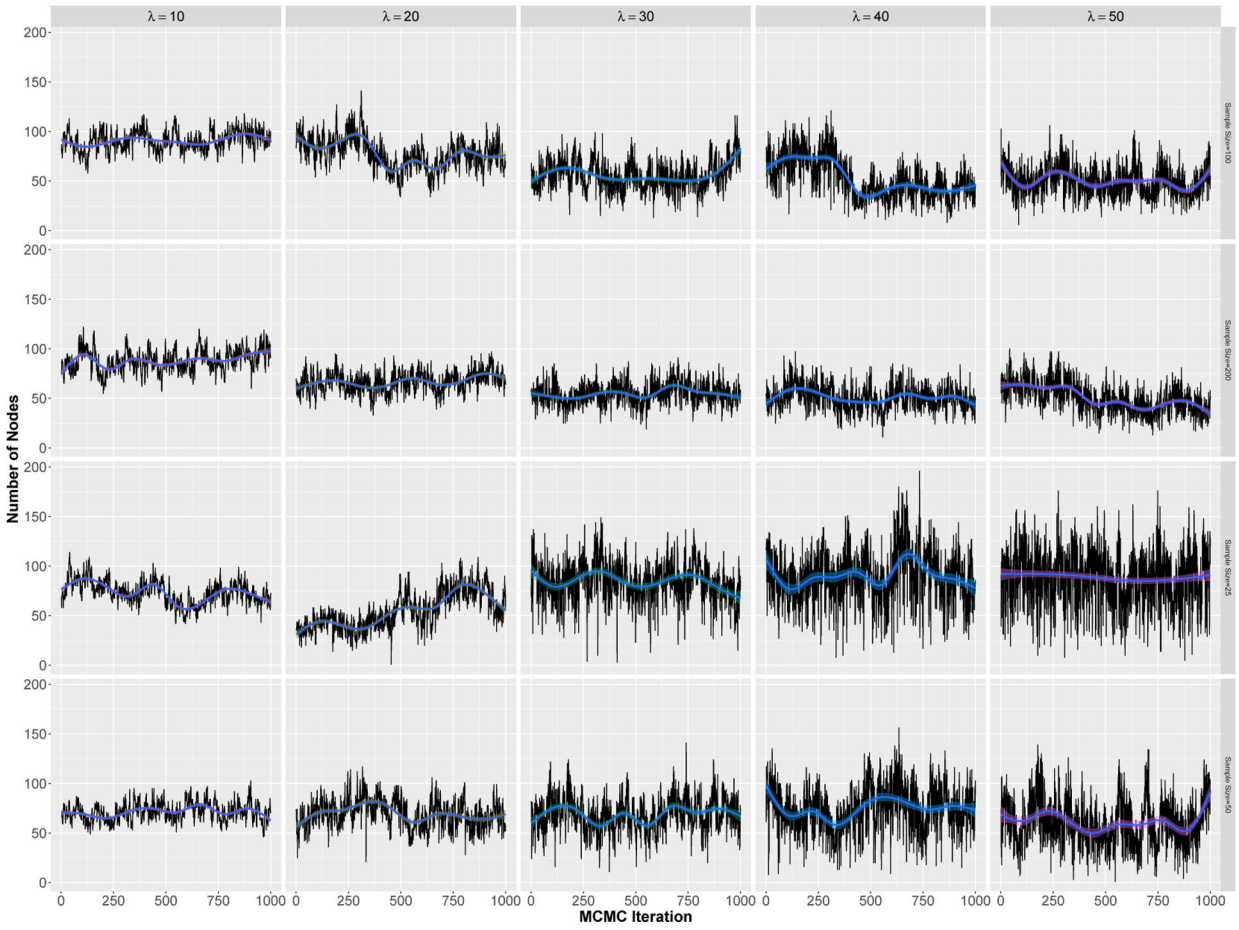

Figure 1:

Trace plots from simulations with varying values for $\lambda(10,20,30,40,50)$ (columns) and sample size for the behavior surveys (25, 50, 100, 200) (rows) are presented. Each plot shows the number of nodes with degree equal to λ (y-axis) for 1000 MCMC iterations (x-axis). Each trace plot presented is from the simulation with the median MSE value for the specific combination of λ and behavior survey sample size. A smooth (blue) line is included across the points.

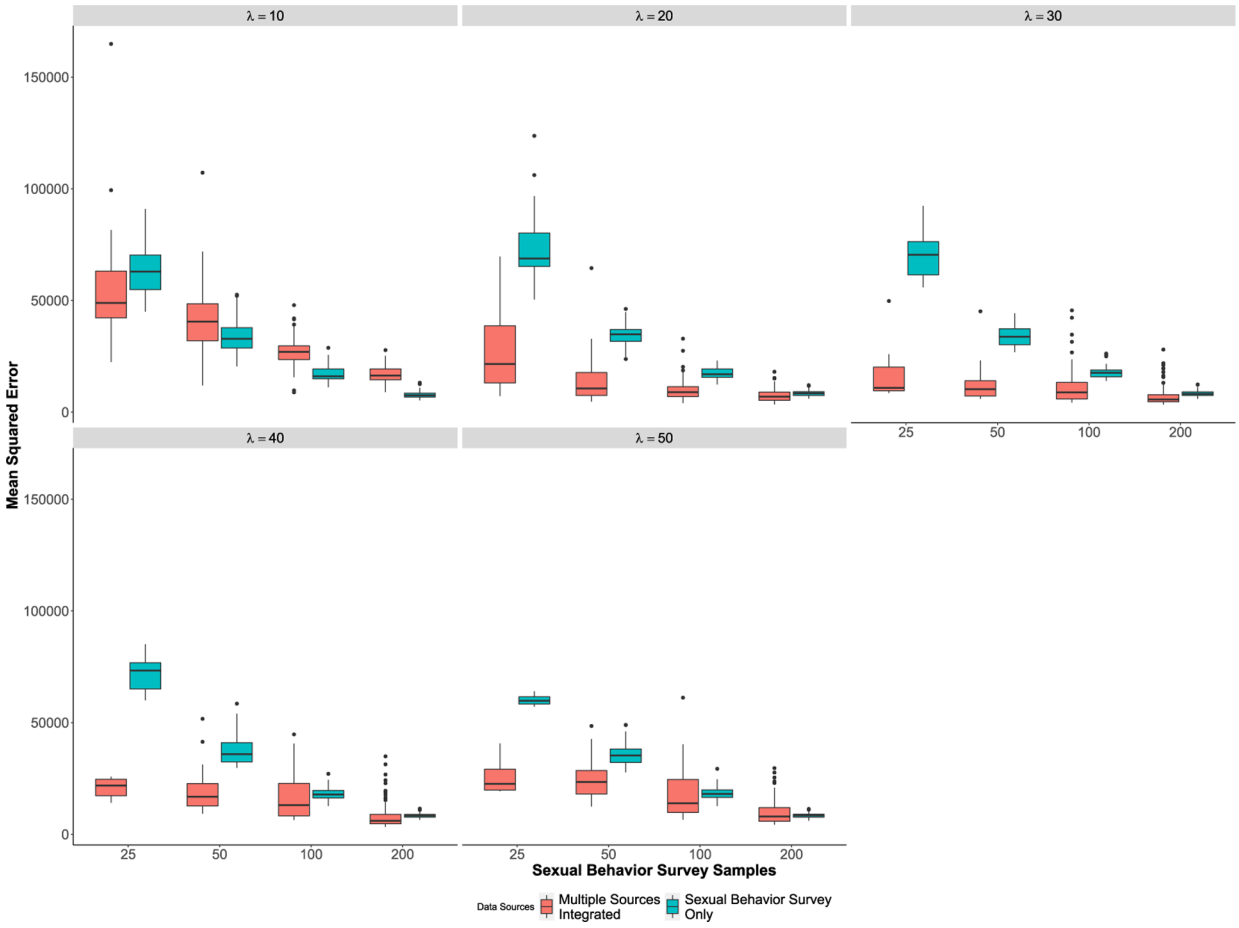

Figure 2:

SEIR epidemic process simulation study boxplots of the MSE values for degree distribution estimates with (posterior) and without (prior) epidemiological and viral genetic data across values for the mean degree of the network (λ).

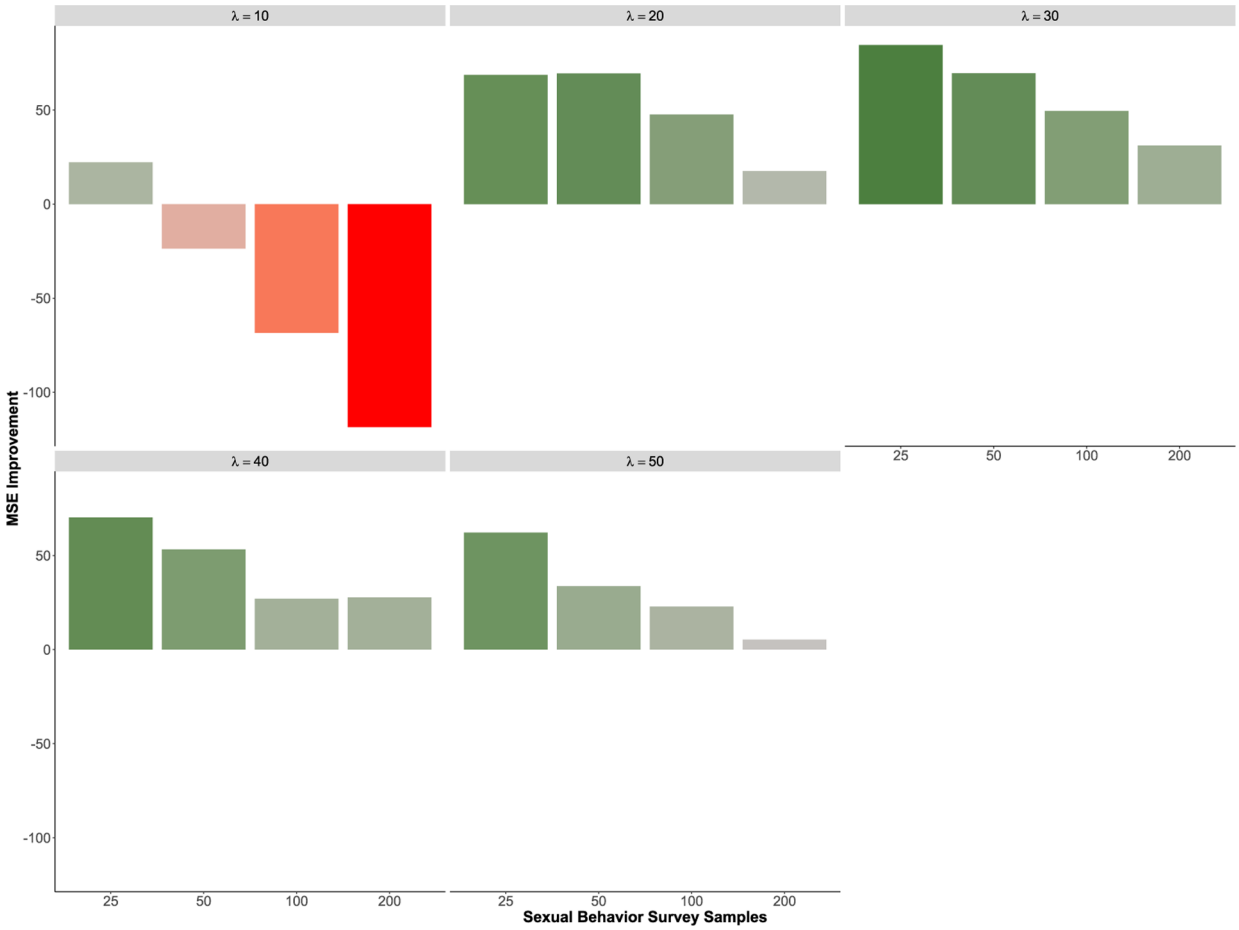

Figure 3:

Improvement in MSE ($MSE_{improve}$) for the *SEIR* epidemic process simulation study. The plots show the median percent improvement in MSE for degree distribution estimates with the inclusion of the epidemiological and viral genetic data compared to without such data across values for the mean degree of the network (λ).

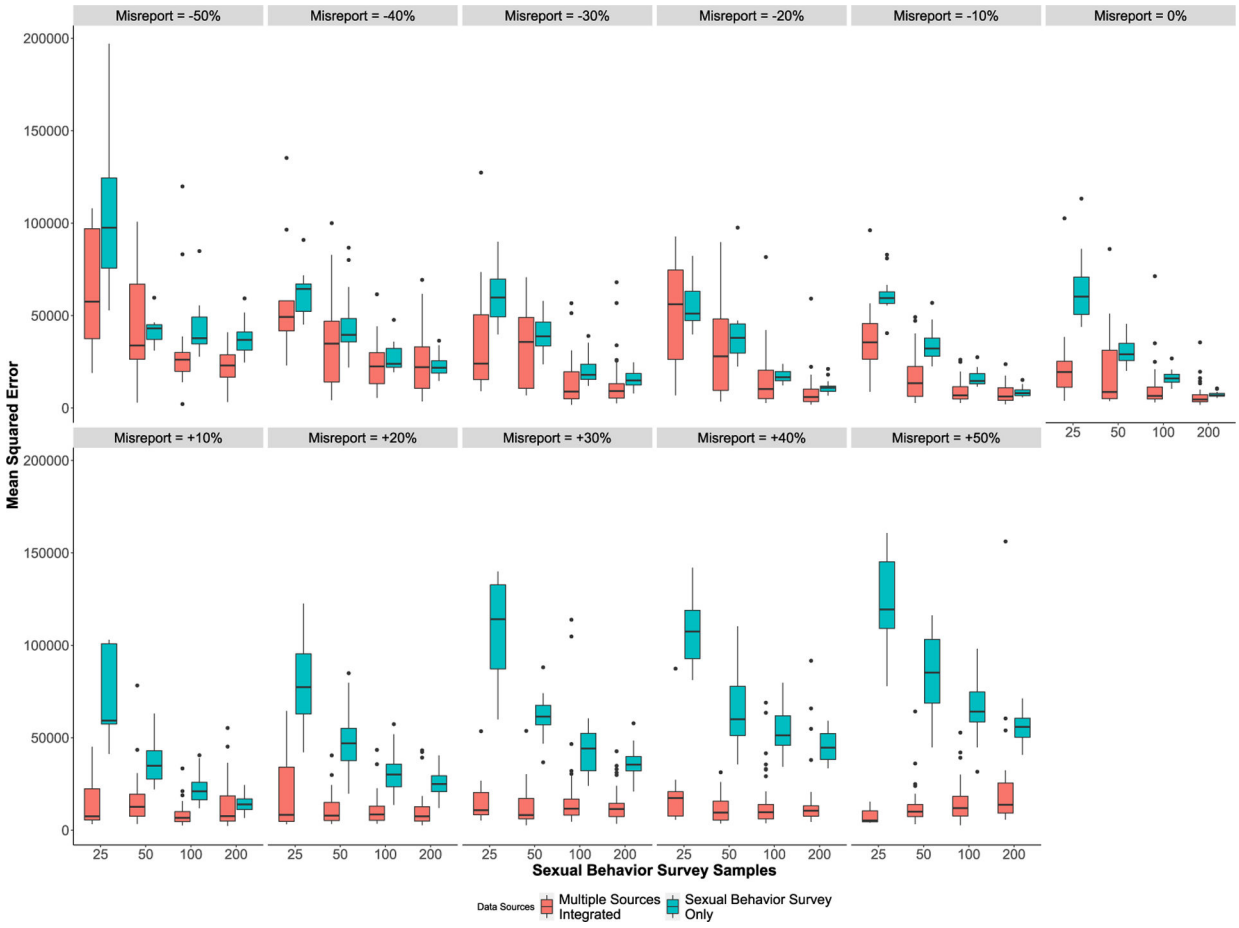

Figure 4:

SI^AI^LR epidemic process simulation study boxplots of the MSE values for degree distribution estimates with (posterior) and without (prior) epidemiological and viral genetic data across values for the misreporting factor.

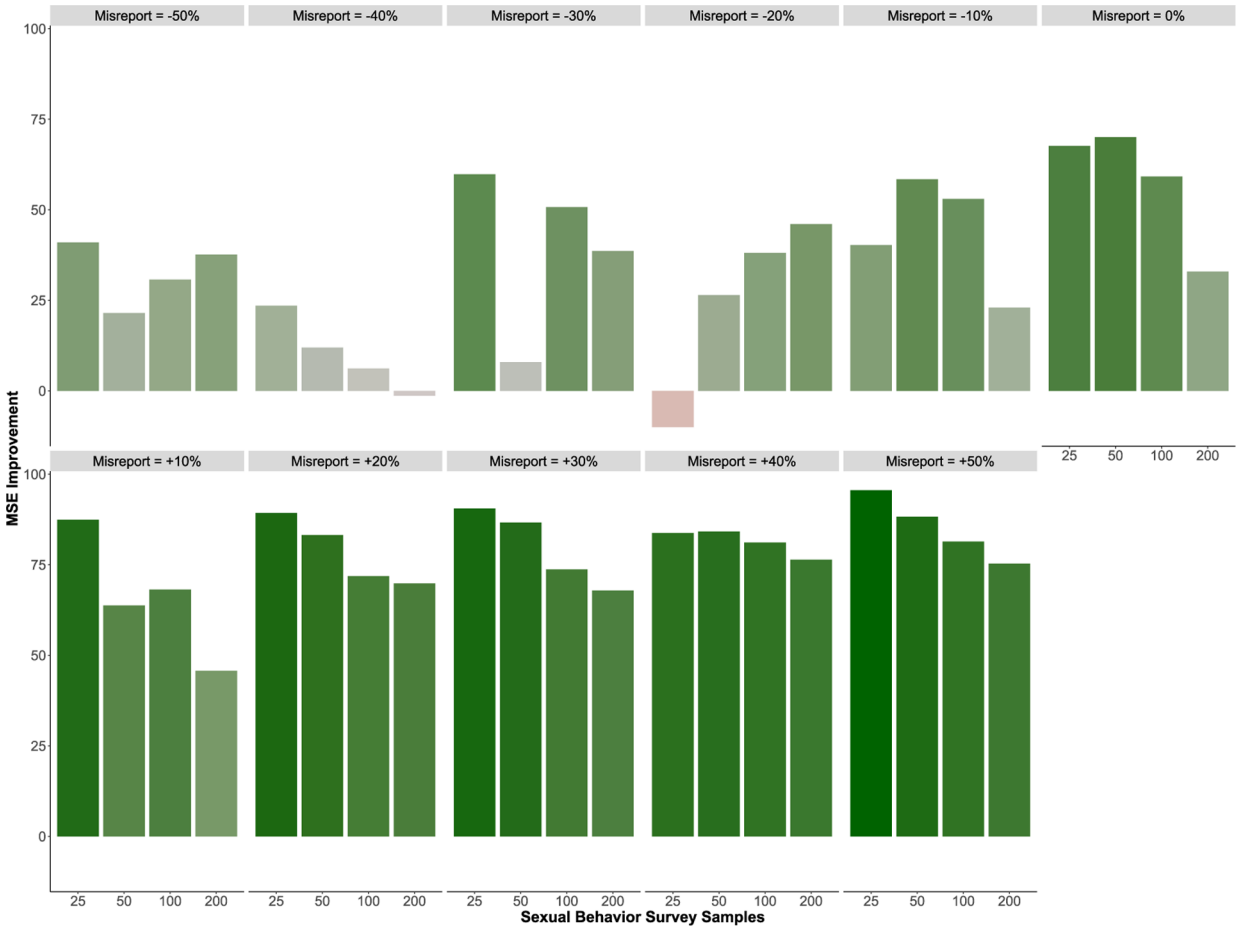

Figure 5:

Improvement in MSE ($MSE_{improve}$) for the SI^AI^LR epidemic process simulation study. The plots show the median percent improvement in MSE for degree distribution estimates with the inclusion of the epidemiological and viral genetic data compared to without such data across values for the measurement error percentage.

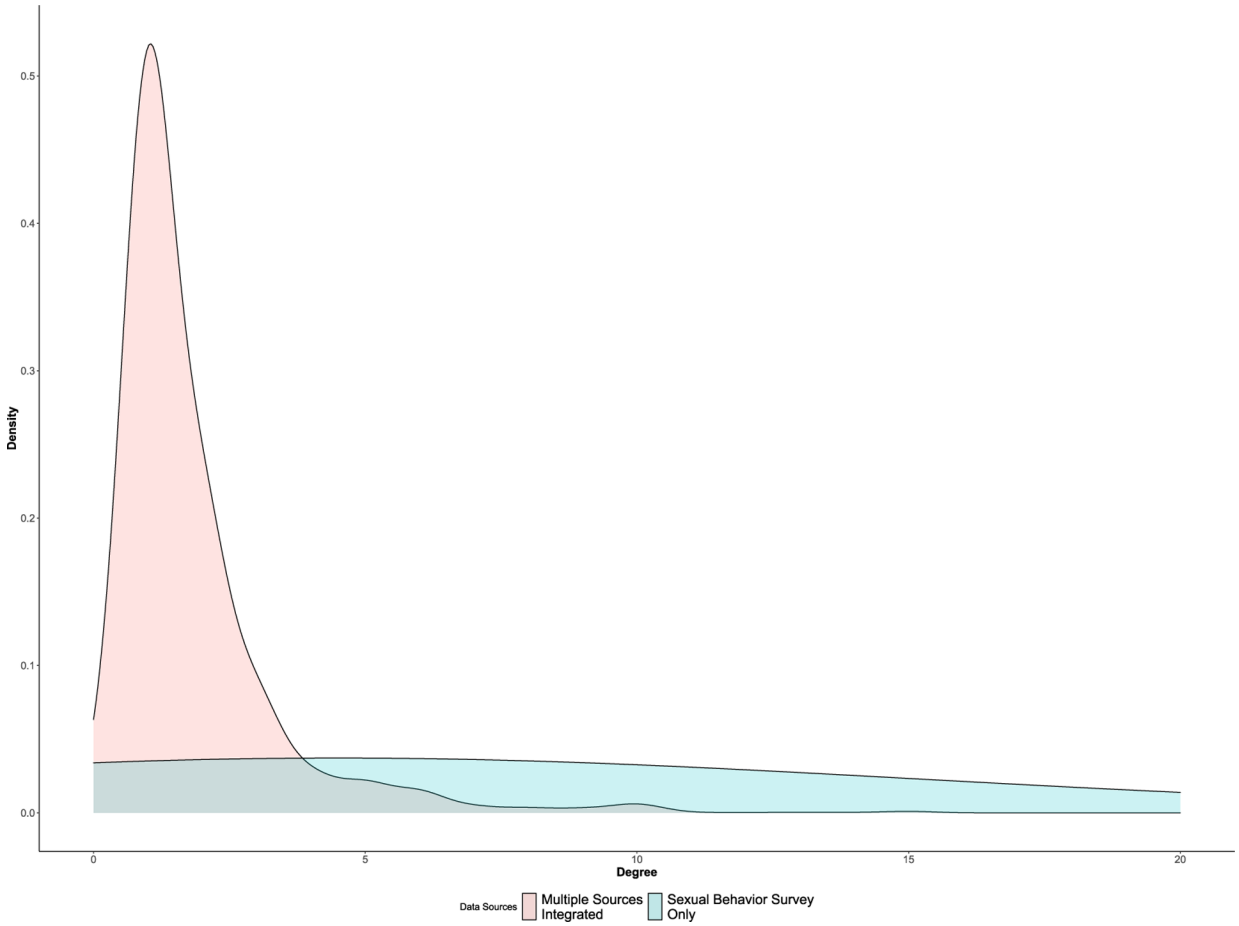

Figure 6:

Density plots for the number of individuals of each degree. Values from the survey are shown in blue, while estimates from the Bayesian approach are shown in red. Only degrees 0 to 20 are shown, as few or no individuals have degrees higher than 20.

Table 1:

Illustration of CCMs modeling the number of edges. Columns 1 and 2 list the possible values for the number of edges and the number of networks with each value (i.e., congruence class size). Columns 3 and 4 show $P_{\phi_1}(x \mid \theta)$ and $P_{\mathcal{G}_n}(g \mid \theta)$ for a CCM defined by a uniform distribution on congruence classes. Columns 5 and 6 show the information for a binomial CCM; Columns 7 and 8 for a bi-modal CCM.

Number of edges (x)	$\lvert c_{\phi_1}(x)\rvert$	Uniform CCM		Binomial CCM		Bi-modal CCM	
		$P_{\phi_1}(x \mid \theta)$	$P_{\mathcal{G}_n}(g \mid \theta)$	$P_{\phi_1}(x \mid \theta)$	$P_{\mathcal{G}_n}(g \mid \theta)$	$P_{\phi_1}(x \mid \theta)$	$P_{\mathcal{G}_n}(g \mid \theta)$
0	1	$\frac{1}{7}$	$\frac{1}{7 \cdot 1}$	0.0156	$\frac{0.0156}{1}$	0.1	$\frac{0.1}{1}$
1	6	$\frac{1}{7}$	$\frac{1}{7 \cdot 6}$	0.0938	$\frac{0.0938}{6}$	0.25	$\frac{0.25}{6}$
2	15	$\frac{1}{7}$	$\frac{1}{7 \cdot 15}$	0.2344	$\frac{0.2344}{15}$	0.1	$\frac{0.1}{15}$
3	20	$\frac{1}{7}$	$\frac{1}{7 \cdot 20}$	0.3125	$\frac{0.3125}{20}$	0.1	$\frac{0.1}{20}$
4	15	$\frac{1}{7}$	$\frac{1}{7 \cdot 15}$	0.2344	$\frac{0.2344}{15}$	0.1	$\frac{0.1}{15}$
5	6	$\frac{1}{7}$	$\frac{1}{7 \cdot 6}$	0.0938	$\frac{0.0938}{6}$	0.25	$\frac{0.25}{6}$
6	1	$\frac{1}{7}$	$\frac{1}{7 \cdot 1}$	0.0156	$\frac{0.0156}{1}$	0.1	$\frac{0.1}{1}$

Table 2:

Illustrations of CCMs modeling degree distribution. The first column provides an index for the congruence classes defined by degree distribution. Columns 2 and 3 list the distinct values for degree distributions and the number of networks with each degree distribution (i.e., congruence class size). Columns 4 and 5 show $P_{\phi_2}(x \mid \theta)$ and $P_{\mathcal{G}_n}(g \mid \theta)$ for a CCM defined by a uniform distribution on congruence classes. Columns 6 and 7 show these values for a multinomial CCM.

Index	Degree distribution $(\boldsymbol{x})^*$	$\lvert c_{\phi_2}(\boldsymbol{x})\rvert$	Uniform CCM		Multinomial CCM	
			$P_{\phi_2}(\boldsymbol{x} \mid \boldsymbol{\theta})$	$P_{\mathcal{G}_n}(\boldsymbol{g} \mid \boldsymbol{\theta})$	$P_{\phi_2}(\boldsymbol{x} \mid \boldsymbol{\theta})$	$P_{\mathcal{G}_n}(\boldsymbol{g} \mid \boldsymbol{\theta})$
1	[4, 0, 0, 0]	1	$\frac{1}{11}$	$\frac{1}{11 \cdot 1}$	0.0014	$\frac{0.0014}{1}$
2	[2, 2, 0, 0]	6	$\frac{1}{11}$	$\frac{1}{11 \cdot 6}$	0.0459	$\frac{0.0459}{6}$
3	[1, 2, 1, 0]	12	$\frac{1}{11}$	$\frac{1}{11 \cdot 12}$	0.2144	$\frac{0.2144}{12}$
4	[0, 4, 0, 0]	3	$\frac{1}{11}$	$\frac{1}{11 \cdot 3}$	0.0417	$\frac{0.0417}{3}$
5	[1, 0, 3, 0]	4	$\frac{1}{11}$	$\frac{1}{11 \cdot 4}$	0.0715	$\frac{0.0715}{4}$
6	[0, 2, 2, 0]	12	$\frac{1}{11}$	$\frac{1}{11 \cdot 12}$	0.2501	$\frac{0.2501}{12}$
7	[0, 3, 0, 1]	4	$\frac{1}{11}$	$\frac{1}{11 \cdot 4}$	0.0715	$\frac{0.0715}{4}$
8	[0, 1, 2, 1]	12	$\frac{1}{11}$	$\frac{1}{11 \cdot 12}$	0.2144	$\frac{0.2144}{12}$
9	[0, 0, 4, 0]	3	$\frac{1}{11}$	$\frac{1}{11 \cdot 3}$	0.0417	$\frac{0.0417}{3}$
10	[0, 0, 2, 2]	6	$\frac{1}{11}$	$\frac{1}{11 \cdot 6}$	0.0459	$\frac{0.0459}{6}$
11	[0, 0, 0, 4]	1	$\frac{1}{11}$	$\frac{1}{11 \cdot 1}$	0.0014	$\frac{0.0014}{1}$

*The degree distribution is denoted as x for consistency of notation across CCM examples. The distribution has the same interpretation as describe in Section 2.2, i.e., it is a vector such that the j^{th} entry represents the number of vertices having degree $j - 1$

Table 3:

The number of simulations stratified by λ and the number of individuals sampled for the behavior survey. The third column shows the total number of simulations for each value of λ and sample size. The next two columns present the number of simulations that we classify with an epidemic (i.e., infected individuals ≥ 50) and the number of simulations in the analysis.

Mean Degree (λ)	Survey Samples	Number of Simulations	Number of Simulations with Epidemic	Number of Simulations in Analysis
10	25	250	123	41
	50	250	124	72
	100	250	112	71
	200	250	105	76
20	25	250	230	35
	50	250	239	59
	100	250	230	85
	200	250	237	119
30	25	250	245	12
	50	250	245	34
	100	250	250	58
	200	250	246	83
40	25	250	247	10
	50	250	248	29
	100	250	248	60
	200	250	249	107
50	25	250	249	4
	50	250	249	40
	100	250	250	66
	200	250	250	128

Table 4:

The number of simulations stratified by the measurement error percentage and the number of behavior survey respondents. The third column shows the total number of simulations. The next two columns present the number of simulations that we classify with an epidemic (i.e., infected individuals ≥ 50) and the number of simulations included in the analysis.

Measurement Error Percentage	Survey Samples	Number of Simulations	Number of Simulations with Epidemic	Number of Simulations in Analysis
−50%	25	250	37	6
	50	250	39	7
	100	250	40	13
	200	250	43	25
−40%	25	250	42	9
	50	250	38	18
	100	250	37	19
	200	250	38	27
−30%	25	250	40	11
	50	250	41	16
	100	250	41	27
	200	250	40	32
−20%	25	250	39	10
	50	250	42	20
	100	250	40	26
	200	250	40	32
−10%	25	250	39	12
	50	250	38	30
	100	250	37	24
	200	250	41	33
0%	25	250	40	12
	50	250	46	20
	100	250	39	23
	200	250	39	34
+10%	25	250	35	9
	50	250	38	24
	100	250	39	29
	200	250	41	32
+20%	25	250	44	14
	50	250	40	23
	100	250	41	29
	200	250	42	33
+30%	25	250	41	12
	50	250	39	21
	100	250	43	32

Measurement Error Percentage	Survey Samples	Number of Simulations	Number of Simulations with Epidemic	Number of Simulations in Analysis
	200	250	42	34
+40%	25	250	44	13
	50	250	36	23
	100	250	43	35
	200	250	41	33
+50%	25	250	34	11
	50	250	40	21
	100	250	43	30
	200	250	42	38