

Approximate Inference for Longitudinal Mechanistic HIV Contact Networks

Octavious Smiley^{1*}, Till Hoffmann¹ and Jukka-Pekka Onnela¹

^{1*}Biostatistics, Harvard University, 677 Huntington Ave, Boston, 02115,
MA, USA.

*Corresponding author(s). E-mail(s): octavioussmiley@gmail.com;
Contributing authors: thoffmann@hsph.harvard.edu ;
onnela@hsph.harvard.edu ;

Abstract

Network models are increasingly used to study infectious disease spread. Exponential Random Graph models have a history in this area, with scalable inference methods now available. An alternative approach uses mechanistic network models. Mechanistic network models directly capture individual behaviors, making them suitable for studying sexually transmitted diseases. Combining mechanistic models with Approximate Bayesian Computation allows flexible modeling using domain-specific interaction rules among agents, avoiding network model oversimplifications. These models are ideal for longitudinal settings as they explicitly incorporate network evolution over time. We implemented a discrete-time version of a previously published continuous-time model of evolving contact networks for men who have sex with men (MSM) and proposed an ABC-based

approximate inference scheme for it. As expected, we found that a two-wave longitudinal study design improves the accuracy of inference compared to a cross-sectional design. However, the gains in precision in collecting data twice, up to 18%, depend on the spacing of the two waves and are sensitive to the choice of summary statistics. In addition to methodological developments, our results inform the design of future longitudinal network studies in sexually transmitted diseases, specifically in terms of what data to collect from participants and when to do so.

Keywords: mechanistic model, networks, HIV, ABC, inference, MSM, agent based modeling

1 Introduction

Networks are used to study a range of systems with interactions or dependencies among their agents, such as the behavior in supply chains and the stock market [1], protein-protein interactions in biological systems [2], and disease transmission on local and global scales [3]. In the study of disease transmission dynamics, the contact structure of a population can be naturally represented as a network, and this representation is especially useful if the contacts persist over time, as is often the case for sexual interactions. Disease dynamics are then driven by interactions (represented by edges) among susceptible and infectious individuals (represented by nodes). More generally, many of these systems arise from stochastic processes forming or dissolving interactions over time that must be accounted for when doing inference.

There are (at least) two main paradigms of networks models: statistical and mechanistic. Statistical network models prioritize tractable likelihoods to facilitate inference at the expense of model flexibility. For example, the Erdős–Rényi graph, also known

as the Bernoulli random graph, assumes that each node pair is connected independently and with identical probability. Hence, the likelihood of the number of edges is the standard binomial likelihood with a fixed number of nodes and inference readily follows because a graph is completely identified by its node and edge sets. It also follows that Erdős–Rényi has a binomial (approximately Poisson) degree distribution.

This generative mechanism however clearly does not map well to most real-world networks. This is easily seen in the World Wide Web (WWW). In this scenario, each website is represented by a node and a directed connection (hyperlink) between websites occurs when one links to the other. Unlike Erdős–Rényi networks, where the degree distribution follows a binomial distribution, the degree distribution here follows a power-law where more successful websites tend to grow their connections faster than others [4]. Exponential Random Graph Models (ERGMs) are generalizations of the Erdős–Rényi model. They represent a probability distribution of graphs on a fixed node set, where the probability of observing a graph is dependent on the presence of the various configurations specified by the model [5]. A typical graph in this distribution can be interpreted as the aggregate of the local configurations, and slight errors in estimating the local configuration counts can alter beliefs about the distribution [6].

Mechanistic models assume that the observed network is generated by a small set of mechanistic rules. The canonical example is the Barabási-Albert (BA) model. Nodes are added one by one to a growing network and each node connects to m previously existing nodes with probability proportional to the nodes' current degree [7]. This so-called preferential attachment mechanism readily generates power-law degree distributions, which are a type of broad-tailed degree distribution that are characteristic of many empirical networks, including that of the WWW. Apart from the target number of nodes, n , the classic BA model only has one free parameter, m . In this case, the fully grown graph has approximately nm edges, and m can be inferred by dividing the number of edges by the number of nodes n , i.e., m is approximately equal to

the average degree of the network. However, even for moderately complex models, the likelihood of the full network becomes intractable due to the fact that the insertion order of the nodes is (usually) not known. Because the graph is sequentially dependent on the previous iteration as it grows, the number of possible graph realizations grows exponentially with the number of added nodes.

Networks have provided insights to major public health problems such as the spread of HIV [8], the opioid crisis [9], and interventions with people who inject drugs [10]. Wertheim et al. noted that HIV is an evolving disease and constructed a disease transmission network using gene sequencing by tracking the evolutionary path of the virus and inferring edges by measuring the similarity of the virus within different individuals. Using an inferred transmission network, they developed a network statistic that was able to detect community level effects of HIV in a clinical trial setting that could help thwart future infections [8]. Aroke et al. showed the benefit of peer influence and concluded that individuals who have a diagnosis of opioid use disorder or use many prescribers may help promote positive health behaviors in an opioid prescription network due to the influence of their direct peers on the network structure. They came to this conclusion by showing the type of opioid that an individual uses and their number of prescribers were identified as significant predictors of high betweenness centrality giving them influence over the network at large [9]. Rolls et al. model network data involving people who inject drugs, using validation techniques, so that these networks can be simulated and intervention strategies could be explored [10].

There are several mechanistic models for studying the impact of men who have sex with men (MSM) contact networks and their impact on HIV transmission [11–13]. Birkett et al. used a data-driven simulation model to understand the impact of network-level mechanisms and STI infections on the spread of HIV among Young Men who have Sex with Men (YMSM) [11]. Mei et al. introduced the concept of a Complex

Agent Network (CAN) to model the HIV epidemics by combining agent-based modelling and complex networks [12]. An especially interesting model was introduced by Hansson et al. to study the role of casual contacts on the HIV epidemic in Stockholm, Sweden. Their research was used to recommend interventions to reduce transmission rates [13]. Padeniya et al. notes Hansson and others' contribution to intervention strategies as they sought to mathematically model the role of female-sex-worker-client interactions for gonorrhoea transmission [14]. Vajdi et al. noted Hansson's choice to model instantaneous casual relationships, and investigated a dynamic model for casual relationships, a two-layer temporal network model, and SIS mean-field equations [15]. A common approach for inference in these papers is to propose mechanisms for contact formation, simulate the spread of disease on the network, and modify parameter values to match disease prevalence to that observed in their respective populations without directly validating their mechanism.

Most scientific studies involving human subjects can be divided into cross-sectional and longitudinal. In cross-sectional studies, measurements are obtained at only a single point in time. The distinguishing feature of longitudinal studies is that the study participants are measured repeatedly (at least twice) throughout the duration of the study, thereby permitting the direct assessment of changes in the response variable over time [16]. To illustrate, participants in a cross-sectional study likely vary in age; however, this type of design cannot be used to study the effect of aging because the effect of aging is potentially confounded with cohort effects. It is important to note here that although we are sampling the evolving network at multiple time points, we are only asking participant information that can maintain privacy.

One example of a longitudinal network study is the work by Birkett et al. [11]. The authors studied the impact of network-level mechanisms and STI infections on the spread of HIV and found that network-level mechanisms and STI infections play a significant role in the spread of HIV and in racial disparities among (YMSM). Their

work shows HIV prevention efforts should target YMSM across race, and interventions focusing on YMSM partnerships with older MSM might be highly effective. In general, one would expect observing a network multiple times to provide more information, and therefore improve accuracy of inference, compared to observing the network just once. In addition, one may address questions that can only be interrogated in a longitudinal study. When growing a network in a simulation, we can track every iteration of the dynamic network and have arbitrarily many observations at our disposal. In an actual study, one is of course constrained by resources and logistics. If the data are obtained from self-administered or staff-administered surveys, too frequent reporting may lead to participant burden and reduce his or her willingness to continue participation, whereas too infrequent reporting may lead to recall bias and participants may be lost to follow up. For example, a person may not remember each individual whom they dated over a five-year period and may not be able to reliably recall the timing of the relationships. Collecting data at different time points that are optimally spaced helps alleviate recall bias while still maintaining an avenue for accurate inference.

Our goal in this paper is to implement a discrete-time version of the mechanistic network model introduced by Hansson et al. and use the model to identify optimal spacing between two data collection points (waves) in a longitudinal network study such that we can achieve the dual goal of accurate inference (learning model parameters as precisely as possible) while minimizing participant burden (using network features that in practice could be elicited from participants with a minimal number of survey questions). These results have implications for study design for HIV and other sexually transmitted diseases, and more broadly they can inform other research questions involving (longitudinal) network data.

This paper is structured as follows. We discuss the discrete-time mechanistic network model in Section 2.1 and explain our ABC-based approach to approximate parameter inference in Section 2.2. We show our results in Section 3 and conclude with a discussion in Section 4.

2 Methods

2.1 Mechanistic network model

As noted in the Introduction, there are several mechanistic models for MSM contact formation in specific populations. We focus on the mechanistic model introduced to study MSM contact networks in Stockholm, Sweden [13]. The model incorporates specific behaviors that guide the formation and dissolution of sexual contacts as well as migration of individuals in and out of the population. While the original model was formulated in continuous time, we consider a discrete time version of the model. This means that rates in the original formulation correspond to probabilities in ours. We note that as the number of the potential discrete time events tends to infinity and the event probabilities tend to zero, our formulation of the model converges to the original. Throughout this paper, each discrete model time step iteration is taken to correspond to one calendar month, and all events are recorded at the end of each iteration. While a constant number of individuals enter the population at each iteration, each individual leaves the population with a fixed probability at each iteration. The size of the network therefore fluctuates around n nodes, where n is the initial number of nodes in the network.

The model incorporates two types of partnerships: steady and casual. Casual relationships are defined to only last one iteration at onset while steady relationships are defined to have the potential to last longer. An individual can have at most one steady partner at any given time. The probability that a single person enters a steady relationship at a given iteration is ρP_0 , where P_0 is the proportion of single individuals in the present iteration. In the original model, where ρ is a rate of steady partnership formation, P_0 fluctuates around an equilibrium; in our version, we fix this parameter and absorb it into ρ for simplicity and to improve identification of model parameters. Our modified probability of a single person entering a steady relationship at a given iteration is therefore ρ . While the number of people willing to form relationships varies

from iteration to iteration, the probability a single person joining a relationship stays the same. In the Hansson paper, the differential equation formulation of the model explicitly considers the fluctuation of the likelihood of new relationships while we do it implicitly as the number of singles changes. The probability of said steady relationship dissolving at each iteration is σ .

In addition to a steady relationship, an individual may also have one casual partnership at each iteration. These casual relationships may occur alongside steady partnerships or during times when the person is single. A single individual enters a casual relationship with probability ω_0 while an individual who is currently in a steady relationship forms a casual relationship with probability ω_1 . For any partnership to form, both individuals must be willing to join that relationship. In the scenario where an odd number of individuals would like to form a relationship, one of them (chosen at random) is left out. Each person migrates from the population with probability μ , and individuals enter into the population at constant rate $n\mu$. The migration of an individual and the formation and dissolution of a sexual contacts are all determined by the outcome of independent Bernoulli trials. In the original continuous time formulation of the model, duration of steady relationships and the time spent in the population both follow exponential distributions. In contrast, for our discrete time formulation both are geometrically distributed. Starting from an empty graph with n nodes, we first run the model until we are confident that it has converged to the target distribution. We set our migration probability to 0 to ensure we are sampling individuals longitudinally and to maintain a closed cohort design. We note that the 'constant' number of nodes being added is largely dependent on only μ and n and easily recoverable. The model is described in Algorithm 1 and a few graph realizations from the model are illustrated in Figure 1. We chose 1000 iterations to ensure we are past the burn-in [13].

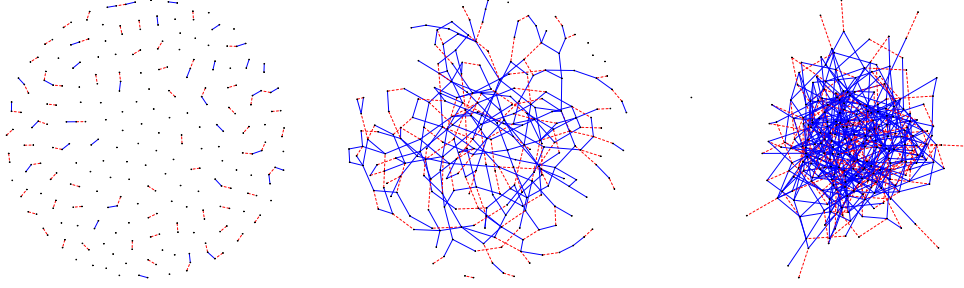


Fig. 1 Network visualizations containing cumulative (from iteration 1) steady (red dashed) and casual (blue solid) edges for iterations 1 (left), 6 (middle), and 12 (right). We used the following parameter values: $\mu = 0$, $\rho = 0.3$, $\sigma = 0.1$, $w_1 = 0.2$, $w_0 = 0.4$.

Algorithm 1 Hansson MSM model [13]

```

1: Inputs:
2:  $n :=$  number of nodes
3:  $G := (V, E)$  the graph has no edges
4:  $\rho :=$  scaling parameter for partnership formation probability
5:  $\sigma :=$  separation probability for steady relationships
6:  $\omega_0 :=$  the probability someone who is single enters into a casual relationship
7:  $\omega_1 :=$  the probability someone who is in a partnership enters into a casual relationship
8:  $iterations :=$  the number of iterations to run algorithm
9: Algorithm:
10: for  $i$  in  $1:iterations$  do
11:   Dissolve all casual relationships generated from the previous iteration if applicable
12:   Identify all current steady relationships
13:   if there are a positive number of steady relationships then
14:     Dissolve each with probability  $\sigma$ 
15:   end if
16:   Identify all nodes with degree 0
17:   Set willingness to form a steady relationship with probability  $\rho$ 
18:   Randomly match the maximum number of even nodes that are willing to form a steady relationship into edges
19:   Identify single nodes  $:=$  nodes with degree 0
20:   Set willingness to form a casual relationship with probability  $\omega_0$ 
21:   Identify steady nodes  $:=$  nodes with degree 1
22:   Set willingness to form a casual relationship with probability  $\omega_1$ 
23:   Randomly match the maximum number of even nodes among the single and steady nodes together that are willing to form a casual relationship
24:   Record Network
25: end for

```

2.2 Inference of model parameters

In Bayesian inference, complete knowledge of the model parameters, given the observed data, is contained in the posterior distribution. Typically, in mechanistic models, the complexity of the model means that the likelihood and corresponding posterior distribution is not available in closed form. In mechanistic models one can nevertheless forward simulate data from the model given parameters, and these parameter values may be obtained from a prior distribution. ABC is an inference framework that has been developed to deal with models that have intractable likelihoods. There are several ABC methods to generate samples from an approximate posterior distribution. For clarity of our objective, we use the simple accept/reject algorithm. In accept/reject, we propose parameter values from a prior distribution to generate data and retain parameter values that produce data that resembles the observed data [17]. If we only kept parameter values that reproduced the observed data exactly, this approach would recover the exact posterior for discrete data. This approach is however impossible for continuous data because the probability of sampling a continuous value exactly is 0. To ensure that our prior distribution’s support is realistic to MSM relationship characteristics, we utilize a uniform distribution on the duration of average time spent for people to be open to joining a relationship $\frac{1}{\rho}$ [1 month, 50 months], average time a steady relationship lasts $\frac{1}{\sigma}$ [1 month, 90 months], average time for a single individual to partake in a casual relationship $\frac{1}{\omega_0}$ [1 month, 40 months], and average time for an individual in a relationship to partake in a casual relationship $\frac{1}{\omega_1}$ [1 month, 61 months]. Figure 2 shows the prior distributions on these inverse parameters and the corresponding implied prior distributions on the parameters themselves. We recognize a variety of definitions for steady and casual relationships in MSM contact networks, as well as a variety of estimates for the support of each duration [18–25]. We chose our support to be consistent with the data the model was originally trained on [13],

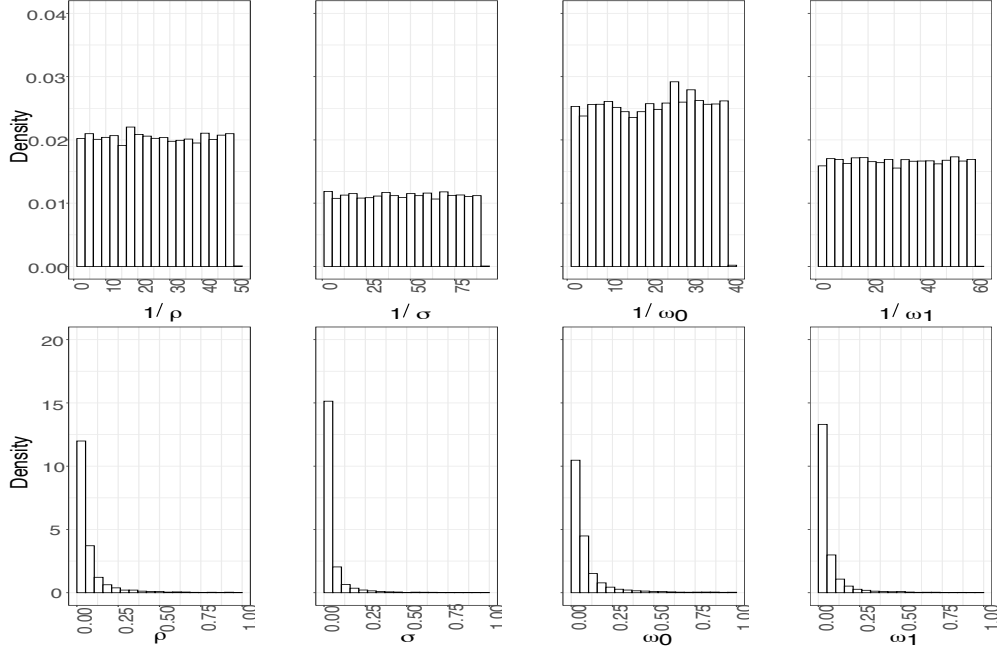


Fig. 2 Prior distributions on the inverse parameters and the corresponding implied prior distributions on parameters themselves. The top row shows the distributions of the inverse parameters, which can be interpreted as distributions of the average values of geometric distributions. The bottom row shows the distributions of the parameter values themselves for our discrete time mechanistic network model for the following parameters: ρ , σ , ω_1 , ω_0 .

and calculate the reciprocal of each parameter sampled from the prior as an input to our model.

There are at least three major considerations in the ABC accept/reject framework: summary statistics, distance measure, and similarity threshold [26]. Given the mechanistic network model of interest, we manually chose a set of summary statistics needed for inference, which renders the network space more manageable [26]. The choice of network summary statistics was guided by the principle that it should be possible to obtain this information from study participants using a questionnaire and they should be informative of the model parameters. At a minimum, one needs at least as many summary statistics as there are parameters to be inferred [26]. Although the model has five parameters (six if one counts n), as previously mentioned, we opted to fix

one of them, the migration probability $\mu = 0$. This leaves us with four parameters to recover: probability of a single person entering a steady relationship ρ , probability of dissolving a steady relationship σ , probability of a single individual to enter a casual relationship ω_0 , and probability of an individual in steady relationship to enter a casual relationship ω_1 . More information on parameters can be found in Table 1. We chose the four summaries, denoted s_1 through s_4 , as listed in Table 2. We chose summaries that could be elicited by asking participants to consider their sexual history in the past year only. Longer histories could potentially be more informative, but longer look-back periods would likely increase recall bias.

Since the mechanistic network model is outside the exponential family, we have no guarantee of sufficiency, i.e., that our summary statistics fully summarize our network. However, we still require the summary statistics to be informative of the model parameters. An informal way to assess the extent of informativeness is to investigate plots of network summary statistics against model parameters. We denote these relationships as $s_i(\theta)$ where $i \in \{1, 2, 3, 4\}$ and $\theta \in \{\rho, \sigma, w_0, w_1\}$. We refer to these relationships as *mapping functions*, and we estimate them using simulations where $s_{i,k}(\theta)$ represents the value of summary statistic s_i with respect to generative parameter θ in simulation run k . The value of $s_i(\theta)$ is given as the median value of $s_{i,k}(\theta)$ taken across all simulations k .

We measured the distance between a simulated network and the observed network by calculating the Euclidean distance within the normalized summary statistic space. The normalized summary statistic value is obtained by first subtracting the mean of the summary statistic from each value and then dividing each value by the standard deviation of the summary statistic. We populated an ABC reference table for each lag by generating 10,000 graphs by sampling parameters from their joint priors and varying the lag between the two network observations between zero and 150 iterations. Next, taking a sample per lag from our joint prior density and its corresponding

graph as our ground truth, we simulated samples from the corresponding approximate posterior distribution. We retained the parameters associated with the 100 (top 1%) smallest distances in the normalized summary statistic space between the observed and generated graphs.

Finally, we performed a regression adjustment on samples from the approximate posterior distribution [27, 28]. The goal of the regression adjustment is to improve our ABC posterior’s convergence to our target posterior. The basis of the method is that we can obtain an estimate of our expected parameter values given the summaries using linear regression in the localized neighborhood around our observed data that we get from the approximate posterior. Then, we can use this relationship to adjust our approximate posterior distribution [27]. We normalized the parameters in our reference table, and utilized the root mean squared error (RMSE) of the approximated posteriors for a fixed set of 500 ground truth parameters to measure accuracy of inference. Then, we averaged over all 500 parameter sets for an estimate of the RMSE, for a given lag, over our prior space [29]. For clarity, consider θ_i as the i th ground truth parameter and $\hat{\theta}_{i,k}$ as the k th sample from the approximate posterior estimating θ_i . Our estimate of RMSE is then

$$RMSE = \frac{1}{500} \sum_{i=1}^{500} \sqrt{\frac{1}{100} \sum_{k=1}^{100} (\theta_i - \hat{\theta}_{i,k})^2}. \quad (1)$$

Finally, we fitted a locally weighted regression (loess) with a 95% confidence interval to the data. We note that while the values of the parameters in the reference table are normalized, the resulting approximate posterior distributions are not. Individually normalizing the reference table parameters is useful because it places all parameters on the same scale when calculating the RMSE. However, the posteriors are displayed on the original scale for ease of interpretation.

Parameter	Hansson et. al [13]	Current Paper
n	Average population size	Population size
μ	Rate of leaving the sexually active population	Probability of leaving the sexually active population
ρ	Partnership desire scale rate	Probability of a single person entering a steady relationship
σ	Separation rate	Probability of dissolving a steady relationship
ω_0	Rate at which an individual who is single tries to have casual sex	Probability of a single individual to enter a casual relationship
ω_1	Rate at which an individual who is in a relationship tries to have casual sex	Probability of an individual in a steady relationship to enter a casual relationship

Table 1 Parameter And Their Interpretation

Name	Description
s_1	The proportion of single individuals
s_2	The average length of steady relationships that start and end in the course of the study
s_3	The proportion of individuals in steady relationships who are also in casual relationships
s_4	The proportion of steady relationships among all relationships

Table 2 Summary Statistic Descriptions

3 Results

We evaluated the mapping functions on a grid along the unit interval by generating 100 graphs per parameter value and using box plots to summarize the results. We investigated mapping functions in two different scenarios. First, we varied each parameter in turn while keeping all others fixed at the values reported in [13], i.e., we fixed $\rho = 0.3$, $\sigma = 0.1$, $\omega_0 = 0.4$, and $\omega_1 = 0.2$, and we also set $\mu = 0$ to keep the cohort closed. Second, we sampled each free parameter from its respective prior distribution. These plots were used to ensure that the chosen summary statistics are informative about the model parameters as can be seen in Figures 3 and 4.

While more summaries could be included, that would increase the computational burden and likely would not significantly increase accuracy. We considered several extra summaries during discovery. Since we would like to obtain the data from questionnaires, one also needs to consider participant burden: all else equal, we would like to ask as few questions as needed to address the scientific question at hand. It

is also worth emphasizing that each of the listed summaries can be obtained using privacy preserving questions only in data collection, i.e., participants do not need to disclose their identity nor the identity of their steady or casual partners. This arguably improves the quality of the collected data as respondents would be expected to be more likely to report their behavior accurately. We note that while a regression adjustment generally improves the results, it can at times generate functionally impossible values, such as negative probabilities, or worsen our inference when the summaries do not accurately represent the network. In the rare occasion the regression adjustment proposes a negative number, we opt to take a conservative approach and set the value at 0. In this study, we did not see any adjusted proposal probabilities above 1.

We visualized the regression adjusted approximate posteriors when looking at the graph once or twice with a lag of 50 iterations in Figure 5. Furthermore, as expected, and as shown in Figure 6, observing a network twice results in a smaller average error compared to observing a network only once. The improvement is largely driven by our ability to recover σ and ρ parameters as shown in Figure 7. We also see the average error steadily decreases with the lag between the two network observations until about 40 to 50 iterations. This lag between the two network observations (data collection waves) is optimal in the sense that extending the gap further does not greatly increase accuracy of inference but does lengthen the duration of the study. In a closed cohort study, all else equal, the longer the duration of the study, the greater the expected attrition of study participants. Attrition of study participants in a setting like ours would lead to incomplete ascertainment of network structure and therefore introduce an additional error to network summary statistics. We also see that implementing a regression adjustment does reduce our average error by nearly an additional 2.6%, while maintaining the overall trend and optimal lag. Finally, we note our overall ability to discern parameters from our joint prior distribution when collecting data twice after

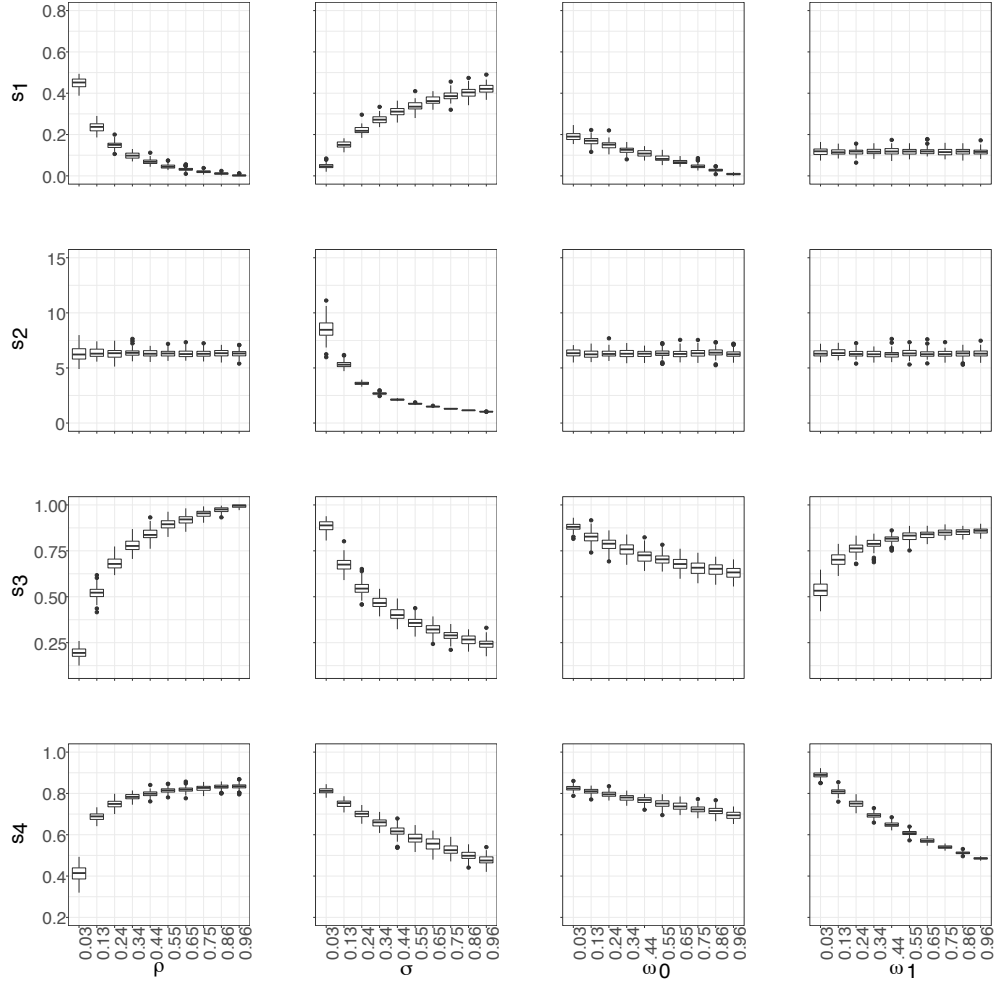


Fig. 3 Pairwise relationships between the model parameters (horizontal axes) and the summary statistics (vertical axes) used in our ABC inference scheme. Free parameters are fixed at $\mu = 0$, $\rho = 0.3$, $\sigma = 0.1$, $\omega_0 = 0.4$, $\omega_1 = 0.2$. The lag between two consecutive network observations is fixed at 15 iterations. Each box plot consists of 100 samples.

a regression adjustment with an optimal drop of roughly 62% from our average prior error and 18% when only collecting data once.

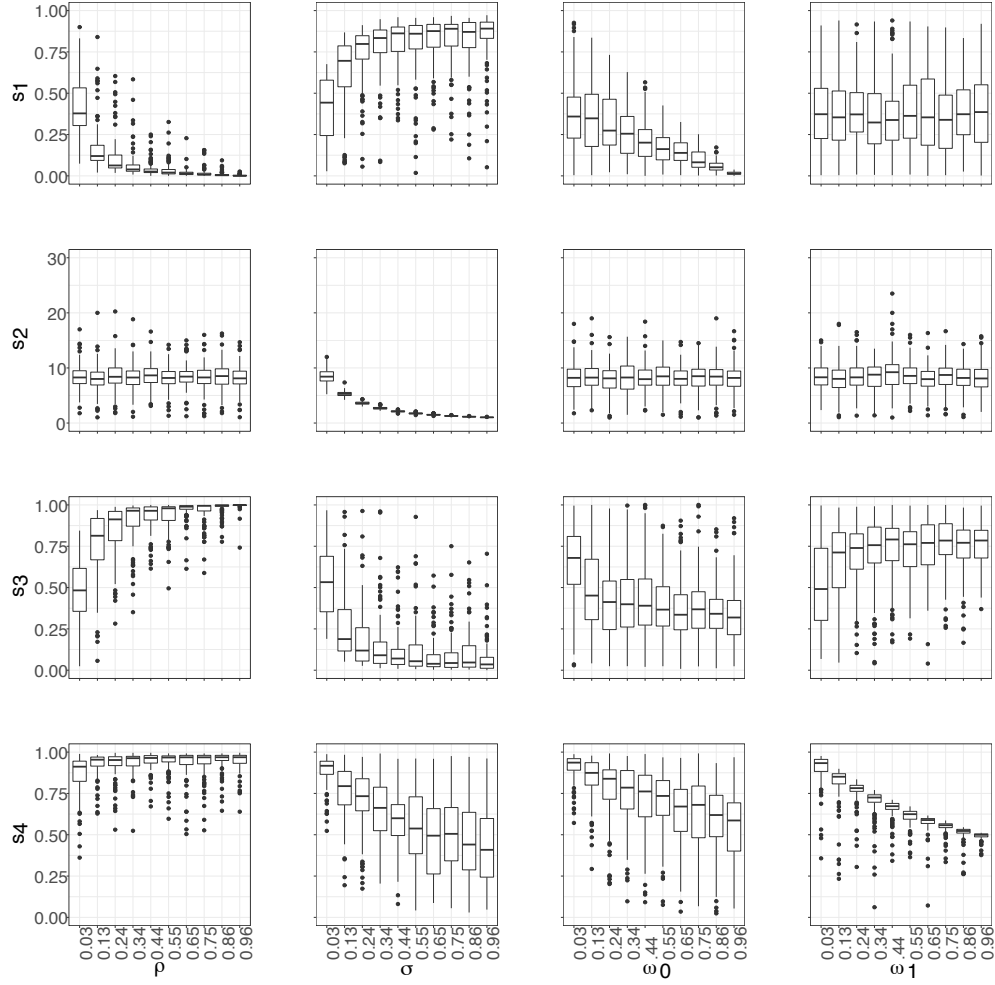


Fig. 4 Pairwise relationships between the model parameters (horizontal axes) and the summary statistics (vertical axes) used in our ABC inference scheme. Free parameters are sampled from the prior distributions. The lag between two consecutive network observations is fixed at 15 iterations. Each box plot consists of 100 samples.

4 Discussion

In this paper, we investigated the accuracy of an approximate inference scheme applied to an evolving mechanistic network model in a setting where the network, representing sexual contacts among people in a closed population, is observed at two different

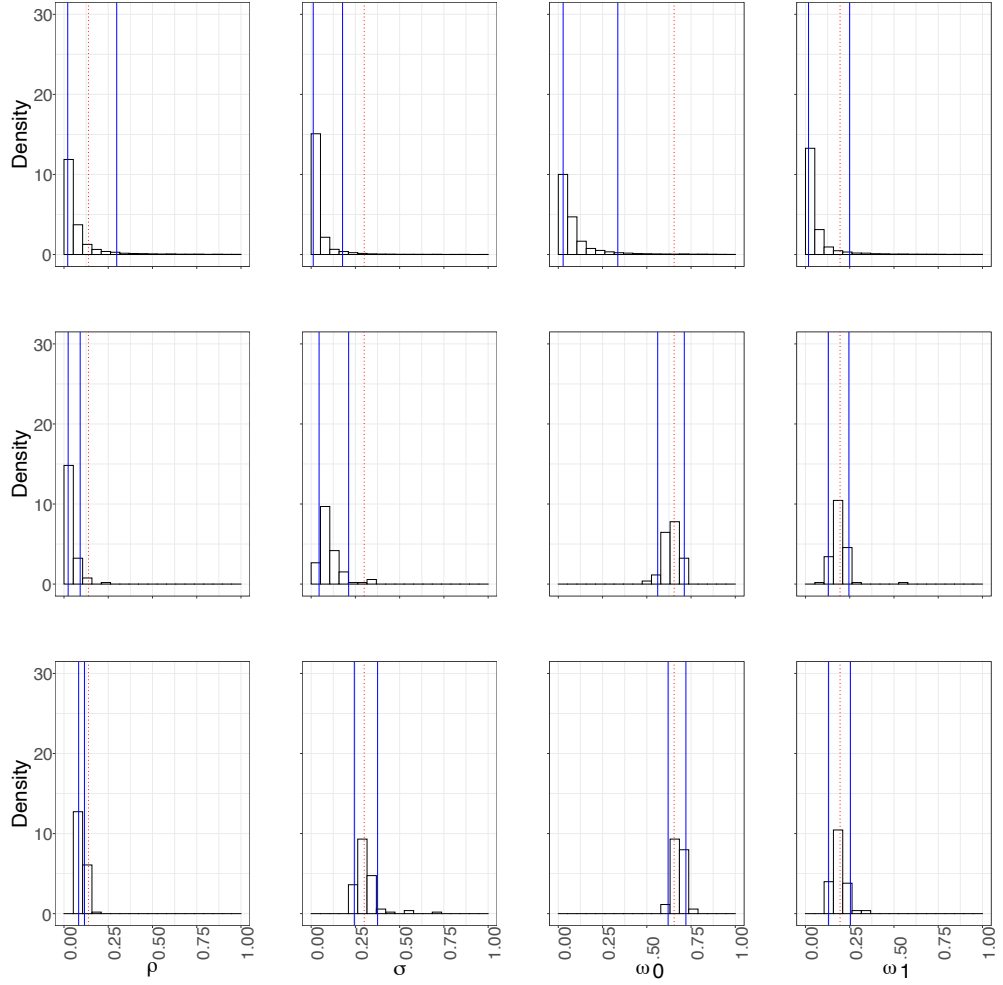


Fig. 5 Approximate marginal posterior distributions of model parameters obtained by retaining the top 1% of proposed prior samples in our ABC accept/reject inference scheme. Different rows correspond to comparing the prior (top), observing the graph once (middle), and observing the graph twice with a lag of 50 iterations (bottom). All posteriors include a regression adjustment. The blue solid lines represent the 95% credible intervals and the red dotted lines represent the true parameter values.

time points. As expected, observing the network twice improves the accuracy of inference, but this reduction in inferential error depends on the time lag between the two observations. Given that collection of real-world sexual network data is expensive and logistically challenging, it pays off to optimize the gap between the two time points to maximize accuracy of inference. If the two network observations are too close in time,

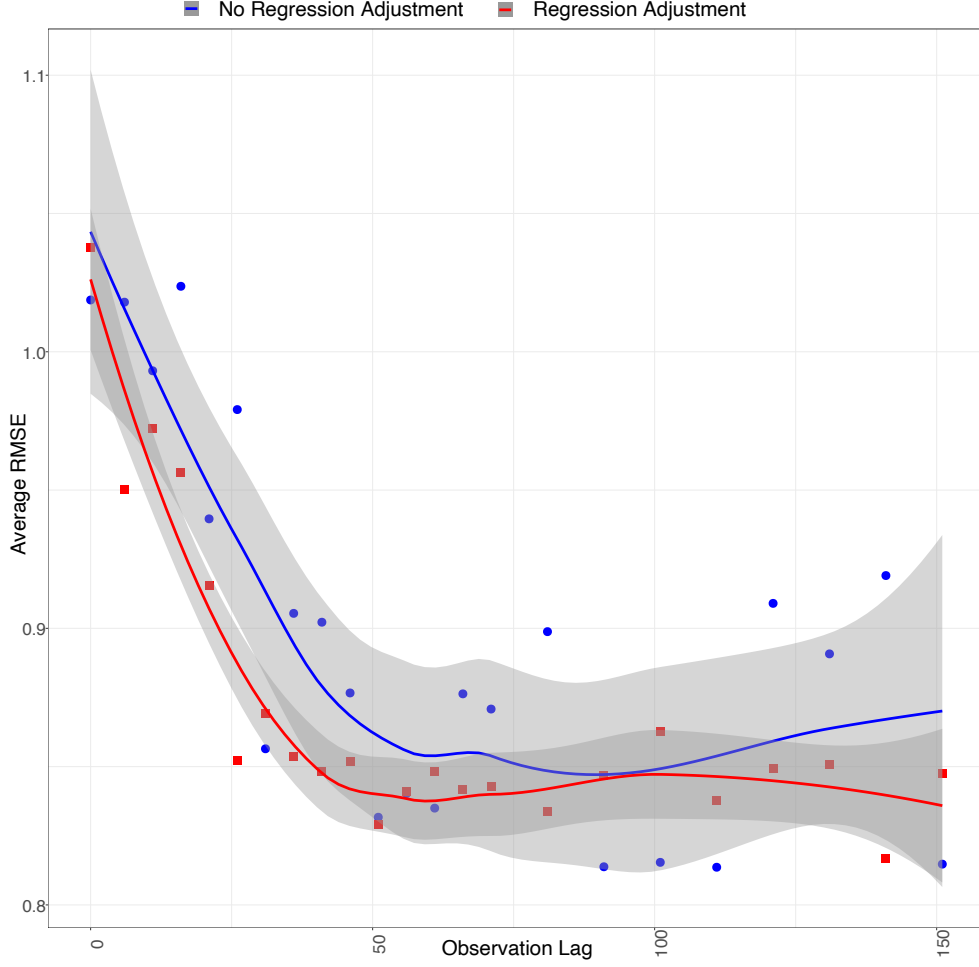


Fig. 6 Estimated average RMSE, where the average is taken across multiple network realizations, as a function of the lag between the two network observations. We also include a loess curve with a 95% confidence interval (shaded areas). The average prior average RMSE is 2.22 (not shown), whereas the corresponding regression adjusted error for a network observed only once is 1.03 that for a network observed twice with a lag of 50 iterations is 0.84.

there may have been only minimal changes in the network structure, and therefore the second observation adds little information. However, if the two network observations are too far apart in time, the study may be logistically difficult to carry out in practice and the population is likely to experience significant churn.

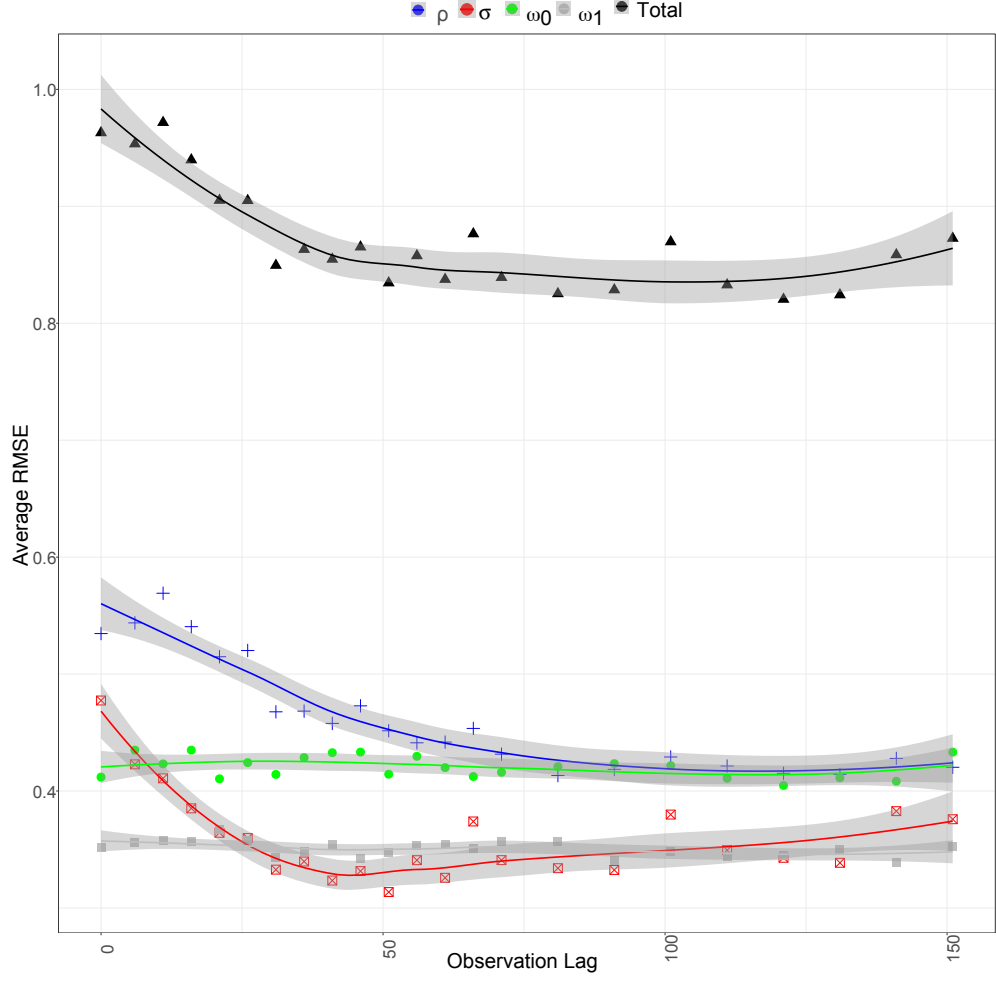


Fig. 7 Estimated regression adjusted average RMSE for the total error (top curve) and separately for the four parameters considered in our study (bottom four curves). These results show that when observing a network twice, the reduction in total RMSE is mainly due to the reduction of RMSE for ρ and σ .

There are a total of six parameters in the model, but we fixed two of them to focus on a closed, fixed-sized cohort. When considering the contribution of the remaining four parameters to inferential error, we observed that the σ (probability of dissolving a steady relationship) and ρ (probability of a single person entering a steady relationship) parameters benefited the most from the lag between the two network

observations. This finding is intuitive as these two parameters influence multiple relationship iterations. However, ω_0 (probability of a single individual to enter a casual relationship) and ω_1 (probability of an individual in steady relationship to enter a casual relationship) both correspond to one-time events and do not benefit as much from a lag. In particular, ω_1 is relatively accurate at all lags while ω_0 would likely see more relative improvement through the consideration of another summary statistic.

The set of summary statistics that may be considered in inference depends on the information obtained from subjects through study questionnaires. The informativeness of questions themselves depends on the mechanisms that drive contact formation in the study population. Depending on the mechanisms, it is possible that any set of individual-level questions (giving rise to so-called egocentric samples of the network) may be inadequate for network inference and instead one may need information about the full network structure. While this type of network-level information could be obtained using a sociocentric design, it is very challenging, and we are aware of only one study that has implemented this in practice. The Likoma Network Study was based on a sociocentric survey of sexual partnerships aimed to investigate the population-level structure of sexual networks connecting the young adult population of several villages on Likoma Island, Malawi [30]. We stress that this notable study is cross-sectional and therefore corresponds to a one-time observation of the network (even if the data collection in this study occurred in two stages for logistical reasons). Obtaining two observations of the network would be logistically nearly impossible, and doing so in larger populations is not feasible.

Our results highlight the importance of using simulation to investigate the hypothesized generative mechanisms of network formation to inform future study designs, here specifically 1) what questions to ask so that maximally informative network summary statistics may be constructed and 2) how to space the two (or possibly more) data collection waves. For example, in our setting, introducing extensive migration in

the population leads to a shorter optimal lag between the two network observations. Our approach is compatible with the recommended paradigm of using simulations for designing and interpreting intervention trials in infectious diseases, particularly with regard to emerging infectious diseases [31]. One of the main goal of such simulations is to more accurately reflect the dynamics of the transmission process. For sexually transmitted diseases, learning about the mechanisms of network formation is an important step in that direction.

In this paper, we have used basic ABC and basic regression adjustment techniques because our goal here is to see whether the ABC approach is effective in its simplest and most interpretable form. More refined variants of these methods, which can substantially improve computational performance, can be studied later on. Finally, at the time of writing, we came across related work on how design choices for egocentric network studies impact statistical estimation and inference for ERGMs [32]. This investigation is relevant for ours, although our focus is specifically on the multiple observation of the evolving network. For a suitably chosen ERGM, i.e., an ERGM with reasonably simple dependence assumptions, it is possible to attain sufficient summary statistics from egocentric network samples. This allows for exact statistical inference, but at the cost of making distributional assumptions that may not hold. For that reason, it is valuable for investigators to have various methods at their disposal so that they may choose the tool that best fits the scientific problem at hand.

5 Declarations

5.1 Availability of data and materials

Data was simulated using a mechanistic model introduced to study MSM contact networks in Stockholm, Sweden [13]. Our code is accessible at: <https://github.com/onnella-lab/longitudinal-inference>

5.2 Competing Interests

The authors have no competing interest to report.

5.3 Funding

NIH Award #R01AI138901.

5.4 Authors' contributions

All authors conceived the study as well as drafted and revised the manuscript; OS implemented the method in code and carried out data analyses. TH and JP supervised.

5.5 Acknowledgments

We would like to acknowledge John Quackenbush and Marcello Pagano for their thoughtful insights on summary statistic exploration.

References

- [1] Macal, C., Sallach, D., North, M.: Emergent structures from trust relationships in supply chains. In: Proc. Agent 2004: Conf. on Social Dynamics, pp. 7–9 (2004)
- [2] Scholtens, D., Gentleman, R.: Making sense of high-throughput protein-protein interaction data. *Statistical Applications in Genetics and Molecular Biology* **3**(1) (2005)
- [3] Le, T.-M., Raynal, L., Talbot, O., Hambridge, H., Drovandi, C., Mira, A., Mengersen, K., Onnela, J.-P.: Framework for assessing and easing global COVID-19 travel restrictions. *Scientific Reports* **12**(1), 1–13 (2022)
- [4] Adamic, L.A., Huberman, B.A.: Power-law distribution of the world wide web. *Science* **287**(5461), 2115–2115 (2000)

- [5] Robins, G., Pattison, P., Kalish, Y., Lusher, D.: An introduction to exponential random graph (p^*) models for social networks. *Social Networks* **29**(2), 173–191 (2007)
- [6] Goyal, R., Onnela, J.: Framework for converting mechanistic network models to probabilistic models. *arXiv 2001.08521* (2020)
- [7] Albert, R., Barabási, A.-L.: Statistical mechanics of complex networks. *Reviews of Modern Physics* **74**(1), 47 (2002)
- [8] Wertheim, J.O., Kosakovsky Pond, S.L., Little, S.J., De Gruttola, V.: Using HIV transmission networks to investigate community effects in HIV prevention trials. *PloS One* **6**(11), 27775 (2011)
- [9] Aroke, H., Katenka, N., Kogut, S., Buchanan, A.: Network-based analysis of prescription opioids dispensing using exponential random graph models (ERGMs). In: *Complex Networks & Their Applications X: Volume 2, Proceedings of the Tenth International Conference on Complex Networks and Their Applications COMPLEX NETWORKS 2021* 10, pp. 716–730 (2022). Springer
- [10] Rolls, D.A., Wang, P., Jenkinson, R., Pattison, P.E., Robins, G.L., Sacks-Davis, R., Daraganova, G., Hellard, M., McBryde, E.: Modelling a disease-relevant contact network of people who inject drugs. *Social Networks* **35**(4), 699–710 (2013)
- [11] Birkett, M., Armbruster, B., Mustanski, B., *et al.*: A data-driven simulation of HIV spread among young men who have sex with men: the role of age and race mixing, and STIs. *Journal of Acquired Immune Deficiency Syndromes* **70**(2), 186 (2015)

- [12] Mei, S., Sloot, P.M., Quax, R., Zhu, Y., Wang, W.: Complex agent networks explaining the HIV epidemic among homosexual men in Amsterdam. *Mathematics and Computers in Simulation* **80**(5), 1018–1030 (2010)
- [13] Hansson, D., Leung, K.Y., Britton, T., Strömdahl, S.: A dynamic network model to disentangle the roles of steady and casual partners for HIV transmission among MSM. *Epidemics* **27**, 66–76 (2019)
- [14] Padeniya, S.M.T.N.: Mathematical modelling to explore the role of the female-sex-worker-client interaction for gonorrhoea transmission and prevention among australian heterosexuals. PhD thesis, UNSW Sydney (2021)
- [15] Vajdi, A., Juher, D., Saldaña, J., Scoglio, C.: A multilayer temporal network model for STD spreading accounting for permanent and casual partners. *Scientific Reports* **10**(1), 1–12 (2020)
- [16] Fitzmaurice, G.M., Laird, N.M., Ware, J.H.: *Applied longitudinal analysis* (2012)
- [17] Csilléry, K., Blum, M.G., Gaggiotti, O.E., François, O.: Approximate bayesian computation (ABC) in practice. *Trends in Ecology & Evolution* **25**(7), 410–418 (2010)
- [18] Malone, J., Syvertsen, J.L., Johnson, B.E., Mimiaga, M.J., Mayer, K.H., Bazzi, A.R.: Negotiating sexual safety in the era of biomedical HIV prevention: relationship dynamics among male couples using pre-exposure prophylaxis. *Culture, Health & Sexuality* **20**(6), 658–672 (2018)
- [19] Down, I., Ellard, J., Bavinton, B.R., Brown, G., Prestage, G.: In Australia, most HIV infections among gay and bisexual men are attributable to sex with ‘new’partners. *AIDS and Behavior* **21**(8), 2543–2550 (2017)

- [20] Vroome, E.M., Stroebe, W., Sandfort, T.G., WIT, J.B., Griensven, G.J.: Safer sex in social context: Individualistic and relational determinants of AIDS-preventive behavior among gay men 1. *Journal of Applied Social Psychology* **30**(11), 2322–2340 (2000)
- [21] Wall, K.M., Stephenson, R., Sullivan, P.S.: Frequency of sexual activity with most recent male partner among young, internet-using men who have sex with men in the United States. *Journal of Homosexuality* **60**(10), 1520–1538 (2013)
- [22] Davidovich, E.: *Liaisons dangereuses: HIV risk behavior and prevention in steady gay relationships* (2006)
- [23] Weiss, K.M., Goodreau, S.M., Morris, M., Prasad, P., Ramaraju, R., Sanchez, T., Jenness, S.M.: Egocentric sexual networks of men who have sex with men in the United States: Results from the ARTnet study. *Epidemics* **30**, 100386 (2020)
- [24] Bavinton, B.R., Duncan, D., Grierson, J., Zablotska, I.B., Down, I.A., Grulich, A.E., Prestage, G.P.: The meaning of ‘regular partner’ in HIV research among gay and bisexual men: implications of an Australian cross-sectional survey. *AIDS and Behavior* **20**(8), 1777–1784 (2016)
- [25] Myers, T., Allman, D., Calzavara, L., Morrison, K., Marchand, R., Major, C.: *Gay and bisexual men’s sexual partnerships and variations in risk behaviour* (1999)
- [26] Sisson, S.A., Fan, Y., Beaumont, M.: *Handbook of approximate bayesian computation* (2018)
- [27] Beaumont, M.A.: Approximate bayesian computation. *Annual Review of Statistics and its Application* **6**, 379–403 (2019)
- [28] Beaumont, M.A., Zhang, W., Balding, D.J.: Approximate bayesian computation

- in population genetics. *Genetics* **162**(4), 2025–2035 (2002)
- [29] Fearnhead, P., Prangle, D.: Constructing summary statistics for approximate bayesian computation: semi-automatic approximate bayesian computation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **74**(3), 419–474 (2012)
- [30] Helleringer, S., Kohler, H.-P.: Sexual network structure and the spread of HIV in Africa: evidence from Likoma Island, Malawi. *Aids* **21**(17), 2323–2332 (2007)
- [31] Halloran, M.E., Auranen, K., Baird, S., Basta, N.E., Bellan, S.E., Brookmeyer, R., Cooper, B.S., DeGruttola, V., Hughes, J.P., Lessler, J., *et al.*: Simulations for designing and interpreting intervention trials in infectious diseases. *BMC Medicine* **15**(1), 1–8 (2017)
- [32] Krivitsky, P.N., Morris, M., Bojanowski, M.: Impact of survey design on estimation of exponential-family random graph models from egocentrically-sampled data. *Social Networks* **69**, 22–34 (2022)