Reconstructing contact network parameters from viral phylogenies

Rosemary M. McCloskey[1], Richard H. Liang[1], and Art F.Y. Poon[1,2]

[1]BC Centre for Excellence in HIV/AIDS, Vancouver, Canada

[2] Department of Medicine, University of British Columbia, Vancouver, Canada

July 14, 2016

Abstract

Models of the spread of disease in a population often make the simplifying assumption that the population is homogeneously mixed, or is divided into homogeneously mixed compartments. However, human populations have complex structures formed by social contacts, which can have a significant influence on the rate of epidemic spread. Contact network models capture this structure by explicitly representing each contact which could possibly lead to a transmission. We developed a method based on approximate Bayesian computation (ABC) for estimating structural parameters of the contact network underlying an observed viral phylogeny. The method combines adaptive sequential Monte Carlo for ABC, Gillespie simulation for propagating epidemics though networks, and a kernel-based tree similarity score. We used the method to fit the Barabási-Albert network model to simulated transmission trees, and also applied it to viral phylogenies estimated from ~~five~~ ten published HIV sequence datasets. On simulated data, we found that the preferential attachment power and the number of infected nodes in the network can often be accurately estimated. On the other hand, the mean degree of the network, as well as the total number of nodes, were not estimable with ABC. ~~We observed substantial heterogeneity in the parameter estimates on real datasets, with point estimates for the preferential attachment power ranging from 0.06 to 1.05.~~ We observed sub-linear preferential attachment (PA) power in all datasets, as well as higher PA power in networks of injection drug users. These results underscore the importance of considering contact structures when performing phylodynamic inference. Our method offers the potential to quantitatively investigate the contact network structure underlying viral epidemics.

Introduction

When an infectious disease spreads through a population, transmissions are generally more likely to occur between certain pairs of individuals. Such pairs must have a particular mode of contact with one another, which varies with the mode of transmission of the disease. For airborne pathogens, physical proximity may be sufficient, while for sexually transmitted diseases, sexual or in some cases blood-to-blood contact is required. The population together with the set of links between individuals along which transmission can occur is called the contact network [1, 2]. The structure of the contact network underlying an epidemic can profoundly impact the speed and pattern of the epidemic's expansion. Network structure can influence the prevalence trajectory [3, 4] and epidemic threshold [5], in turn affecting the estimates of quantities such as effective viral population size [6]. From a public health perspective, contact networks have been explored as tools for curtailing epidemic spread, by way of interventions targeted to well-connected nodes [7]. True contact networks are a challenging type of data to collect, requiring extensive epidemiological investigation [8, 9].

Viral sequence data, on the other hand, has become ~~relatively inexpensive and straightforward to collect on a population level.~~ easier to collect as the cost of sequencing has declined. In the case of HIV, genotyping has become part of routine clinical care in several health regions. Due to the high mutation rate of RNA viruses, epidemiological processes impact the course of viral evolution, thereby shaping the inter-host viral phylogeny [10]. The term "phylodynamics" was coined to describe this interaction, as well as the growing family of inference methods to estimate epidemiological parameters from viral phylogenies [11]. These methods have revealed diverse properties of local viral outbreaks, from basic reproductive number [12], to the degree of clustering [13], to the elevated transmission risk during acute infection [14]. On the other hand, although sophisticated methods have been developed for fitting complex population genetic models to phylogenies [15, 16], inference of structural network parameters has to date been limited. However, it has been shown that network structure has a tangible impact on phylogeny shape [6, 17–20], suggesting that such statistical inference might be possible [8]. In the context of networks, sequence data have the advantage of being objective, in that they are not affected by misreporting. However, just as with survey data, it is important to collect a representative sample from the population to perform accurate inference [21].

Survey-based studies of sexual networks [22–27] have found that these networks tend to have a degree distribution which follows a power law [although there has been some disagreement, see 28]. That is, the number of nodes of degree k is proportional to $k^{-\gamma}$ for some constant γ. These networks are also referred to as "scale-free" [29]. One process by which scale-free networks can be generated is preferential attachment, where nodes with a high number of contacts attract new connections at an elevated rate. The first contact network model incorporating prefer-

ential attachment was introduced by Barabási and Albert [29], and is now referred to as the Barabási-Albert (BA) model. Under this model, networks are formed by iteratively adding nodes with m new edges each. In the most commonly studied formulation, these new edges are joined to existing nodes of degree k with probability proportional to k, so that nodes of high degree tend to attract more connections. Barabási and Albert suggested an extension where the probability of attaching to a node of degree k is k^α for some non-negative constant α, and we use this extension in this work. When $\alpha \neq 1$, the degree distribution is no longer a power law: for $\alpha < 1$, the distribution is a stretched exponential, while for $\alpha > 1$, it is a "gelation" type distribution where one or a few hub nodes are connected to nearly every other node in the graph [30].

Previous work offers precedent for the possibility of statistical inference of structural network parameters. Britton and O'Neill [31] develop a Bayesian approach to estimate the edge density in an Erdős-Rényi network [32] given observed infection dates, and optionally recovery dates. Their approach was later extended by Groendyke, Welch, and Hunter [33] and applied to a much larger data set of 188 individuals. Volz and Meyers [34] and Volz [35] developed differential equations describing the spread of a susceptible-infected (SI) epidemic on static and dynamic contact networks with several degree distributions, which could in principle be used for inference if observed incidence trajectories were available. Leigh Brown et al. [36] analyzed the degree distribution of an approximate transmission network, estimated based on genetic similarity and estimated times of infection, relating 60% of HIV-infected men who have sex with men (MSM) in the United Kingdom. The transmission network is a subgraph of the contact network which includes only those edges which have already led to a new infection. The authors found that a Waring distribution, which is produced by a more sophisticated preferential attachment model, was a good fit to their estimated network.

Standard methods of model fitting involve calculation of the likelihood of observed data under the model. In maximum likelihood estimation, a quantity proportional to the likelihood is optimized, often through a standard multi-dimensional numerical optimization procedure. Bayesian methods integrate prior information by optimizing the posterior probability instead. To avoid calculation of a normalizing constant, Bayesian inference is often performed using Markov chain Monte Carlo (MCMC), which uses likelihood *ratios* in which the normalizing constants cancel out. Unfortunately, it is generally difficult to explicitly calculate the likelihood of an observed transmission tree under a contact network model, even up to a normalizing constant. To do so, it would be necessary to integrate over all possible networks, and also over all possible labellings of the internal nodes of the transmission tree. While it is not known (to us) whether such integration is tractable, a simpler alternative is offered by likelihood-free methods, namely approximate Bayesian computation (ABC) [37, 38]. ABC leverages the fact that, although calculating the likelihood may be impractical, generating simulated datasets

according to a model is often straightforward. If our model fits the data well, the simulated data it produces should be similar to the observed data. More formally, if D is the observed data, the posterior distribution $f(\theta \mid D)$ on model parameters θ is replaced as the target of statistical inference by $f(\theta \mid \rho(\hat{D},D) < \varepsilon)$, where ρ is a distance function, $\hat{D}$ is a simulated dataset according to θ, and ε is a small tolerance [39]. ~~In the specific case when ρ is a kernel function, the approach is known as ABC [40, 41].~~ Our group [41] and others [42] have demonstrated that taking ρ to be a well-chosen kernel function can produce a more accurate ABC approximation than the typical choice of a difference of summary statistics.

Here, we develop a method using ABC to estimate the parameters of contact network models from observed phylogenetic data. The distance function we use is the tree kernel developed by Poon et al. [43], which computes a weighted dot product of the trees' representations in the space of all possible subset trees. We apply the method to investigate the parameters of the BA network model on a variety of simulated and real datasets. Our results show that some network parameters can be inferred with reasonable accuracy, while others ~~have a minimal detectable impact on tree shape and therefore cannot be estimated accurately~~ are weakly- or non-identifiable with ABC . We also find that these parameters can vary considerably between real epidemics from different settings.

Methods

Netabc: phylogenetic inference of contact network parameters with ABC

We have developed an ABC-based method to perform statistical inference of contact network parameters from a transmission tree estimated from an observed viral phylogeny. We implemented the adaptive sequential Monte Carlo (SMC) algorithm for ABC developed by Del Moral, Doucet, and Jasra [44]. The SMC algorithm keeps track of a population of parameter "particles", which are initially sampled from the parameters' joint prior distribution. Several datasets are simulated under the model of interest for each of the particles. In this case, the datasets are transmission trees, which are generated by a two-step process. First, a contact network is simulated according to the network model being fit. Second, a transmission tree is simulated over that network with a Gillespie simulation algorithm [45], in the same fashion as several previous studies [*e.g.* 17, 19]. Tips of the simulated transmission tree are randomly removed until the simulated tree has the same number of tips as the input tree. The particles are weighted according to the similarity between their associated simulated trees and the observed tree. To quantify this similarity, we used the tree kernel developed by Poon et al. [43].

Particles are iteratively perturbed by applying a Metropolis-Hastings kernel and, if the move is accepted, simulating new datasets under the new parameters. When a particle's weight drops to zero, because its simulated trees are too dissimilar to the observed tree, the particle is dropped from the population, and eventually replaced by a resampled particle with a higher weight. As the algorithm progresses, the population converges to a Monte Carlo approximation of the ABC target distribution, which is assumed to approximate the desired posterior [39, 44].

In the original formulation of ABC-SMC [46, 47], the user is required to specify a decreasing sequence of tolerances $\{\varepsilon_i\}$. At iteration i, particles with no associated simulated datasets within distance ε_i of the observed data are removed from the population. In the adaptive version of Del Moral, Doucet, and Jasra [44], the sequence of tolerances is determined automatically by fixing the decay rate of the population's expected sample size (ESS) to a user-defined value. Del Moral, Doucet, and Jasra call this value α, but we will refer to it here as α_{ESS} to avoid confusion with the preferential attachment power parameter of the BA model.

To check that our implementation of Gillespie simulation was correct, we reproduced Figure 1A of Leventhal et al. [17] (our fig. S1), which plots the imbalance of transmission trees simulated over four network models at various levels of pathogen transmissibility. Our implementation of adaptive ABC-SMC was tested by applying it to the same mixture of Gaussians used by Del Moral, Doucet, and Jasra to demonstrate their method (originally used by Sisson, Fan, and Tanaka [46]). We were able to obtain a close approximation to the function (see fig. S2), and attained the stopping condition used by the authors in a comparable number of steps.

Nodes in our networks followed simple SI dynamics, meaning that they became infected at a rate proportional to their number of infected neighbours, and never recovered. For all analyses, the transmission trees' branch lengths were scaled by dividing by their mean. We used the *igraph* library's implementation of the BA model [48] to generate the graphs. The analyses were run on Westgrid (`https://www.westgrid.ca/`) and a local computer cluster. A computer program implementing our method is freely available at `https://github.com/rmcclosk/netabc` (last accessed July 14, 2016).

~~Kernel classifiers~~ Classifiers for BA model parameters from tree shapes

We considered four parameters related to the BA model, denoted N, m, $alpha$, and I. The first three of these parameterize the network structure, while I is related to the simulation of transmission trees over the network. However, we will refer to all four as BA parameters. N denotes the total number of nodes in the network, or equivalently, susceptible individuals in the population. m is the number of new undirected edges added for each new vertex,

or equivalently one-half of the average degree. α is the power of preferential attachment – new nodes are attached to existing nodes of degree d with probability proportional to $d^\alpha + 1$. Finally, I is the number of infected individuals at the time when sampling occurs. The α parameter is unitless, while m has units of edges or connections per vertex, and N and I both have units of nodes or individuals.

Before proceeding with a full validation of *netabc* on simulated data, we undertook an experiment designed to assess the identifiability of the BA parameters. One parameter of the BA model was investigated at a time while holding all others fixed, a strategy commonly used when performing sensitivity analyses of mathematical models. This allowed us to perform a fast preliminary analysis without dealing with the "curse of dimensionality" of the full parameter space. We simulated trees under three different values of each parameter, and asked how well we could tell the different trees apart. The better we are able to distinguish the trees, the more identifiability we might expect for the corresponding parameter when we attempt to estimate it with ABC.

This experiment also had the secondary purpose of validating our choice of the tree kernel as a distance measure in ABC. To tell the trees apart, we used a classifier based on the tree kernel, but we also tested two other tree shape statistics. Sackin's index [49] is a measure of tree imbalance which not take branch lengths into account, considering only the topology. The normalized lineages-through-time [nLTT, 50] compares two trees based on normalized distributions of their branching times, and does not explicitly consider the topology. Since the tree kernel incorporates both of these sources of information, we expected it to outperform the other two statistics. Finally, the tree kernel can be tuned by adjusting the values of the meta-parameters λ and σ (the "decay factor" and "radial basis function variance", see Poon et al. [43]). λ is used to penalize ween large subset trees which tend to dominate the kernel score. When $\lambda = 0$, all but the root branches of each subset tree are ignored, while when $\lambda = 1$, no penalty is applied. σ controls how strictly the notion of similarity is applied to branch lengths. When $\sigma = 0$, branch lengths must match exactly, while as $\sigma \to \infty$, branch lengths are not considered at all.

~~We used the phylogenetic kernel developed by Poon et al. [43]~~ To test whether the parameters of the BA model had a measurable effect on tree shape, 100 networks were simulated under each of three different values of α: 0.5, 1.0, and 1.5 (300 networks total). The other parameters were fixed to the following values: $N = 5000$, $I = 1000$, and $m = 2$. A transmission tree with 500 tips was simulated over each network (300 transmission trees total). The 300 trees were compared pairwise with the tree kernel to form a 300×300 kernel matrix. The kernel meta-parameters λ ~~(the "decay factor"),~~ and σ ~~(the "radial basis function variance") [see 43],~~ were set to 0.3 and 4 respectively. We also computed a 300×300 matrix of pairwise nLTT values, and a 1×300 vector of Sackin's index values. We constructed three classifiers for

α: a kernel support vector regression (kSVM) from the kernel matrix with the *kernlab* package [51], and two ordinary SVMs from the nLTT matrix and Sackin's index vector with the e1071 package [52]., and a linear regression from the Sackin's index values. The accuracy of each classifier was evaluated with 1000 two-fold cross validations with equally-sized folds .

Three similar experiments were performed for the other BA model parameters (one experiment per parameter). m was varied between 2, 3, and 4; I between 500, 1000, and 2000; and N between 3000, 5000, and 8000. The parameters not being tested were fixed at the values $N = 5000$, $I = 1000$, $m = 2$, and $\alpha = 1$. Thus, we performed a total of four cross-validations for each classifier , one for each of the BA model parameters α, I, m, and N. We repeated these four cross-validations with different values of λ (0.2, 0.3, and 0.4) and σ ($2^{-3}, 2^{-2}, \ldots, 2^{3}$), as well as on trees with differing numbers of tips (100, 500, and 1000). For the structural parameters α, m, and N, the experiments were repeated with three different fixed values of I (500, 1000, and 2000). and in epidemics of differing size (500, 1000, and 2000). The combination of the number of sampled individuals (*i.e.* the number of tips) and the epidemic size (*i.e.* I) will be referred to as an "epidemic scenario". When evaluating the classifier for I, we did not consider trees with 1000 tips, because one of the tested I values was 500, and the number of tips cannot be larger than I.

For each of the four parameters, we also tested a linear regression against Sackin's index [49] and an ordinary SVM based on the normalized lineages-through-time (nLTT) statistic [50].

ABC simulations

We tested *netabc* by jointly estimating the four parameters of the BA model. We used the standard validation approach of simulating transmission trees under the model with known parameter values and attempting to recover those values with *netabc*. The algorithm was not informed of any of the true parameter values for the main set of simulations. We simulated three transmission trees, each with 500 tips, under every element of the Cartesian product of these parameter values: $N = 5000$, $I = \{1000, 2000\}$, $m = \{2, 3, 4\}$, and $\alpha = \{0.0, 0.5, 1, 1.5\}$. This produced a total of 24 parameter combinations $\times$ three trees per combination = 72 trees total. The adaptive ABC algorithm was applied to each tree with these priors: $m \sim$ DiscreteUniform(1, 5), $\alpha \sim$ Uniform(0, 2), and (N, I) jointly uniform on the region $\{500 \leq N \leq 15000, 500 \leq I \leq 5000, I \leq N\}$. Proposals for α, N, and I were Gaussian, while proposals for m were Poisson. Following Del Moral, Doucet, and Jasra [44] and Beaumont et al. [47], the variance of all proposals was equal to the empirical variance of the particles.

The SMC settings used were 1000 particles, 5 simulated datasets per particle, and the "quality" parameter controlling the decay rate of the tolerance ϵ set to $\alpha_{\text{ESS}} = 0.95$. We used the same stopping criterion as Del Moral, Doucet, and Jasra, namely when the MCMC

acceptance rate dropped below 1.5%. ~~Point estimates for the parameters were obtained by taking the highest point of an estimated kernel density on the final set of particles, calculated using the *density* function with the default parameters in *R*.~~ Approximate posterior means for the parameters were obtained by taking the weighted average of the final set of particles. Highest posterior density (HPD) intervals were calculated with the *HPDinterval* function from the *R* package *coda* [53].

To evaluate the effects of the true parameter values on the accuracy of the posterior mean estimates, we analyzed the α and I parameters individually using generalized linear models (GLMs) The response variable was the error of the point estimate, and the predictor variables were the true values of α, I, and m. We did not test for differences across true values of N, because N was not varied in these simulations. The distribution family and link function for the GLMs were Gaussian and inverse, respectively, chosen by examination of residual plots and Akaike information criteria (AIC). The p-values of the estimated GLM coefficients were corrected using Holm-Bonferroni correction [54] with $n = 6$ (two GLMs with three predictors each). Because there was clearly little to no identifiability of N and m with ABC (see results in next section), we did not construct GLMs for those parameters.

Two further simulations were performed to address ~~potential sources of error~~ the possible impact of two types of model misspecification . To evaluate the effect of model misspecification in the case of heterogeneity among nodes, we generated a network where half the nodes were attached with power $\alpha = 0.5$, and the other half with power $\alpha = 1.5$. The other parameters for this network were $N = 5000$, $I = 1000$, and $m = 2$. To investigate the effects of potential sampling bias, we simulated a transmission tree where the tips were sampled in a peer-driven fashion, rather than at random. That is, the probability to sample a node was twice as high if any of that node's network peers had already been sampled. The parameters of this network were $N = 5000$, $I = 2000$, $m = 2$, and $\alpha = 0.5$.

Despite the fact that the parameter values used to generate the simulated transmission trees were known, the true posterior distributions of the BA parameters were unknown. Therefore, any apparent errors or biases in the estimates could be due to either poor performance of our method, or to real features of the posterior distribution. Two retrospective experiments were performed to disambiguate some of the observed errors. To assess the impact of the SMC settings on *netabc*'s accuracy, we ran *netabc* twice on the same simulated transmission tree. For the first run, the SMC settings were the same as in the other simulations: 1000 particles, 5 simulated transmission trees per particle, and $\alpha_{\text{ESS}} = 0.95$. The second run was performed with 2000 particles, 10 simulated transmission trees per particle, and $\alpha_{\text{ESS}} = 0.97$. To investigate the extent to which errors in the estimated BA parameters were due to true features of the posterior, rather than an inaccurate ABC approximation, we performed marginal estimation for one set of parameter values. Each combination of 1, 2, or 3 model parameters

Reference	Sequences (n)	Location	Risk group	Gene
Zetterberg et al. [55]	171	Estonia	IDU	*env*
Niculescu et al. [56]	136	Romania	IDU	*pol*
Novitsky et al. [57] Novitsky et al. [21]	180	Mochudi, Botswana	HET	*env*
McCormack et al. [58]	141/154	Karonga District, Malawi	HET	*env*/*gag*
Grabowski et al. [59]	225	Rakai District, Uganda	HET	*env*/*gag*
Wang et al. [7]	173	Beijing, China	MSM	*pol*
Kao et al. [60]	275	Taiwan	MSM	*pol*
Little et al. [61]	180	San Fransisco, USA	MSM	*pol*
Li et al. [62]	280	Shanghai, China	MSM	*pol*
Cuevas et al. [63]	287	Basque Country, Spain	mixed	*pol*

Table 1: Characteristics of published datasets investigated with ABC. Acronyms: MSM, men who have sex with men; IDU, injection drug users; HET, heterosexual. The HET data were sampled from a primarily heterosexual risk environment, but did not explicitly exclude other risk factors. The number of sequences column indicates how many sequences were included in our analysis; there may have been additional sequences linked to the study which we excluded for various reasons (see methods).

(14 combinations total) was fixed to their known values, and the remaining parameters were estimated with *netabc*. The parameter values were $\alpha = 0.0$, $m = 2$, $I = 2000$, and $N = 5000$.

Investigation of published data

We applied our ABC method to ten published HIV datasets. Because the BA model generates networks with a single connected component, we specifically searched for datasets which originated from existing clusters, either phylogenetically or geographically defined. Characteristics of the datasets we investigated are given in table 1. For clarity, we will refer to each dataset by its risk group and location of origin in the text. For example, the Zetterberg et al. [55] data will be referred to as IDU/Estonia.

We downloaded all sequences associated with each published study from GenBank. For the IDU/Romania data, only sequences from injection drug users (IDU, whose sequence identifiers included the letters "DU") were included in the analysis. Kao et al. [60] (MSM/Taiwan) found a strong association in their study population between subtype and risk group - subtype B was most often associated with men who have sex with men (MSM), whereas IDU were usually infected with a circulating recombinant form. Since there were many more subtype B sequences in their data than sequences of other subtypes, we restricted our analysis to the subtype B sequences and labelled this dataset as MSM. Two datasets (HET/Uganda and HET/Malawi) included both *env* and *gag* sequences. Each gene was analyzed separately

to assess the robustness of *netabc* to the particular HIV gene sequence used to estimate a transmission tree. The IDU/Estonia data also sequenced both genes, but the highly variable coverage and high homology of the *gag* sequences made it impossible to obtain a sufficiently large block of non-identical sequences to analyze. Therefore, we analyzed only *env* for this dataset.

~~For the Novitsky et al. [21] data,~~ Each *env* sequence was aligned pairwise to the HXB2 reference sequence (GenBank accession number K03455), and the hypervariable regions were clipped out with *BioPython* version 1.66+ [64]. Sequences were multiply aligned using *MUSCLE* version 3.8.31 [65], and alignments were manually inspected with *Seaview* version 4.4.2 [66]. Duplicated sequences were removed with *BioPython*. Phylogenies were constructed from the nucleotide alignments by approximate maximum likelihood using *FastTree2* version 2.1.7 [67] with the generalized time-reversible (GTR) model [68]. Transmission trees were estimated by rooting and time-scaling the phylogenies by root-to-tip regression, using a modified version of *Path-O-Gen* (distributed as part of *BEAST* [69]) as described previously [41]. Due to the removal of duplicated sequences, all estimated transmission trees were fully binary.

To check if our results were robust to the choice of phylogenetic reconstruction method, we built and reanalyzed phylogenies for the datasets with the lowest and highest estimated α values (mixed/Spain and IDU/Estonia) with *RAxML* [70] with the GTR+Γ model of sequence evolution and rate heterogeneity. The trees were rooted and time-scaled with *Least Square Dating* [*LSD*, 71]. For expediency, the analysis was run with the prior $m \sim \text{DiscreteUniform}(2,5)$, which defines a smaller total search space than the prior allowing $m = 1$. For both of these datasets, we also analyzed five bootstrap replicate alignments generated by resampling alignment columns with replacement.

~~Two~~ Four of the datasets ~~[21, 62]~~ (MSM/Shanghai, HET/Botswana, HET/Uganda, and MSM/USA) were initially much larger than the others, containing 1265, 1299, 1026/915 (*env*/*gag*), and 648 sequences respectively. To ensure that the analyses were comparable, we reduced these to a number of sequences similar to the smaller datasets. For the MSM/Shanghai data, we detected a cluster of size 280 using a patristic distance cutoff of 0.02 as described previously [72]. Only sequences within this cluster were carried forward. For the HET/Uganda, HET/Botswana, and MSM/USA data, no large clusters were detected using the same cutoff, so we analyzed ~~a subtree~~ subsets of sizes 255, 180, and 180 respectively. The subset of the HET/Uganda data was chosen by eye such that the individuals were monophyletic in both the *gag* and *env* trees. The other subsets were arbitrarily chosen subtrees from phylogenies of the complete datasets.

For all datasets, we used the priors $\alpha \sim \text{Uniform}(0, 2)$ and N and I jointly uniform on the region $\{n \leq N \leq 10000, n \leq I \leq 10000, I \leq N\}$, where n is the number of tips in the tree. Since the value $m = 1$ produces networks with no cycles, which we considered fairly implausible,

we ran one analysis with the prior $m \sim$ DiscreteUniform(1, 5), and one with the prior $m \sim$ DiscreteUniform(2, 5). The other parameters to the SMC algorithm were the same as used for the simulation experiments, except that we used 10000 particles instead of 1000 to increase the accuracy of the estimated posterior for all analyses except the bootstrap replicates. This was computationally feasible due to the small number of runs required for this analysis.

Results

~~Kernel classifiers~~ Classifiers for BA model parameters from tree shapes

We investigated the identifiability of four parameters of the BA network model [29]: the number of nodes N, the preferential attachment power α, the number of edges added per vertex m, and the number of infected nodes I. ~~In addition to m and α (see Introduction), we considered N, which denotes the total number of nodes in the network, and I, which is the number of infected nodes at which to stop the simulation and sample the transmission tree.~~ To examine the effect of these parameters on tree shape, we simulated transmission trees under different parameter values, calculated pairwise tree kernel scores between them, and attempted to classify the trees using a kernel support vector machine (kSVM). We also tested classifiers based on Sackin's index [49] and the normalized lineages-through-time (nLTT) statistic [50]. We report the accuracy of the classifiers, which is simply the proportion of trees which were assigned the correct parameter value. Since there were three possible values, random guessing would produce an accuracy of 0.33. ~~The accuracy of each classifier varied based on the parameter being tested~~ The results are shown in fig. 1. Classifiers based on ~~two other tree statistics,~~ the nLTT and Sackin's index generally exhibited worse performance than the tree kernel, although the magnitude of the disparity varied between the parameters (fig. 1, centre and right). ~~The results were largely robust to variations in the tree kernel meta-parameters λ and σ, although accuracy varied between different epidemic and sampling scenarios~~ Larger datasets were generally classified more accurately (figs. S3 to S6), although large values of λ produced worse estimates on large datasets. Extremely low σ values, which require nearly-exact matches between branch lengths, resulted in low accuracy in some cases (*e.g.* fig. S3, center row).

The kSVM classifier for α had an average accuracy of 0.92, compared to 0.6 for the nLTT, and 0.77 for Sackin's index. ~~There was little variation about the mean for different tree and epidemic sizes.~~ No classifier could accurately identify m in any epidemic scenario, with average accuracy values of 0.35 for kSVM, 0.32 for the nLTT, and 0.38 for Sackin's index. There was little variation in accuracy between epidemic scenarios, although the accuracy of

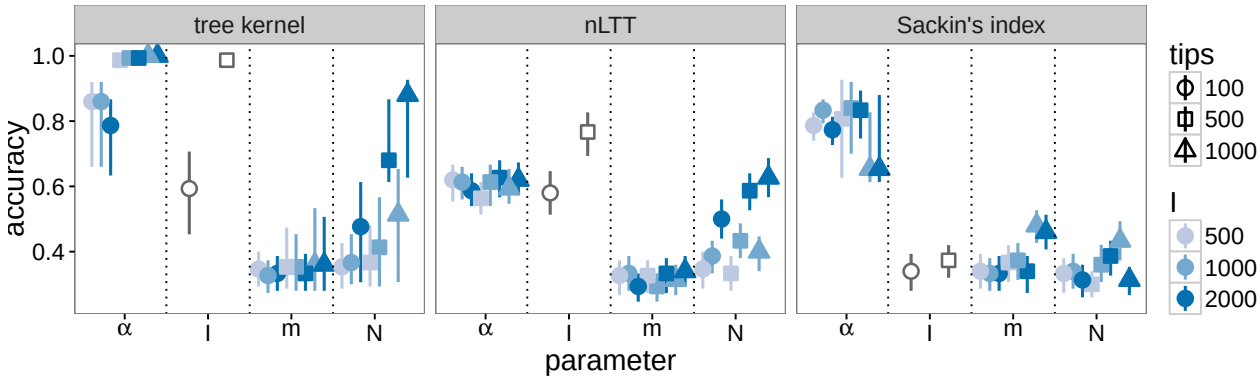

Figure 1: Cross-validation accuracy of kernel-SVM classifier (left), SVM classifiers using nLTT (centre) and Sackin's index (right) for BA model parameters. Kernel meta-parameters were set to $\lambda = 0.3$ and $\sigma = 4$. Each point was calculated based on 300 simulated transmission trees over networks with three different values of the parameter being tested, assuming perfect knowledge of the other parameters. Vertical lines are empirical 95% confidence intervals based on 1000 two-fold cross-validations. The classifiers for I were not evaluated with 1000-tip trees, because one of the tested I values was 500, and it is not possible to sample a tree of size 1000 from 500 infected individuals.

the kSVM was slightly higher on 1000-tip trees (fig. 1, left).

The accuracy of classifiers for I varied significantly with the number of tips in the tree. For 100-tip trees, the average accuracy was 0.59, 0.58, and 0.34 for the tree kernel, nLTT, and Sackin's index respectively. For 500-tip trees, the values increased to 0.99, 0.76, and 0.37. Finally, the performance of classifiers for N depended heavily on the epidemic scenario. The accuracy of the kSVM classifier ranged from 0.36 for the smallest epidemic and smallest sample size, to 0.81 for the largest. Accuracy for the nLTT ranged from 0.33 to 0.63. Sackin's index did not accurately classify N in any scenario, with an average accuracy of 0.35 and little variation between scenarios.

ABC simulations

Figure 2 shows ~~maximum *a posteriori* (MAP)~~ stratified posterior mean point estimates of the BA model parameters α and I , obtained with ABC on simulated data. The parameters m and N were not identifiable with ABC for any parameter combinations (fig. S7). ~~The estimates shown correspond only to the simulations where the m parameter was set to 2, however the results for $m = 3$ and $m = 4$ were similar.~~ Average boundaries of 95% HPD intervals for all parameters are given in table 2.

~~The accuracy of the parameter estimates obtained with ABC paralleled the results from the kSVM classifier. Of the four parameters, α was the most accurately estimated, with point~~

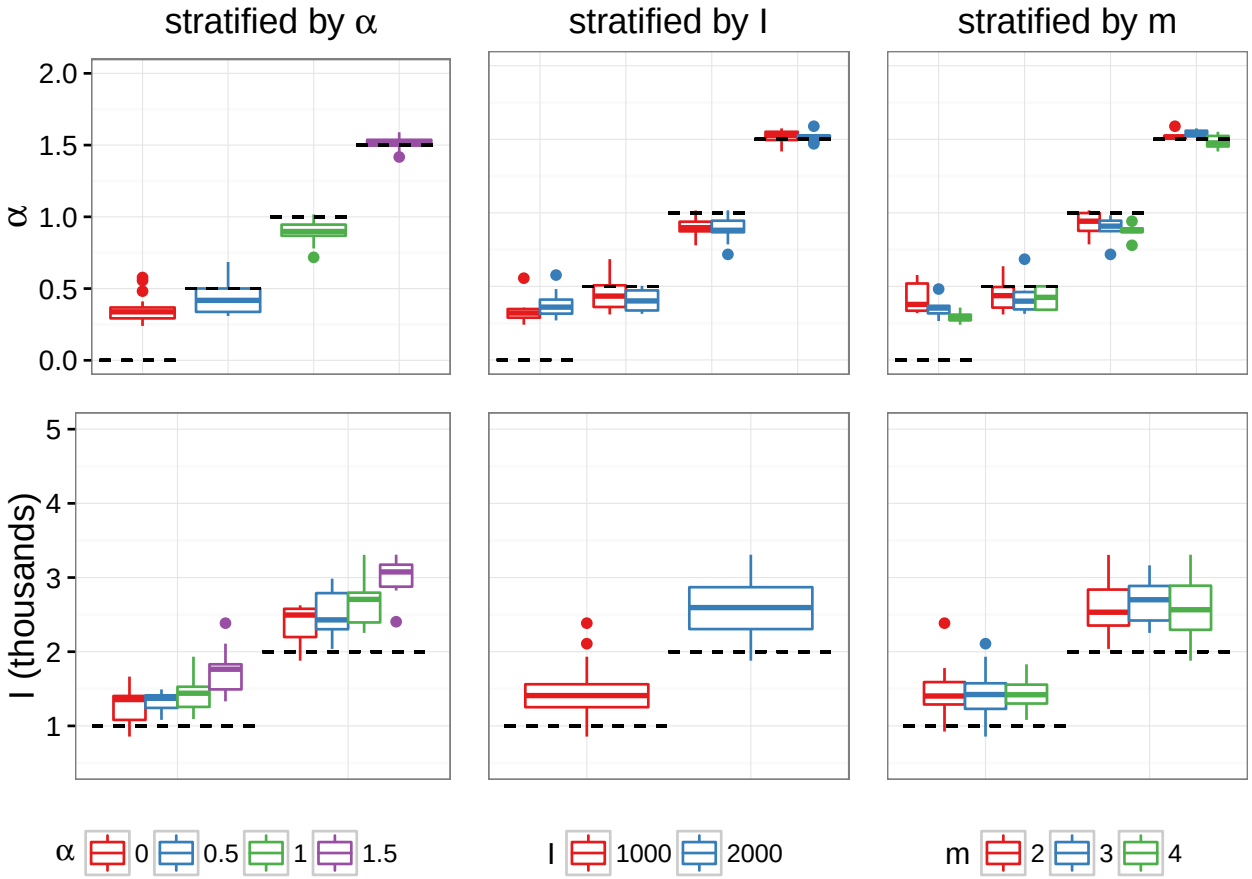

Figure 2: Posterior mean point estimates for BA model parameters α and I obtained by running *netabc* on simulated data, stratified by true parameter values. First row of plots contains true versus estimated values of α; second row contains true versus estimated values of I. Columns are stratified by α, I, and m respectively. Dashed lines indicate true values.

~~estimates having a median [IQR] absolute error of 0.11 [0.03 - 0.25]. The errors when the true value of α was zero were significantly greater than those for the other values (Wilcoxon rank-sum test, $p = 0$). Errors in estimating α also varied with the true value of m just at the threshold of statistical significance ($p = 0.5$), but did not vary across the true values of N or I (both one-way ANOVA). Estimates for I were relatively accurate, with point estimate errors of 492 [294 - 782] individuals. These errors were significantly higher when the true value of α was at least 1 (Wilcoxon rank-sum test, $p = 0$) and when the true value of I was 2000 ($p < 10^{-5}$). The true value of m did not affect the estimates of I (one-way ANOVA).~~

~~The m parameter was estimated correctly in only 37 % of simulations, barely better than random guessing. The true values of the other parameters did not significantly affect the estimates of m (both one-way ANOVA). Finally, the total number of nodes N was consistently over-estimated by about a factor of two (error 4153 [3660 - 4489] individuals). No parameters~~

Parameter	True value	Mean point estimate	Mean HPD lower bound	Mean HPD upper bound
α	0.0	0.36	0.01	0.81
	0.5	0.43	0.04	0.83
	1.0	0.90	0.51	1.09
	1.5	1.52	1.26	1.81
I	1000	1450	651	2592
	2000	2622	1114	4080
m	2	2.96	2.00	5.00
	3	3.04	2.04	4.96
	4	3.17	1.88	5.00
N	5000	9041	2613	14659

Table 2: Average posterior mean point estimates and 95% highest posterior density interval widths for BA model parameter estimates obtained with *netabc* on simulated data. Three transmission trees were simulated under each combination of the listed parameter values, and the parameters were estimated with ABC without training.

~~influenced the accuracy of the N estimates (all one-way ANOVA).~~

Across all simulations, the median [IQR] absolute errors of the parameter estimates obtained with *netabc* were 0.11 [0.03 - 0.25] for α, 492 [294 - 782] for I, 1 [0 - 1] for m, and 4153 [3660 - 4489] for N. These errors comprised, respectively, 6%, 11%, 17%, and 29% of the regions of nonzero prior density. For I and N, relative errors were 38% [20 - 50%] and 83% [73 - 90%]. Average 95% HPD interval widths were 0.68, 2454, 3.01, and 12046, representing 34%, 55%, 50%, and 83% of the nonzero prior density regions. Point estimates of I were upwardly biased: I was overestimated in 69 out of 72 simulations (96%). The estimates for m and N were similar across all simulations (median [IQR] point estimates 3 [3 - 3] and 9153 [8660 - 9489]) regardless of the true values of any of the BA parameters (fig. S7).

To analyze the effects of the true parameter values on the accuracy our estimates of α and I, we fitted one GLM for each of these two parameters, with error rate as the dependent variable and the true parameter values as independent variables. Since the estimates of m and N were roughly equal across all simulations (fig. S7), GLMs were not fitted for these parameters. The estimated coefficients are shown in table 3. The GLM analysis indicated that the error in estimates of α decreased with larger true values of α ($p<10^{-5}$) and m ($p = 0.01$) but was not significantly affected by I. Qualitatively, α seemed to be only weakly identifiable between the values of 0 and 0.5 (fig. 2). The error in the estimated I value was slightly lower for smaller values of α ($p<10^{-5}$) and I ($p = 0.05$), but was not significantly affected by the true value of m.

The dispersion of the ABC approximation to the posterior also varied between the pa-

Dependent variable	Independent variable	Estimate	Standard error	p-value
α	(Intercept)	2	0.6	0.01
	α	10	2	$<10^{-5}$
	I	-3×10^{-4}	2×10^{-4}	0.7
	m	0.5	0.2	0.01
I	(Intercept)	0.004	5×10^{-4}	$<10^{-5}$
	α	-0.001	2×10^{-4}	$<10^{-5}$
	I	-4×10^{-7}	2×10^{-7}	0.05
	m	-7×10^{-5}	8×10^{-5}	1

Table 3: Parameters of fitted GLMs relating error in estimated α and I to true values of BA parameters. GLMs ere fitted with the Gaussian distribution and inverse link function. Coefficients are interpretable as additive effects on the inverse of the mean error.

rameters~~, with narrower HPD intervals for the parameters with more accurate point estimates~~ (table 2). HPD intervals around α and I were often narrow relative to the region of nonzero prior density, whereas the intervals for m and N were more widely dispersed. Figures 3 and 4 ~~shows the distributions for one simulation.~~ show one- and two-dimensional marginal distributions for a simulation with α and I errors close to their respective medians. The parameters for this simulation were $\alpha = 1$, $I = 1000$, $m = 3$, and $N = 5000$. The two-dimensional marginals indicate some dependence between pairs of parameters, particularly I and N which show a diagonally shaped region of high posterior density.

To test the effect of model misspecification, we simulated one network where the nodes exhibited heterogeneous preferential attachment power (half 0.5, the other half 1.5), with m = 2, N = 5000, and I = 1000. The posterior mean [95% HPD] estimates for each parameter were: α, 1.03 [0.67 - 1.18]; I, 1474 [511 - 2990]; m, 3 [1 - 5]; N, 9861 [3710- 14977]. ~~The approximate one-dimensional marginal posterior distributions for this simulation are shown in ??.~~ To test the effect of sampling bias, we sampled one transmission tree in a peer-driven fashion, where the probability to sample a node was twice as high if one of its peers had already been sampled. The parameters for this experiment were N = 5000, m = 2, α = 0.5, and I = 2000. The estimated values were α, 0.3 [0 - 0.63]; I, 2449 [1417 - 3811]; m, 3 [2 - 5]; N, 9132 [2852 - 14780]. ~~The approximate one-dimensional marginal posterior distributions are shown in ??.~~ Both of these results were in line with estimates obtained on other simulated datasets (table 2), although the estimate of α for peer-driven sampling was somewhat lower than typical.

Figure S8 shows the effect of performing marginal ABC estimation of each of the BA parameters on the same simulated transmission tree. The estimates of m were apparently unaffected by marginalizing out the other parameters, corroborating the previous experiments'

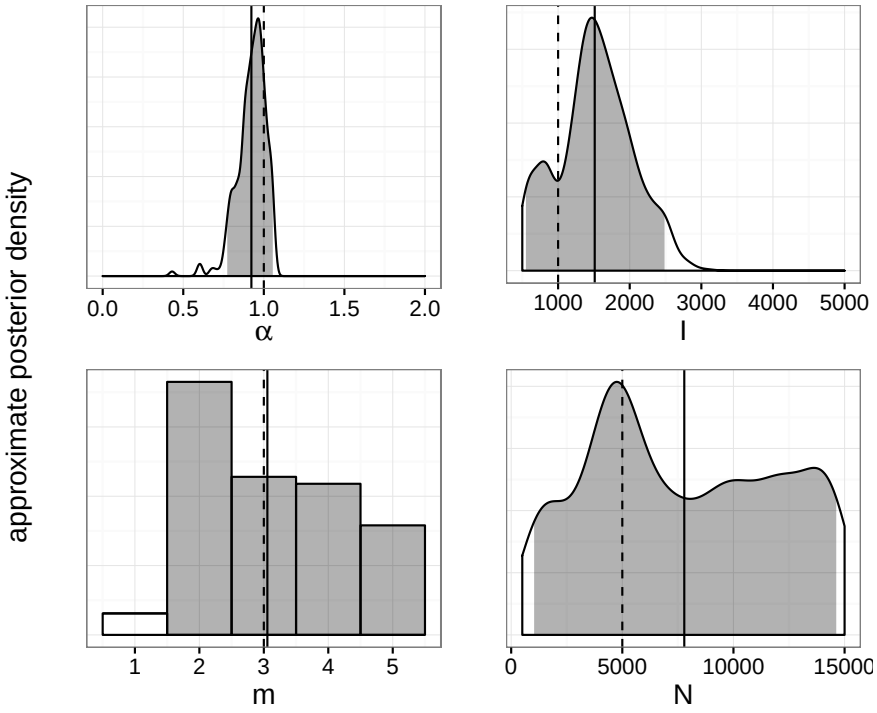

Figure 3: One-dimensional marginal posterior distributions of BA model parameters estimated by *netabc* from a simulated transmission tree. Dashed lines indicate true values, solid lines indicate posterior means, and shaded areas show 95% highest posterior density intervals.

findings that m is not an identifiable parameter from scaled tree shapes. Compared to allowing all parameters to vary, estimates of α, I, and N were improved by 41%, 59%, and 46% when all other parameters were fixed. Figure S9 shows the impact of increasing the number of particles, simulated datasets, and α_{ESS} parameter on the accuracy of a single simulation. The number of iterations until the stopping condition was reached was 81 with the basic settings and 124 with the higher settings. The results of the two simulations were similar, but surprisingly, the results with higher SMC settings were slightly worse (by 10%, 8%, and 11% for α, I, and N respectively). However, the 50% HPD interval for I was closer to the true value of 2000 with the improved settings (2338 - 3423, vs. 2810 - 3767 with basic settings). The estimate of m, 3 in both cases, was unaffected by the settings.

Published HIV data

We applied ABC to five published HIV datasets (table 1), and found substantial heterogeneity among the parameter estimates (figs. 5 and S12). Posterior mean point estimates and 50% and 95% HPD intervals for each parameter are shown in fig. 5. ~~Plots of the marginal posterior distributions for each dataset are shown in ??????????.~~ Figure S12 shows point estimates and

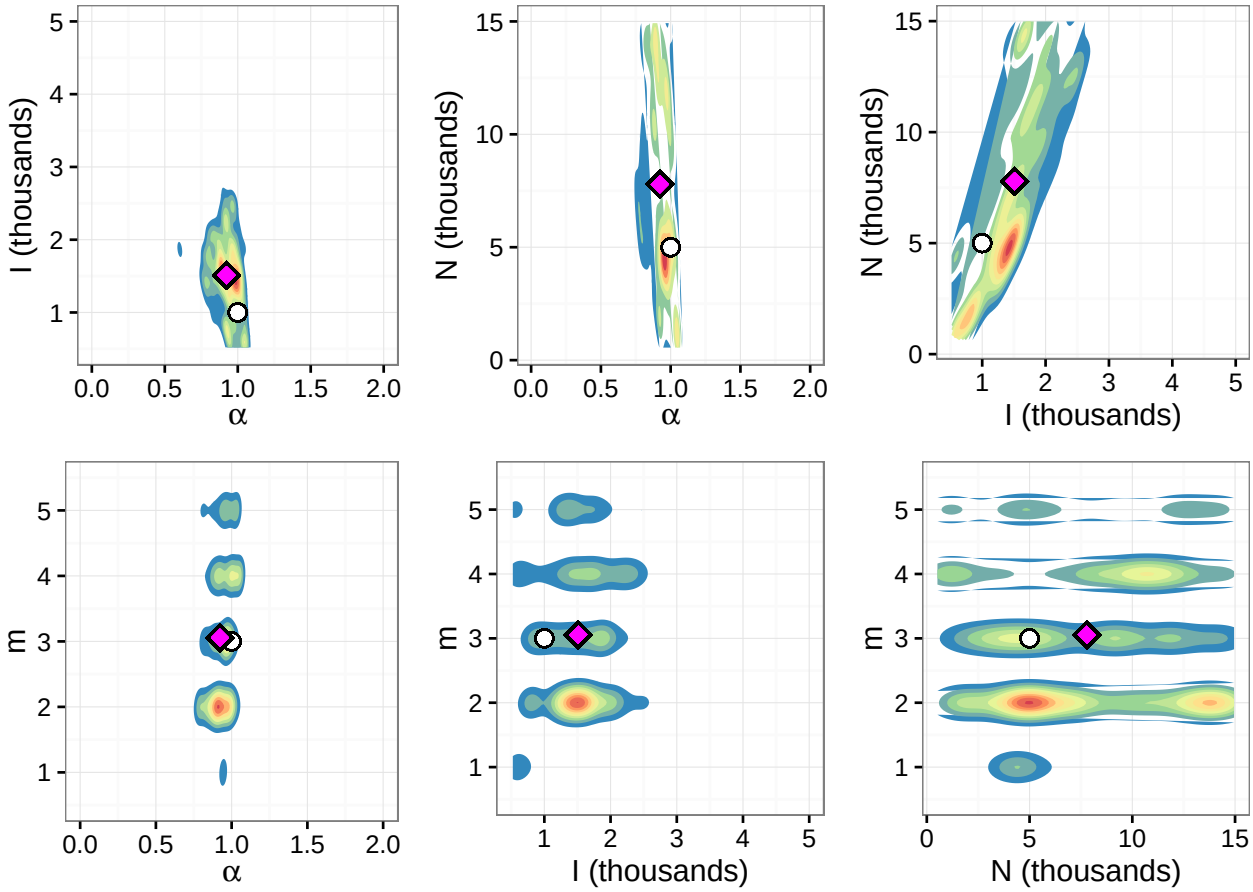

Figure 4: Two-dimensional marginal posterior distributions of BA model parameters estimated by *netabc* from a simulated transmission tree. White circles indicate true values, magenta diamonds indicate posterior means.

HPD intervals obtained when the value $m = 1$ was disallowed by the prior. Since the results indicated that $m = 1$ was the most credible value for several datasets, all results discussed henceforth apply to the prior $m \sim \text{DiscreteUniform}(1,5)$ unless otherwise stated.

~~Two of the datasets [7, 56] had estimated α values near unity for the prior allowing $m = 1$ (MAP estimates [95% HPD] 0.73 [0.05 - 1.18] and 0.55 [0.01 - 0.99] respectively). The MAP estimates did not change appreciably when $m = 1$ was disallowed by the prior, although the credible interval of the Niculescu et al. [56] data was narrower (0.05 - 1.18). When $m = 1$ was permitted, the Li et al. [62] and Cuevas et al. [63] both had low estimated α values (0.33 [0 - 0.76] and 0.27 [0 - 0.59]). However, the MAP estimates increased when $m = 1$ was not permitted, although the HPD intervals remained roughly the same (0.58 [0.06 - 0.99] and 0.48 [0.02 - 0.87]). The Novitsky et al. [21] data had a fairly low estimated α for both priors on m (0.55 for $m \geq 1$; 0.53 for $m \geq 2$). However, the confidence interval was much wider when~~

~~$m = 1$ was allowed ([0 - 1.75] for $m \geq 1$ vs. 0 - 1.75 for $m \geq 2$).~~

Posterior mean point estimates for the preferential attachment power α were all sub-linear, ranging from 0.27 (mixed/Spain) to 0.83 (IDU/Estonia). When aggregated by risk group, the average estimates were 0.78 for IDU, 0.41 for primarily heterosexual risk, and 0.37 for MSM. 95% HPD intervals were very wide for most datasets, often encompassing nearly the entire range from 0 to 1 (fig. 5). As shown in fig. S13, the estimates of α were quite robust to the gene analyzed.

~~For all the datasets except Novitsky et al., estimated values of I were below 2000 when $m = 1$ was allowed, with relatively narrow HPD intervals compared to the nonzero prior density region (Cuevas et al., 701 [289 - 1279]; Niculescu et al., 747 [136 - 2378]; Li et al., 1390 [310 - 2821]; Wang et al., 675 [175 - 1400]). The Novitsky et al. data was the outlier, with a very high estimated I, and HPD interval spanning almost the entire prior region (5431 [183 - 8739]). The I estimates and HPD intervals were generally robust to the choice of prior on m, with slightly narrower HPD intervals (compare figs. 5 and S12).~~

For all but the HET/Botswana data, the posterior mean estimates for I were between 373 (IDU/Estonia) and 1391 (MSM/Shanghai). The HET/Botswana data had a much higher estimated I value (5432) than the other datasets, with a very wide 95% HPD interval covering almost the entire prior region (fig. 5). There was no significant correlation between the number of sequences in the tree and the estimated I value (Spearman correlation, $p = 0.9$), indicating that the higher estimates were not simply due to increased sampling density. When both *gag* and *env* sequences were analyzed, the estimates from the *env* data were higher (HET/Uganda, 939 for *gag* vs. 1615 for *env*; HET/Malawi, 724 for *gag* vs. 845 for *env*).

~~The MAP estimate of m was equal to 1 for all but the Novitsky et al. data, when this value was allowed. However, the upper bound of the HPD interval was different for each dataset (Niculescu et al., 4; Wang et al., 1; Li et al., 1; Cuevas et al., 1). When $m = 1$ was disallowed, the MAP for all datasets was either 2 or 3, with HPD intervals spanning the entire prior region. The estimates for the total number of nodes N were largely uninformative for all samples, with almost all MAP estimates greater than 7500 and HPD intervals spanning almost the entire nonzero prior density region. The only exception was the Li et al. data, for which the MAP estimate was lower (5916) when $m = 1$ was allowed.~~

The posterior means of m were equal to one for zero of the datasets analyzed. The widths of the 95% HPD intervals varied from 0 (all the mass on the estimated value) to 5 (the entire prior region). Estimates of N were mostly uninformative, with very similar estimates for all datasets (mean 6202, range 5881 - 6882). This was similar to the pattern observed for the synthetic data, where the posterior mean always fell around the upper two-thirds mark of the range (fig. S7).

When the value $m = 1$ was disallowed by the prior, the separation in α between the IDU datasets and the others became more striking (fig. S12). Both IDU datasets had estimated α values at or above 1. The estimate for the MSM/Beijing data was slightly lower (0.85) and the estimates for the seven remaining non-IDU datasets were bounded above by 0.58. The values of I were fairly robust to the choice of prior (compare figs. 5 and S12), although the 95% HPD intervals were slightly narrower (average width 2159 for $m \geq 1$ and 1874 for $m \geq 2$). The posterior means of m for all but the HET/Botswana data took on the value 3 with this prior, with the HPD intervals spanning the entire prior region. This is very similar to the results observed for m on simulated data (table 2), and suggests that m is not identifiable from these data with this prior. The results for N did not change appreciably between the two choices of prior.

For the two datasets we reanalyzed using RAxML [70] and LSD [71], α was relatively robust to the choice of method (fig. S14, posterior means 0.48 vs. 0.48 for mixed/Spain and 1.02 vs. 1.12 for IDU/Estonia). However, the estimates of I were about twice as high when RAxML was used instead of FastTree to reconstruct the trees (228 vs. 437 for IDU/Estonia, 816 vs. 1949 for mixed/Spain). Figure S15 shows estimates obtained for five bootstrap replicate alignments for each of these two datasets. For the mixed/Spain data, the estimated posterior mean [range of bootstrap posterior means] was 0.48 [0.54 - 0.67] for α, 816 [403 - 886] for I, 2.76 [2.61 - 3.24] for m, and 6639 [6652 - 7245] for N. For the IDU/Estonia data, values were 1.02 [0.78 - 1.07] for α, 228 [313 - 741] for I, 3.11 [3.41 - 3.46] for m, and 6803 [5941 - 6913] for N.

Discussion

Contact networks can have a strong influence on epidemic progression, and are potentially useful as a public health tool [7, 61]. Despite this, few methods exist for investigating contact network parameters in a phylodynamic framework [although see 17, 33, 35, 36, for related work]. ABC is a model-agnostic method which can be used to investigate any quantity that affects tree shape [41]. In this work, we developed a ABC-based method to infer the parameters of a contact network model. The method is general, and could be applied to any model from which contact networks can be simulated. We demonstrated the method on the BA model, which is a simple preferential attachment model. For some parameter choices, the BA model gives rise to the power law degree distributions commonly observed in real-world networks.

Analysis of BA model with synthetic data

~~By training a kernel-SVM classifier, we found that the α and I parameters, representing preferential attachment power and number of infected nodes, had a strong influence on tree shape. This was reflected in the relative accuracy of the ABC estimates of these parameters. The total number of nodes N had a weak influence on tree shape, which was most prominent when the epidemic size I and number of sampled tips were both large. The m parameter, representing the number of edges created in the network per vertex, did not produce much variation in tree shape, resulting in in both poorly performing classifiers and uninformative ABC estimates.~~

The preferential attachment power α had a strong influence on tree shape in the range of values we considered. Although the tree kernel was the most effective classifier for α, a Sackin's index of tree imbalance performed nearly as well (fig. 1). High α values produce networks with few well-connected "superspreader" nodes which are involved in a large number of transmissions, resulting in a highly unbalanced ladder-like tree structure. There appeared to be weaker identifiability for $\alpha < 1$ than for $\alpha \geq 1$ (fig. 2 and table 2), which may be partially explained by the relationship between α and the power law exponent γ (fig. S10). Although the degree distributions do not truly follow a power law for $\alpha \neq 1$, the fitted exponent still captures the shape of the degree distribution reasonably well (fig. S11). The γ values fitted to $\alpha = 0$ and $\alpha = 0.5$ are nearly identical (about 2.28 for $\alpha = 0$ and 2.33 for $\alpha = 0.5$ with $N = 5000$ and $m = 2$). In other words, the degree distributions of networks with $\alpha < 1$ are similar to each other, which may result in similarity of corresponding transmission trees as well.

I, representing the number of infected individuals at the time of sampling, was also identifiable, albeit over-estimated with ABC for both values we considered. Sackin's index was better able to discern I from tree shape than the nLTT (figs. 1 and S5), suggesting that this parameter impacts the distribution of branching times in the tree more than the topology. In a homogeneously-mixed population, branching times can be modelled by the coalescent process [73], in our case under the SI model [74]. Although networks are not homogeneously mixed, the forces which affect the distribution of branching times still apply. In our simulations, all discordant edges shared the same transmission rate, so that the waiting time until the next transmission in the entire network was always inversely proportional to the number of discordant edges. In the initial phase of the epidemic, when I is small, each new transmission results in many new discordant edges. Hence, there is an early exponential growth phase, producing many short branches near the root of the tree. As the epidemic gets closer to saturating the network, the number of discordant edges decays, causing longer waiting times.

The number of nodes in the network, N, exhibited the most variation in terms of its effect on tree shape. There was almost no measurable difference between trees simulated under different N values when the number of infected nodes I was small (fig. S6). In retrospect, it is unreasonable to expect good estimation of N, in many cases, because adding additional nodes

does not change the edge density or overall shape of a BA network. This can be illustrated by imagining that we add a small number of nodes to a network after the epidemic simulation has already been completed. If I is small relative to N, very few of the infected nodes will gain any new neighbours. Thus, the outcome of a second simulation on the same network will likely be very similar. ~~It is possible that none of these new nodes attains a connection to any infected node. Thus, running the simulation again on the new, larger network could produce the exact same transmission tree as before.~~ On the other hand, when I is large relative to N, the coalescent dynamics discussed above also apply. The waiting times until the next infection increase, resulting in longer coalescence times toward the tips. The relative accuracy of the nLTT in these situations (figs. 1 and S6) corroborates this hypothesis, as the nLTT uses only information about the coalescence times. When all BA parameters were simultaneously estimated with ABC, N was nearly always over-estimated by approximately a factor of two (fig. S7 and table 2). One factor which may have contributed to this bias was our choice of prior distribution. Since the prior for I and N was jointly uniform on a region where $I \leq N$, more prior weight was assigned to higher N values. We note also that ~~our accurate estimates of I may have been influenced by~~ this prior~~, which~~ places more mass on low I values. However, the estimate of I was very high for the HET/Botswana data, suggesting that a strong enough signal in the data can overcome the prior. Furthermore, when I was estimated marginally with fixed N, the accuracy of the estimate improved even though there was no longer any extra prior mass on low I values.

Another possible contributing factor to the overestimation of I and N relates to the dynamics of the SI model and the coalescent process. The number of infected individuals follows a logistic growth curve under the SI model. This kind of growth curve has three qualitative phases: a slow ramp-up, an exponential growth phase, and a slow final phase when the susceptible population is almost depleted. The waiting times until the next transmission, which determine the coalescence times in the tree, are dependent on the growth phase of the epidemic. Therefore, we hypothesize that it is the growth phase at the time of sampling which most affects tree shape, rather than the specific values of I or N. To investigate this hypothesis, we simulated transmission trees over networks on a grid of I and I values and fit logistic growth curves to the proportion of infected individuals over time. We and calculated the first and second derivatives of these curves (with respect to time) at the time of transmission tree sampling. As shown in fig. S16, there are bands along which both derivatives are similar which contain the values we tested. These bands span mostly higher values of N and I than the true values. Thus, if N and I are free to vary (as is the case in ABC), both parameters will tend to be overestimated due to being less identifiable within their own band. However, when N is fixed at 5000, the derivatives vary substantially along the I-axis, which explains why a marginal estimate of I was more accurate (fig. S8). We also note the resemblance of the contour surface of fig. S16 to

the two-dimensional marginal posterior distribution on I and N obtained with simulated data (fig. 4).

The m parameter, which controls the number of connections added to the network per vertex, did not have a measurable impact on tree shape and was not identifiable with ABC. It was pointed out to us by an anonymous reviewer that for a fixed I, an infected node may only end up transmitting along a fraction of its outgoing edges, which could mask the presence of the extra edges associated with higher m. If m were a continuous variable, it is possible that we would observe stronger identifiability for lower values (say between 0 and 2), where extra edges are more likely to be involved in the epidemic and have an impact on the transmission tree. One way to achieve this would be to draw m separately for each node from a distribution parameterized by a continuous variable.

As noted by Lintusaari et al. [75], uniform priors on model parameters may translate to highly informative priors on quantities of interest. We observed a non-linear relationship between the preferential attachment power α and the power law exponent γ (fig. S10). Therefore, placing a uniform prior on α between 0 and 2 is equivalent to placing an informative prior that γ is close to 2. Therefore, if we were primarily interested in γ rather than α, a more sensible choice of prior might have a shape informed by fig. S10 and be bounded above by approximately $\alpha = 1.5$. This would uniformly bound γ in the region $2 \leq \gamma \leq 4$ commonly reported in the network literature [22–24, 36]. We note however that Jones and Handcock [76] estimated γ values greater than four for some datasets, in one case as high as 17, indicating that a wider range of permitted γ values may be warranted.

Analysis of real world HIV datasets

Our investigation of published HIV datasets indicated heterogeneity in the contact network structures underlying several distinct local epidemics. When interpreting these results, we caution that the BA model is quite simple and most likely misspecified for these data. In particular, the average degree of a node in the network is equal to $2m$, and therefore is constrained to be a multiple of 2. Furthermore, we considered the case $m = 1$, where the network has no cycles, to be implausible and therefore assigned it zero prior probability in one set of analyses. This forced the average degree to be at least four, which may be unrealistically high for sexual networks. The fact that the estimated values of α differed substantially for several datasets depending on whether or not $m = 1$ was allowed by the prior is further evidence of this potential misspecification. However, we note that the ordering of the datasets with respect to α was similar between the two priors, ~~for two of the datasets, the estimated values of α did not change much between priors,~~ and the estimates of I were robust to the choice of prior for all datasets studied (compare figs. 5 and S12). More sophisticated models, for example models

incorporating heterogeneity in node behaviour, are likely to provide a better fit to these data.

With respect to the preferential attachment power α, the five datasets analyzed fell into three categories (fig. 5). First, we estimated a preferential attachment power close to 1, indicating linear preferential attachment, for the outbreaks studied by Niculescu et al. [56] and Wang et al. [7]. These values were robust to specifying different priors for m. Both studies were of populations in which we would expect a high degree of epidemiological relatedness: Niculescu et al. [56] studied a recent outbreak among Romanian injection drug users (IDU), while Wang et al. sampled acutely infected MSM in Beijing, China. Both these are contexts in which we would expect some of the assumptions of the BA model, such as a connected network, relatively high mean degree, and preferential attachment dynamics, to hold.

The remaining three datasets (Novitsky et al. [21], Li et al. [62], and Cuevas et al. [63]) had estimated values of α below 0.5 when $m = 1$ was included in the prior, but these were not robust to changing the prior to exclude $m = 1$. For the Cuevas et al. data, model misspecification is likely partially responsible. While the authors found that a large proportion of the samples were epidemiologically linked, these were mainly in small local clusters rather than the single large component postulated by the BA model. In addition, the mixed risk groups in the dataset would be unlikely to significantly interact, further weakening any global preferential attachment dynamics. The dataset studied by Novitsky et al. [21] originated from a densely sampled population where the predominant risk factor was believed to be heterosexual exposure. Although the MAP estimate of α was almost unchanged when the value $m = 1$ was excluded from the prior, the confidence interval shrank significantly. For both priors, the estimated I value was extremely high, in fact higher than the estimated HIV prevalence in the sampled region. The authors indicated that the source of the samples was a town in close proximity to the country's capital city, and suggested that there may have been a high degree of migration and partner interchange between the two locations. It is possible that the contact network underlying the subtree we investigated includes a much larger group based in the capital city, which would explain the high estimate of I. There is no clear explanation for the discrepancy between the two priors for the Li et al. [62] data, as the subset we analyzed formed a phylogenetic cluster and therefore was a good candidate for the BA model. However, nearly all the posterior density was assigned to $m = 1$ when this value was allowed, indicating that the network was more likely to have an acyclic tree structure.

Preferential attachment power is sub-linear and higher for IDU networks

For all datasets we examined, the posterior mean estimates for α were sub-linear, ranging from 0.27 to 0.83. The sub-linearity is consistent with the results of de Blasio, Svensson, and Liljeros [77], who developed a statistical inference method to estimate the parameters of a more sophisticated

preferential attachment model incorporating heterogeneous node behaviour. When used to analyze population-level longitudinal partner count data, they found α values ranging from 0.26 to 0.62 depending on the gender and time period considered.

Both de Blasio, Svensson, and Liljeros [77] and the HET/Botswana data studied populations whose primary risk factor for HIV infection was heterosexual contact. de Blasio, Svensson, and Liljeros explicitly excluded reported homosexual contacts; Novitsky et al. did not, but noted that heterosexual contact is the primary mode of transmission in Botswana where the study was done. In the first of the two papers describing the Botswana study [57], the authors noted that their sample was gender-biased, being composed of approximately 75% women. Our estimate of α for these data was 0.55 or 0.53, depending on the prior on m. Similarly, de Blasio, Svensson, and Liljeros [77] estimated 0.54, 0.57, and 0.29 for 3-year, 5-year, and lifetime partnership networks respectively for the female portion of their sample.

The datasets derived from IDU populations had a higher estimated preferential attachment power than the other datasets (figs. 5 and S12). This finding is in line with Dombrowski et al. [78], who reanalyzed a network of IDUs in Brooklyn, USA, collected between 1991 and 1993 [79]. They found that the IDU network resembled a BA network much more closely than other social and sexual networks, and offered sociological explanations for the apparent preferential attachment dynamics in this population. Importantly, from a public health perspective, the authors asserted that the removal of *random* individuals from IDU networks may have the paradoxical effect of increasing the network's epidemic susceptibility. When low-degree nodes are removed, as would occur during a police crackdown, their network neighbours may turn to well-known community members for advice or supplies, thus increasing the connectivity of these high-degree nodes.

Unfortunately, the sub-linear region for α identified by both de Blasio, Svensson, and Liljeros [77] and *netabc* is also the region of poorest identifiability (fig. 2). This was reflected in the high level of uncertainty in the estimates, with most 95% HPD intervals covering the majority of the range [0, 1]. The value $\alpha = 0.5$ was contained in the 95% HPD interval for every dataset; consequently, it is not possible to say with high confidence that any of the α values are different from each other. In synthetic data, the confidence intervals around α narrowed when other parameters were marginalized out (fig. S8). Thus, it is possible that estimates of α could be made more precise by specifying either exact values or informative priors on the other BA parameters when these are known.

Other BA parameters

The true HIV prevalence in a population can be difficult to estimate for several reasons. HIV-infected individuals may be asymptomatic for months or years, possibly delaying their

awareness of their status. In many contexts, the risk factors for acquisition of HIV are illegal or stigmatized, which may represent a barrier to testing, treatment, and/or disclosure of status. Our simulation study showed that I is weakly identifiable from tree shapes, however, the estimates of I obtained with *netabc* were upwardly biased (fig. 2). In addition, our initial exploratory analysis showed that the identifiability of I decreases with the number of sampled tips (fig. S5); in real world studies, the proportion of infected individuals sampled is usually low. The estimated I values for the HIV datasets ranged from 373 (IDU/Estonia) to 5432 (HET/Botswana).

We were not able to discern and trends toward over- or underestimation of I from available prevalence data. For example, the authors of the HET/Botswana data [21, 57] estimated that there were 1731 HIV-positive individuals in the study area. HIV sequences were obtained from approximately 70% of these individuals. The estimated prevalence we obtained was much higher (5432), with a 95% HPD interval spanning nearly the entire prior region. On the other hand, the study which produced the MSM/USA data [61] enrolled 648 HIV-positive MSM; thus, our estimate of I (482) was clearly an underestimate. Post-hoc explanations can be imagined for both of these results. The HET/Botswana data were collected from a town located proximally to the country's capital city; the authors suggested that frequent travel between the two locations may have facilitated linking of their sexual networks. Thus, the high estimate of I we obtained for these data may include a larger network component based in the capital city. The MSM/USA result may indicate that the individuals genotyped for the study are members of a smaller subnetwork which does not include the entire local MSM population. Unfortunately, none of these hypotheses can be easily tested.

Over half of the datasets were estimated to have $m = 1$, which produces tree-like networks without cycles. Since the average degree of a Barabási-Albert network is $2m$, this value may simply reflect the fact that most people have a small number of sexual partners, especially when only recent partnerships are considered [23]. In fact, in one survey, the most common number of partnerships in the past twelve months was one [23]; the BA model does not allow any nodes with degree 1 when $m \geq 2$. For this reason, the choice of whether to allow $m = 1$ in the prior is problematic, as we must choose between an unrealistic topology (no cycles) and an unrealistic minimum degree. Extensions to the BA model which relax this constraint can be imagined and may offer improved parameter resolution. Estimates of N were not informative for any of the datasets under either choice of prior, consistent with our simulation results.

Modelling assumptions

In addition to the aforementioned possibility of misspecification, additional modelling assumptions include the network being connected and static, all transmission rates being equal, no

removal after infection, identical behaviour of all nodes, and random sampling. The last two were addressed with small-scale experiments. We simulated a network where some nodes exhibited a higher attachment power than others, and found that the estimated attachment power was simply the average of the two values. This indicated that, although we could characterize the network in aggregate, the estimated parameters could not be said to apply to any individual node. The effect of biased sampling was investigated by analyzing a transmission tree which had been sampled in a peer-driven fashion. The results were roughly in line with those for random sampling, however the estimated value of α was lower than the average for randomly-sampled trees. Further experiments would be necessary to fully explore the impact of these assumptions on the method's accuracy. However, despite these issues, we felt it was best to demonstrate the method first on a simple model. It is possible to use this framework to fit more complex models which address some of these issues, such as one incorporating heterogeneous node behaviour, which may prove a fruitful avenue for future investigations.

Our method has a number of caveats, perhaps the most significant being that it takes a transmission tree as input. In reality, true transmission trees are not available and must be approximated, often by way of a viral phylogeny. Although this has been demonstrated to be a fair approximation [e.g. 80], and is frequently used in practice [e.g. 81], the topologies of a viral phylogeny and transmission tree can differ significantly [82] due to within-host evolution and the sampling process [83]. The ABC-SMC algorithm is computationally intensive, taking about a day when run on 20 cores in parallel with the settings we described in the methods. Nevertheless, our method is potentially useful to epidemiological researchers interested in the general characteristics of the network structure underlying disease outbreaks. This work, and previous work by our group [41], has demonstrated that ABC is a broadly applicable and effective framework in which to perform phylodynamic inference.

Acknowledgements

We are grateful to Dr. Sally Otto, Dr. Alexandre Bouchard-Côté, Dr. Richard Harrigan, and the two anonymous reviewers for many helpful suggestions. This work was supported by grants from the Canadian Institutes of Health Research (CIHR, operating grant HOP-111406), and the Bill & Melinda Gates Foundation (award number OPP1110049). R.M.M. was supported by a scholarship from the CIHR Strategic Training Program in Bioinformatics. A.F.Y.P. was supported by a CIHR New Investigator Award (Canadian HIV Vaccine Initiative, Vaccine Discovery and Social Research) and by a Career Investigator Scholar Award from the Michael Smith Foundation for Health Research, in partnership with the Providence Health Care Research Institute and St. Paul's Hospital Foundation.

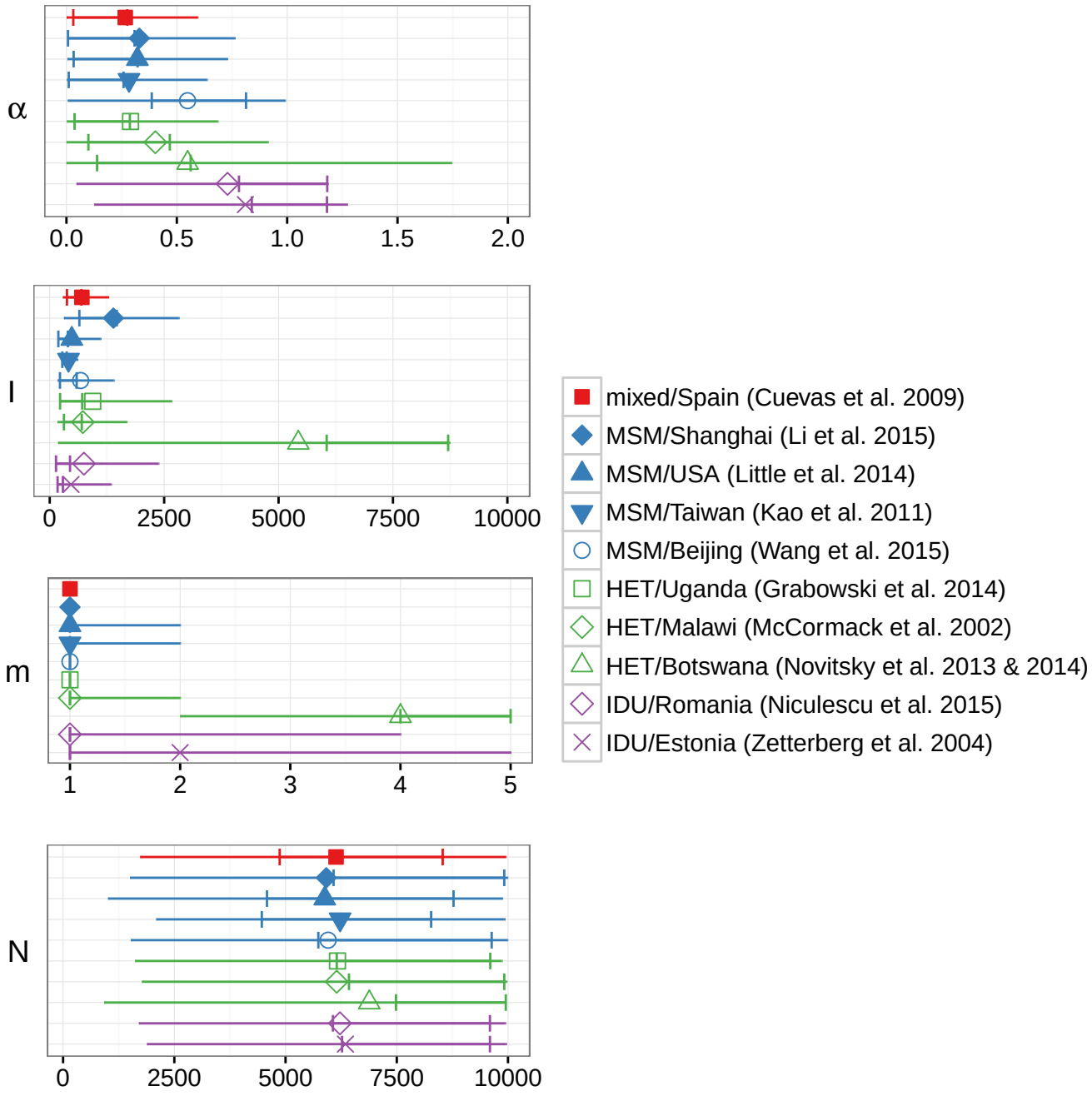

Figure 5: Posterior means (points), 50% HPD intervals (notches), and 95% HPD intervals (lines) for parameters of the BA network model, fitted to ten HIV datasets with *netabc*. Legend labels indicate risk group and country of origin. Abbreviations: IDU, injection drug users; MSM, men who have sex with men; HET, heterosexual. Note that posterior means can fall outside of the HPD interval if the distribution is diffuse.

References

[1] Alden S Klovdahl. “Social networks and the spread of infectious diseases: the AIDS example”. In: *Social Science & Medicine* 21.11 (1985), pp. 1203–1216.

[2] Martina Morris. “Epidemiology and social networks: modeling structured diffusion”. In: *Sociological Methods & Research* 22.1 (1993), pp. 99–126.

[3] Eamon B O’Dea and Claus O Wilke. “Contact heterogeneity and phylodynamics: how contact networks shape parasite evolutionary trees”. In: *Interdisciplinary Perspectives on Infectious Diseases* (2011), p. 238743.

[4] Junling Ma, P van den Driessche, and Frederick H Willeboordse. “The importance of contact network topology for the success of vaccination strategies”. In: *Journal of Theoretical Biology* 325 (2013), pp. 12–21.

[5] Marc Barthélemy et al. “Dynamical patterns of epidemic outbreaks in complex heterogeneous networks”. In: *Journal of Theoretical Biology* 235.2 (2005), pp. 275–288.

[6] Steven M Goodreau. “Assessing the effects of human mixing patterns on human immunodeficiency virus-1 interhost phylogenetics through social network simulation”. In: *Genetics* 172.4 (2006), pp. 2033–2045.

[7] Xicheng Wang et al. “Targeting HIV prevention based on molecular epidemiology among deeply sampled subnetworks of men who have sex with men”. In: *Clinical Infectious Diseases* 61.9 (2015), p. 1462.

[8] David Welch, Shweta Bansal, and David R Hunter. “Statistical inference to advance network models in epidemiology”. In: *Epidemics* 3.1 (2011), pp. 38–45.

[9] K Eames et al. “Six challenges in measuring contact networks for use in modelling”. In: *Epidemics* 10 (2015), pp. 72–77.

[10] Alexei J Drummond et al. “Measurably evolving populations”. In: *Trends in Ecology & Evolution* 18.9 (2003), pp. 481–488.

[11] Bryan T Grenfell et al. “Unifying the epidemiological and evolutionary dynamics of pathogens”. In: *Science* 303.5656 (2004), pp. 327–332.

[12] Tanja Stadler et al. “Estimating the basic reproductive number from viral sequence data”. In: *Molecular Biology and Evolution* 29.1 (2012), pp. 347–357.

[13] Gareth J Hughes et al. “Molecular phylodynamics of the heterosexual HIV epidemic in the United Kingdom”. In: *PLoS Pathogens* 5.9 (2009), e1000590.

[14] Erik M Volz et al. “Simple epidemiological dynamics explain phylogenetic clustering of HIV from patients with recent infection”. In: *PLoS Computational Biology* 8.6 (2012), e1002552.

[15] Erik M Volz. “Complex population dynamics and the coalescent under neutrality”. In: *Genetics* 190.1 (2012), pp. 187–201.

[16] David A Rasmussen, Erik M Volz, and Katia Koelle. “Phylodynamic inference for structured epidemiological models”. In: *PLoS Computational Biology* 10.4 (2014), e1003570.

[17] Gabriel E Leventhal et al. “Inferring epidemic contact structure from phylogenetic trees”. In: *PLoS Computational Biology* 8.3 (2012), e1002413.

[18] Caroline Colijn and Jennifer Gardy. “Phylogenetic tree shapes resolve disease transmission patterns”. In: *Evolution, Medicine, and Public Health* 2014.1 (2014), pp. 96–108.

[19] Katy Robinson et al. “How the dynamics and structure of sexual contact networks shape pathogen phylogenies”. In: *PLoS Computational Biology* 9.6 (2013), e1003105.

[20] Luc Villandre et al. “Assessment of overlap of phylogenetic transmission clusters and communities in simple sexual contact networks: applications to HIV-1”. In: *PloS ONE* 11.2 (2016), e0148459.

[21] Vlad Novitsky et al. “Impact of sampling density on the extent of HIV clustering”. In: *AIDS Research and Human Retroviruses* 30.12 (2014), pp. 1226–1235.

[22] Stirling A Colgate et al. “Risk behavior-based model of the cubic growth of acquired immunodeficiency syndrome in the United States”. In: *Proceedings of the National Academy of Sciences* 86.12 (1989), pp. 4793–4797.

[23] Fredrik Liljeros et al. “The web of human sexual contacts”. In: *Nature* 411.6840 (2001), pp. 907–908.

[24] Anne Schneeberger et al. “Scale-free networks and sexually transmitted diseases: a description of observed patterns of sexual contacts in Britain and Zimbabwe”. In: *Sexually Transmitted Diseases* 31.6 (2004), pp. 380–387.

[25] Vito Latora et al. “Network of sexual contacts and sexually transmitted HIV infection in Burkina Faso”. In: *Journal of Medical Virology* 78.6 (2006), pp. 724–729.

[26] Richard Rothenberg and Stephen Q Muth. “Large-network concepts and small-network characteristics: fixed and variable factors”. In: *Sexually Transmitted Diseases* 34.8 (2007), pp. 604–612.

[27] Stéphan Clémençon et al. “A statistical network analysis of the HIV/AIDS epidemics in Cuba”. In: *Social Network Analysis and Mining* 5.1 (2015), pp. 1–14.

[28] Mark S Handcock and James Holland Jones. “Likelihood-based inference for stochastic models of sexual network formation”. In: *Theoretical Population Biology* 65.4 (2004), pp. 413–422.

[29] Albert-László Barabási and Réka Albert. “Emergence of scaling in random networks”. In: *Science* 286.5439 (1999), pp. 509–512.

[30] Paul L Krapivsky, Sidney Redner, and Francois Leyvraz. “Connectivity of growing random networks”. In: *Physical Review Letters* 85.21 (2000), p. 4629.

[31] Tom Britton and Philip D O’Neill. “Bayesian inference for stochastic epidemics in populations with random social structure”. In: *Scandinavian Journal of Statistics* 29.3 (2002), pp. 375–390.

[32] Paul Erdős and Alfred Rényi. “On the evolution of random graphs”. In: *Publications of the Mathematical Institute of the Hungarian Academy of Sciences* 5 (1960), pp. 17–61.

[33] Chris Groendyke, David Welch, and David R Hunter. “Bayesian inference for contact networks given epidemic data”. In: *Scandinavian Journal of Statistics* 38.3 (2011), pp. 600–616.

[34] Erik Volz and Lauren Ancel Meyers. “Susceptible–infected–recovered epidemics in dynamic contact networks”. In: *Proceedings of the Royal Society of London B: Biological Sciences* 274.1628 (2007), pp. 2925–2934.

[35] Erik Volz. “SIR dynamics in random networks with heterogeneous connectivity”. In: *Journal of Mathematical Biology* 56.3 (2008), pp. 293–310.

[36] Andrew J Leigh Brown et al. “Transmission network parameters estimated from HIV sequences for a nationwide epidemic”. In: *The Journal of Infectious Diseases* 204.9 (2011), p. 1463.

[37] Simon Tavaré et al. “Inferring coalescence times from DNA sequence data”. In: *Genetics* 145.2 (1997), pp. 505–518.

[38] Mark A Beaumont, Wenyang Zhang, and David J Balding. “Approximate Bayesian computation in population genetics”. In: *Genetics* 162.4 (2002), pp. 2025–2035.

[39] Mikael Sunnåker et al. “Approximate Bayesian computation”. In: *PLoS Computational Biology* 9.1 (2013), e1002803.

[40] Shigeki Nakagome, Kenji Fukumizu, and Shuhei Mano. “Kernel approximate Bayesian computation in population genetic inferences”. In: *Statistical Applications in Genetics and Molecular Biology* 12.6 (2013), pp. 667–678.

[41] Art FY Poon. “Phylodynamic inference with kernel ABC and its application to HIV epidemiology”. In: *Molecular Biology and Evolution* 32.9 (2015), pp. 2483–2495.

[42] Mijung Park et al. “K2-ABC: Approximate Bayesian Computation with Kernel Embeddings”. In: *stat* 1050 (2015), p. 24.

[43] Art FY Poon et al. “Mapping the shapes of phylogenetic trees from human and zoonotic RNA viruses”. In: *PLoS ONE* 8.11 (2013), e78122.

[44] Pierre Del Moral, Arnaud Doucet, and Ajay Jasra. “An adaptive sequential Monte Carlo method for approximate Bayesian computation”. In: *Statistics and Computing* 22.5 (2012), pp. 1009–1020.

[45] Daniel T Gillespie. “A general method for numerically simulating the stochastic time evolution of coupled chemical reactions”. In: *Journal of Computational Physics* 22.4 (1976), pp. 403–434.

[46] Scott A Sisson, Yanan Fan, and Mark M Tanaka. “Sequential Monte Carlo without likelihoods”. In: *Proceedings of the National Academy of Sciences* 104.6 (2007), pp. 1760–1765.

[47] Mark A Beaumont et al. “Adaptive approximate Bayesian computation”. In: *Biometrika* 96.4 (2009), pp. 983–990.

[48] Gabor Csardi and Tamas Nepusz. “The igraph software package for complex network research”. In: *InterJournal, Complex Systems* 1695.5 (2006), pp. 1–9.

[49] Kwang-Tsao Shao. “Tree balance”. In: *Systematic Biology* 39.3 (1990), pp. 266–276.

[50] Thijs Janzen, Sebastian Höhna, and Rampal S Etienne. “Approximate Bayesian computation of diversification rates from molecular phylogenies: introducing a new efficient summary statistic, the nLTT”. In: *Methods in Ecology and Evolution* 6.5 (2015), pp. 566–575.

[51] Achim Zeileis et al. “kernlab-an S4 package for kernel methods in R”. In: *Journal of Statistical Software* 11.9 (2004), pp. 1–20.

[52] David Meyer et al. *e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien*. R package version 1.6-7. 2015.

[53] Martyn Plummer et al. “CODA: Convergence diagnosis and output analysis for MCMC”. In: *R News* 6.1 (2006), pp. 7–11.

[54] Sture Holm. “A simple sequentially rejective multiple test procedure”. In: *Scandinavian Journal of Statistics* (1979), pp. 65–70.

[55] Veera Zetterberg et al. “Two viral strains and a possible novel recombinant are responsible for the explosive injecting drug use-associated HIV type 1 epidemic in Estonia”. In: *AIDS Research and Human Retroviruses* 20.11 (2004), pp. 1148–1156.

[56] Iulia Niculescu et al. “Recent HIV-1 outbreak among intravenous drug users in Romania: evidence for cocirculation of CRF14_BG and subtype F1 strains”. In: *AIDS Research and Human Retroviruses* 31.5 (2015), pp. 488–495.

[57] Vladimir Novitsky et al. “Phylogenetic relatedness of circulating HIV-1C variants in Mochudi, Botswana”. In: *PLoS ONE* 8.12 (2013), e80589.

[58] Grace P McCormack et al. “Early evolution of the human immunodeficiency virus type 1 subtype C epidemic in rural Malawi”. In: *Journal of Virology* 76.24 (2002), pp. 12890–12899.

[59] Mary K Grabowski et al. “The role of viral introductions in sustaining community-based HIV epidemics in rural Uganda: evidence from spatial clustering, phylogenetics, and egocentric transmission models”. In: *PLoS Medicine* 11.3 (2014), e1001610.

[60] Cheng-Feng Kao et al. “Surveillance of HIV type 1 recent infection and molecular epidemiology among different risk behaviors between 2007 and 2009 after the HIV type 1 CRF07_BC outbreak in Taiwan”. In: *AIDS Research and Human Retroviruses* 27.7 (2011), pp. 745–749.

[61] Susan J Little et al. “Using HIV networks to inform real time prevention interventions”. In: *PLoS ONE* 9.6 (2014), e98443.

[62] Xiaoyan Li et al. “HIV-1 genetic diversity and its impact on baseline CD4+ T cells and viral loads among recently infected men who have sex with men in Shanghai, China”. In: *PLoS ONE* 10.6 (2015), e0129559.

[63] MT Cuevas et al. “HIV-1 transmission cluster with T215D revertant mutation among newly diagnosed patients from the Basque Country, Spain”. In: *Journal of Acquired Immune Deficiency Syndromes* 51.1 (2009), p. 99.

[64] Peter JA Cock et al. “Biopython: freely available Python tools for computational molecular biology and bioinformatics”. In: *Bioinformatics* 25.11 (2009), pp. 1422–1423.

[65] Robert C Edgar. “MUSCLE: multiple sequence alignment with high accuracy and high throughput”. In: *Nucleic Acids Research* 32.5 (2004), pp. 1792–1797.

[66] Manolo Gouy, Stéphane Guindon, and Olivier Gascuel. “SeaView version 4: a multiplatform graphical user interface for sequence alignment and phylogenetic tree building”. In: *Molecular Biology and Evolution* 27.2 (2010), pp. 221–224.

[67] Morgan N Price, Paramvir S Dehal, and Adam P Arkin. “FastTree 2–approximately maximum-likelihood trees for large alignments”. In: *PloS ONE* 5.3 (2010), e9490.

[68] Simon Tavaré. “Some probabilistic and statistical problems in the analysis of DNA sequences”. In: *Lectures on Mathematics in the Life Sciences* 17 (1986), pp. 57–86.

[69] Alexei J Drummond and Andrew Rambaut. “BEAST: Bayesian evolutionary analysis by sampling trees”. In: *BMC Evolutionary Biology* 7.1 (2007), p. 214.

[70] Alexandros Stamatakis. “RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies”. In: *Bioinformatics* 30.9 (2014), p. 1312.

[71] Thu-Hien To et al. “Fast Dating Using Least-Squares Criteria and Algorithms”. In: *Systematic Biology* 65.1 (2016), p. 82.

[72] Art FY Poon et al. “The impact of clinical, demographic and risk factors on rates of HIV transmission: a population-based phylogenetic analysis in British Columbia, Canada”. In: *The Journal of Infectious Diseases* 211.6 (2015), pp. 926–935.

[73] John Frank Charles Kingman. “The coalescent”. In: *Stochastic Processes and their Applications* 13.3 (1982), pp. 235–248.

[74] Erik M Volz et al. “Phylodynamics of infectious disease epidemics”. In: *Genetics* 183.4 (2009), pp. 1421–1430.

[75] Jarno Lintusaari et al. “On the identifiability of transmission dynamic models for infectious diseases”. In: *Genetics* (2016).

[76] James Holland Jones and Mark S Handcock. “An assessment of preferential attachment as a mechanism for human sexual network formation”. In: *Proceedings of the Royal Society of London B: Biological Sciences* 270.1520 (2003), pp. 1123–1128.

[77] Birgitte Freiesleben de Blasio, Åke Svensson, and Fredrik Liljeros. “Preferential attachment in sexual networks”. In: *Proceedings of the National Academy of Sciences* 104.26 (2007), pp. 10762–10767.

[78] Kirk Dombrowski et al. “Topological and historical considerations for infectious disease transmission among injecting drug users in bushwick, Brooklyn (USA)”. In: *World Journal of AIDS* 3.1 (2013), p. 1.

[79] Samuel R Friedman et al. *Social Networks, Drug Injectors’ Lives, and HIV/AIDS*. Springer Science & Business Media, 2006.

[80] Thomas Leitner et al. “Accurate reconstruction of a known HIV-1 transmission history by phylogenetic tree analysis”. In: *Proceedings of the National Academy of Sciences* 93.20 (1996), pp. 10864–10869.

[81] Tanja Stadler and Sebastian Bonhoeffer. “Uncovering epidemiological dynamics in heterogeneous host populations using phylogenetic methods”. In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 368.1614 (2013).

[82] Rolf JF Ypma, W Marijn van Ballegooijen, and Jacco Wallinga. “Relating phylogenetic trees to transmission trees of infectious disease outbreaks”. In: *Genetics* 195.3 (2013), pp. 1055–1062.

[83] Federica Giardina. “Inference of epidemic contact networks from HIV phylogenetic trees”. Oral presentation at HIV Dynamics and Evolution. 2016.

Supplementary Figures

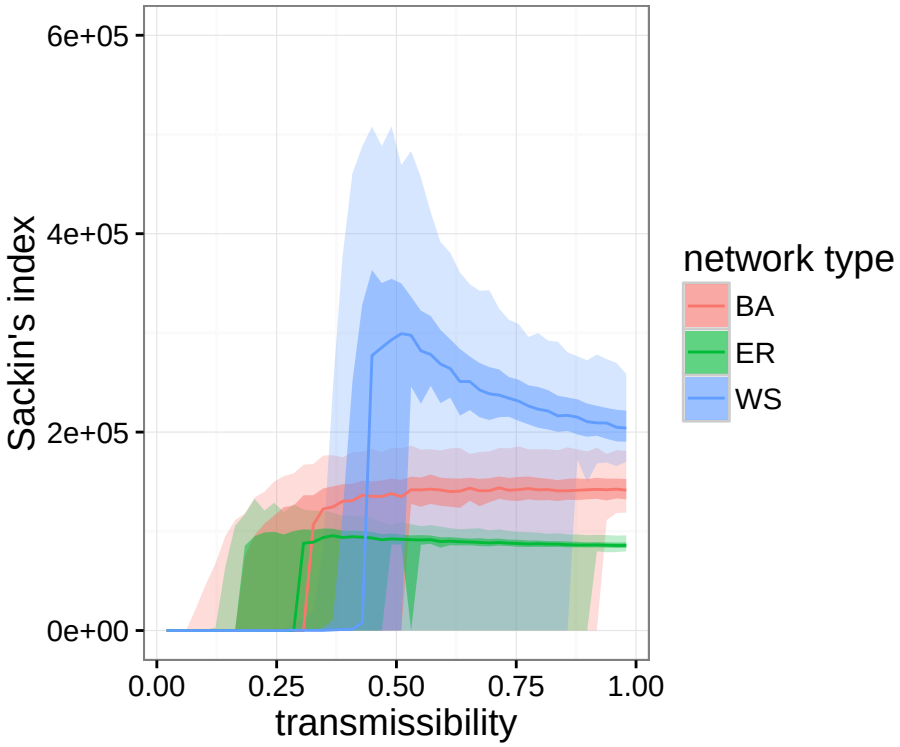

Figure S1: Reproduction of Figure 1A from Leventhal *et al.* (2012) used to check the accuracy of our implementation of Gillespie simulation. Transmission trees were simulated over three types of network, with pathogen transmissibility varying from 0 to 1 (1000 trees per transmissibility value). Sackin's index was calculated for each simulated transmission tree. Lines are means, light shaded areas are 95% quantile range, and dark shaded areas are interquartile ranges.

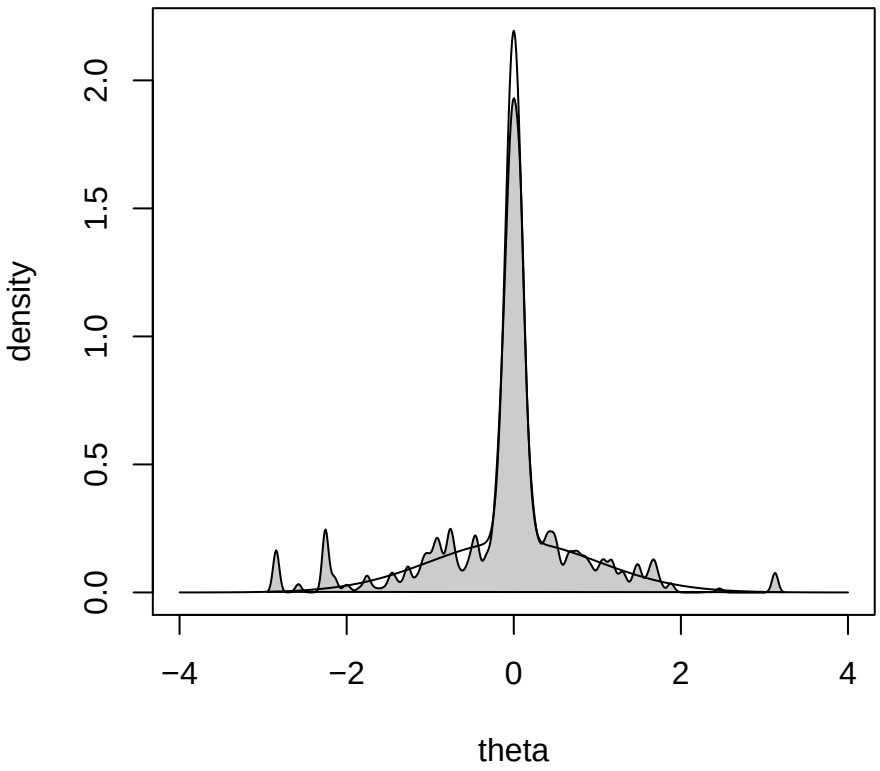

Figure S2: Approximation of mixture of Gaussians used by Del Moral *et al.* (2012) and Sisson *et al.* (2009) to test SMC. Solid black line indicates true distribution. Grey shaded area shows SMC approximation obtained with our implementation, using 10000 particles with one simulated data point per particle.

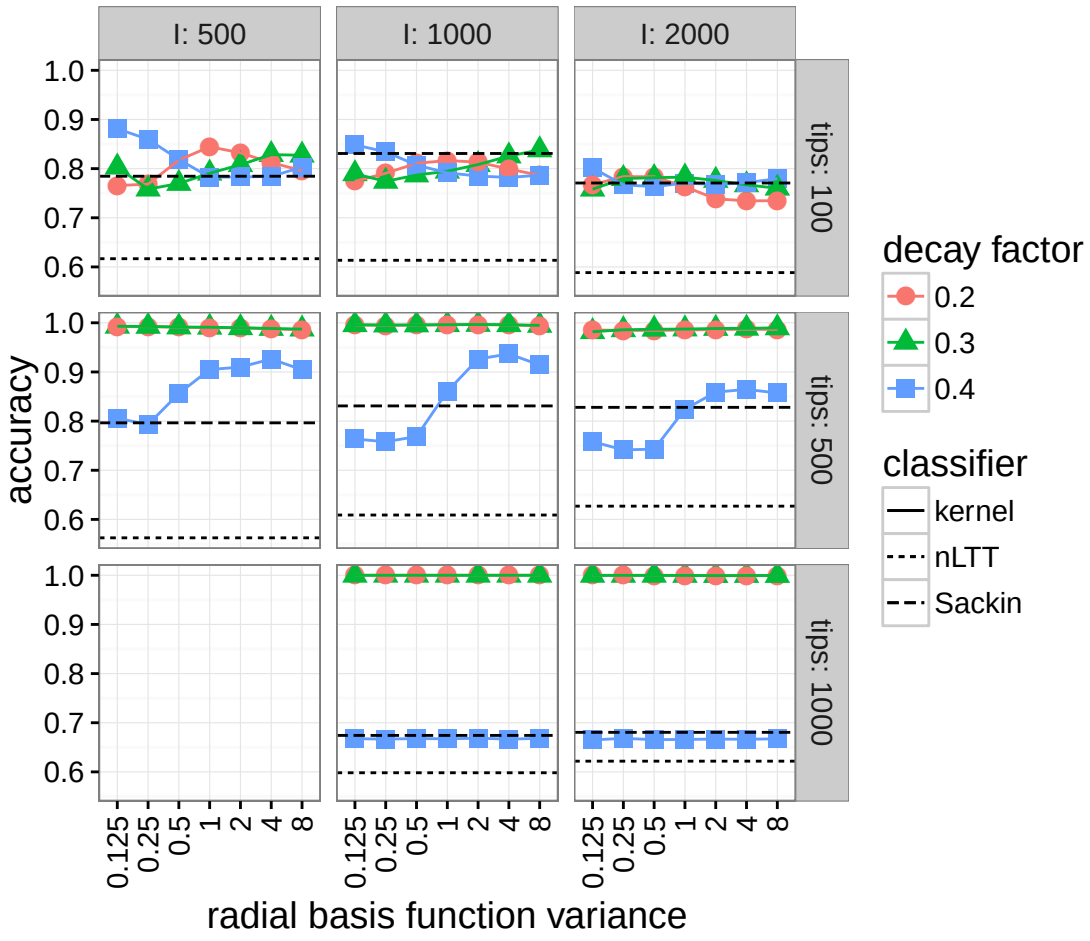

Figure S3: Cross-validation accuracy of kernel-SVM classifiers for α parameter of BA network model, for various tree kernel meta-parameters and epidemic scenarios. Each point was calculated based on 300 simulated transmission trees over networks with $\alpha = 0.5$, 1.0, or 1.5. Dotted and and dashed lines indicate, respectively, performance of SVMs using the nLTT statistic and Sackin's index. Facets are number of infected nodes before the simulation was stopped (I) and number of tips in the sampled transmission tree.

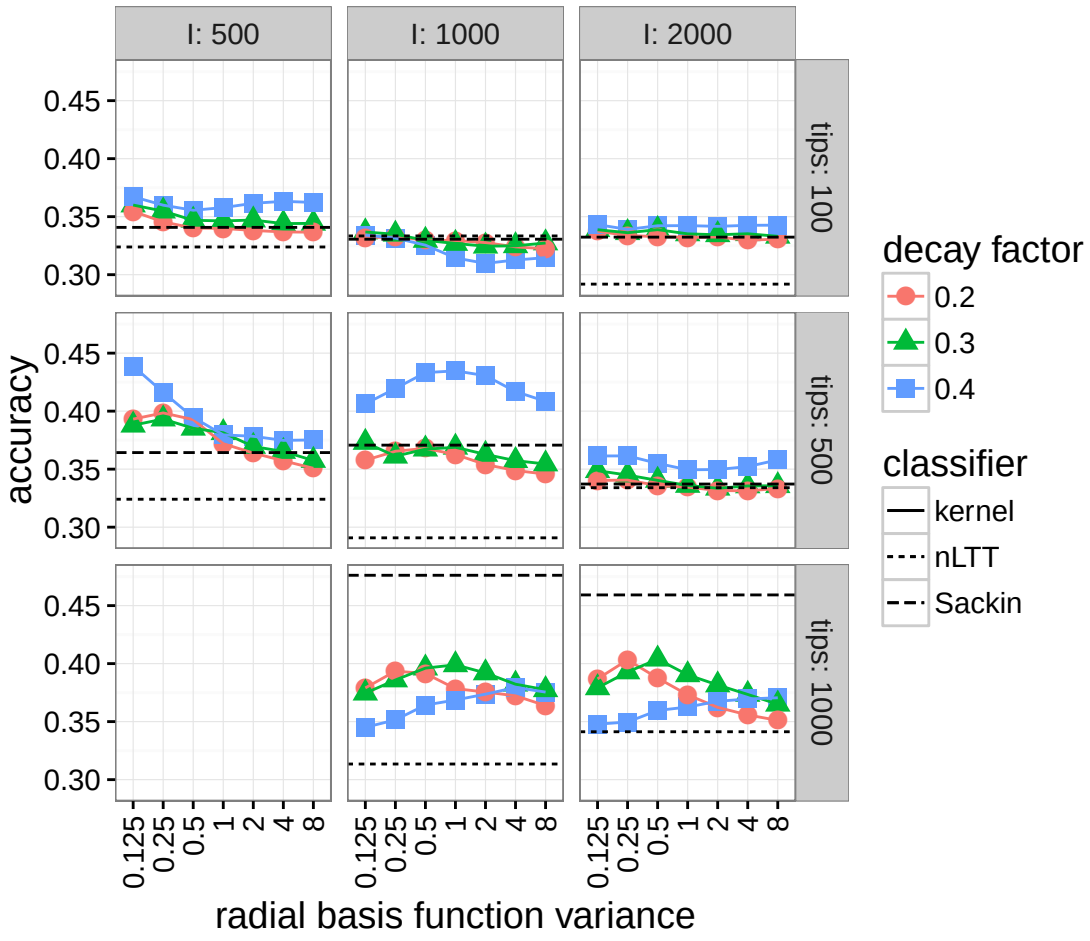

Figure S4: Cross-validation accuracy of kernel-SVM classifiers for m parameter of BA network model, for various tree kernel meta-parameters and epidemic scenarios. Each point was calculated based on 300 simulated transmission trees over networks with m = 2, 3, or 4. Dotted and and dashed lines indicate, respectively, performance of SVMs using the nLTT statistic and Sackin's index. Facets are number of infected nodes before the simulation was stopped (I) and number of tips in the sampled transmission tree.

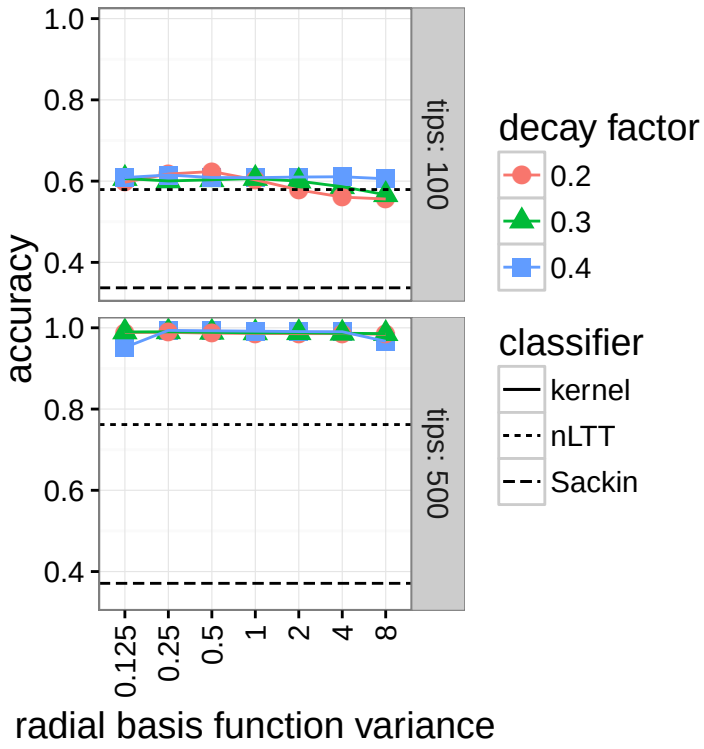

Figure S5: Cross-validation accuracy of kernel-SVM classifiers for number of infected nodes (I) under BA network model, for various tree kernel meta-parameters and two tree sizes. Each point was calculated based on 300 simulated transmission trees over networks with I = 500, 1000, or 2000. Dotted and and dashed lines indicate, respectively, performance of SVMs using the nLTT statistic and Sackin's index. Facets are the number of tips in the sampled transmission tree.

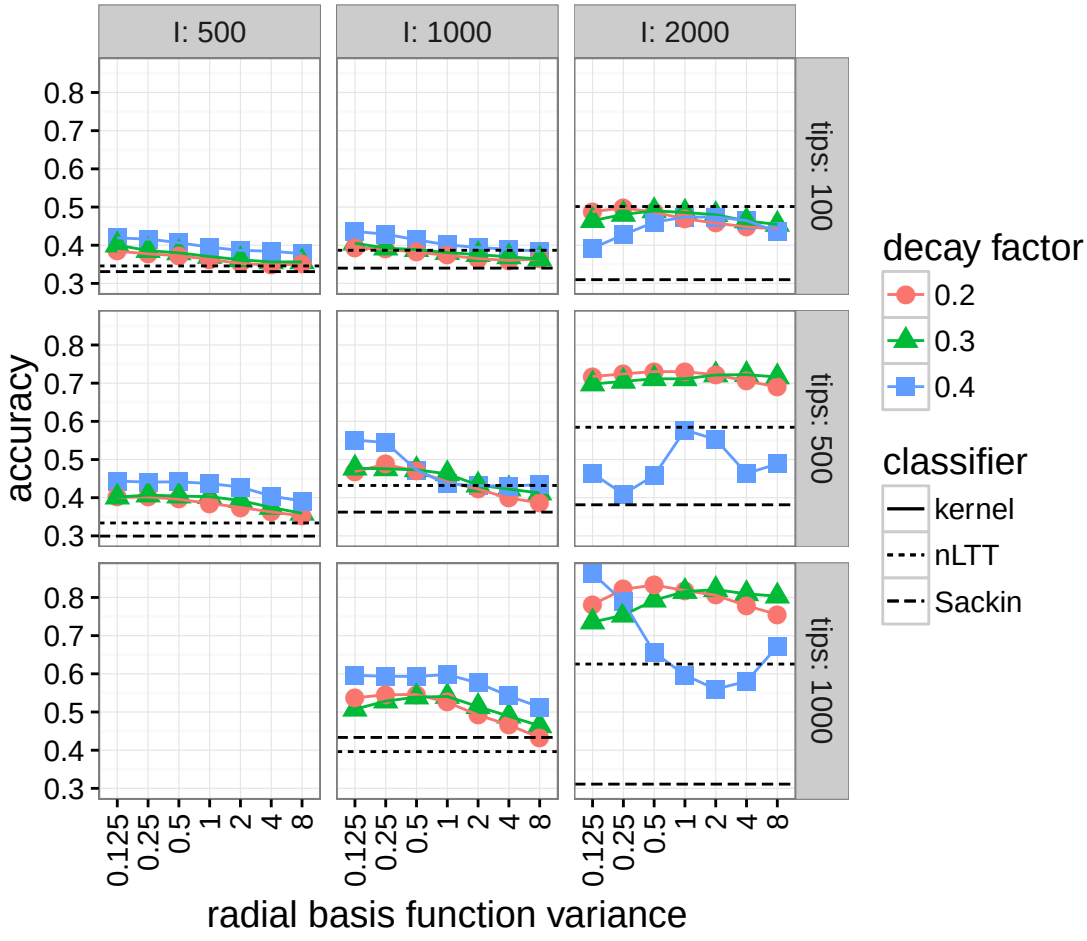

Figure S6: Cross-validation accuracy of kernel-SVM classifiers for total number of nodes (N) under BA network model, for various tree kernel meta-parameters and epidemic scenarios sizes. Each point was calculated based on 300 simulated transmission trees over networks with $N = 3000$, 5000, or 8000. Dotted and and dashed lines indicate, respectively, performance of SVMs using the nLTT statistic and Sackin's index. Facets are the number of tips in the sampled transmission tree.

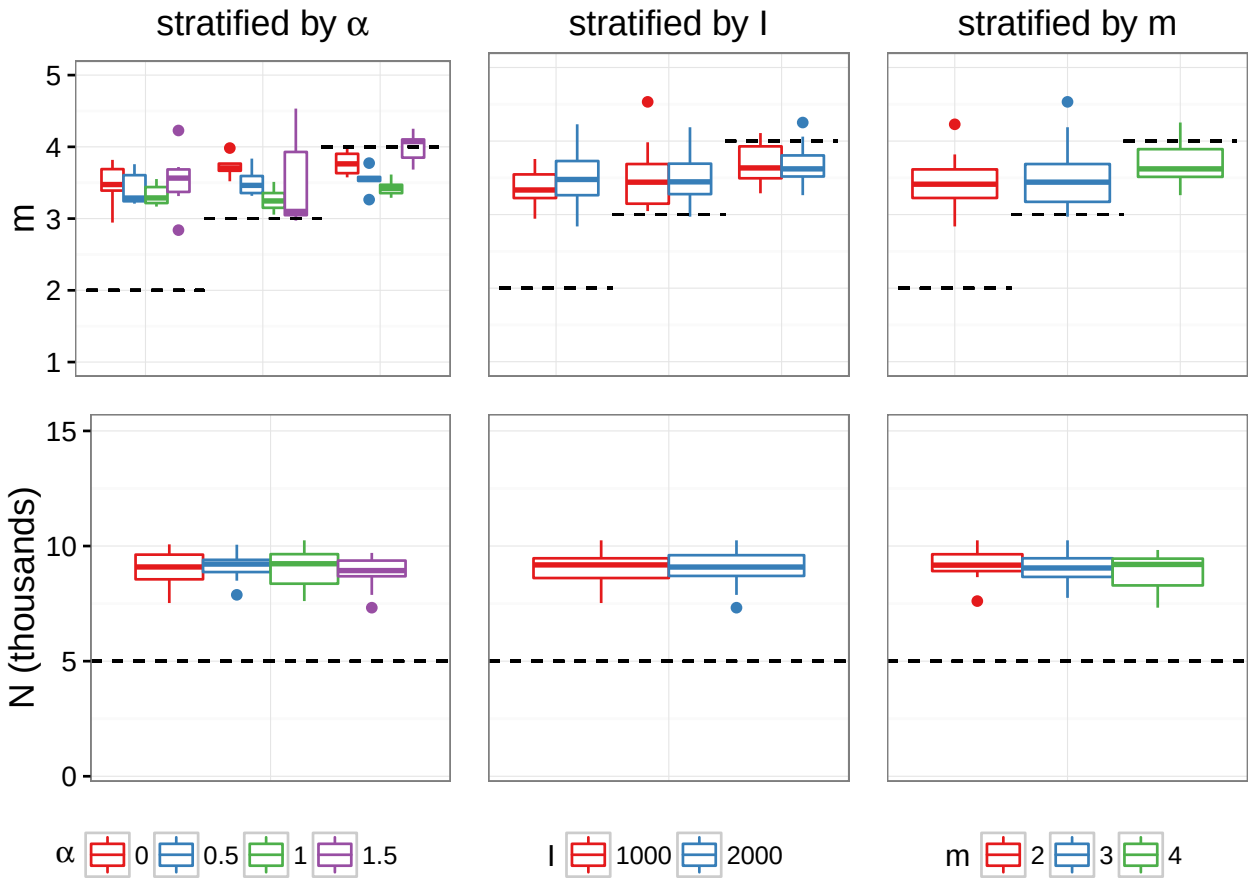

Figure S7: Posterior mean point estimates for BA model parameters m and N obtained by running *netabc* on simulated data, stratified by true parameter values. First row of plots contains true versus estimated values of m; second row contains true versus estimated values of N. Columns are stratified by α, I, and m respectively. Dashed lines indicate true values.

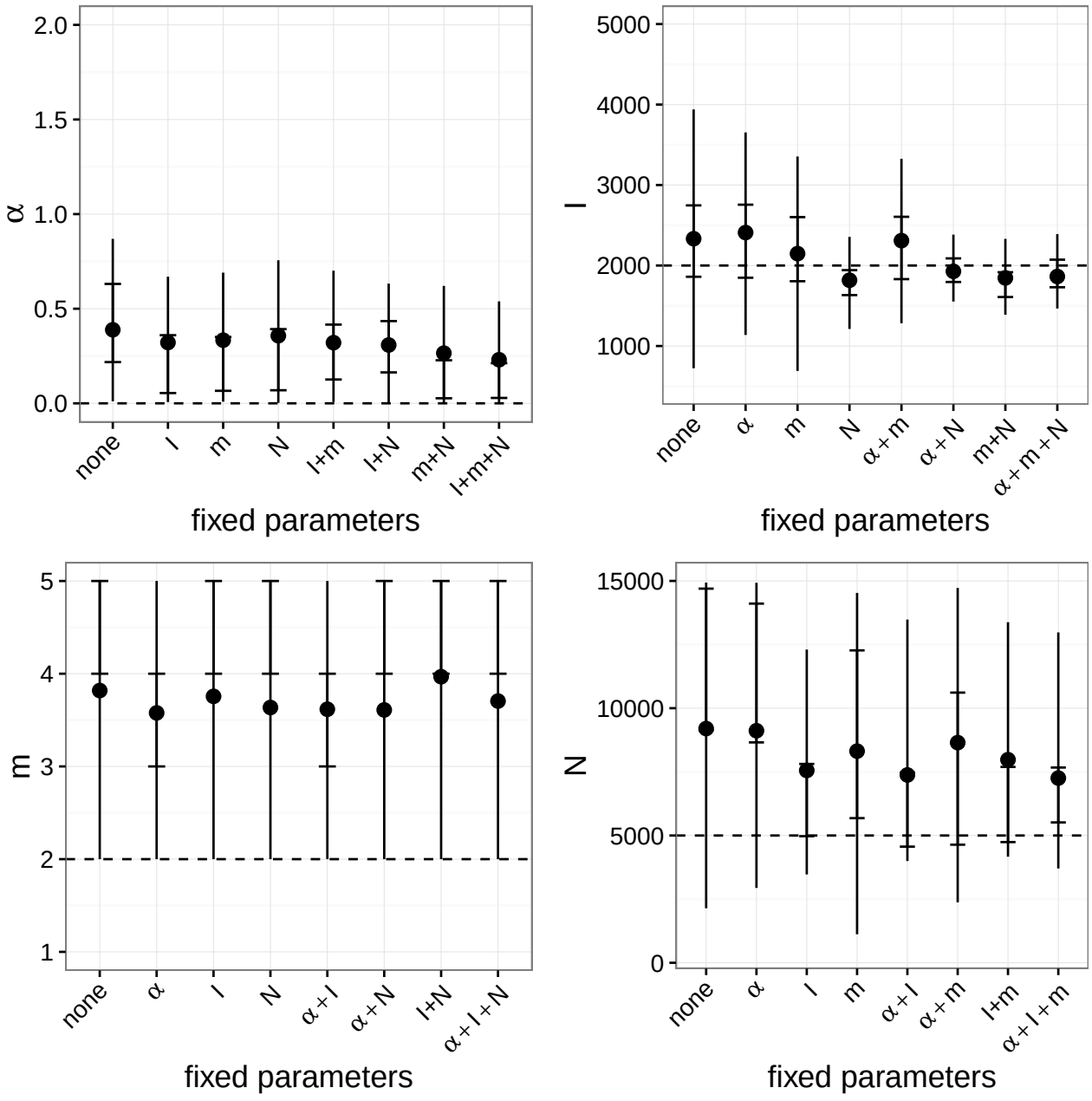

Figure S8: Posterior means (points), 50% HPD intervals (notches), and 95% HPD intervals (lines) for BA model parameters estimated marginally with ABC. *x*-axis labels indicate parameters which were fixed to their true values by specifying Dirac-delta priors.

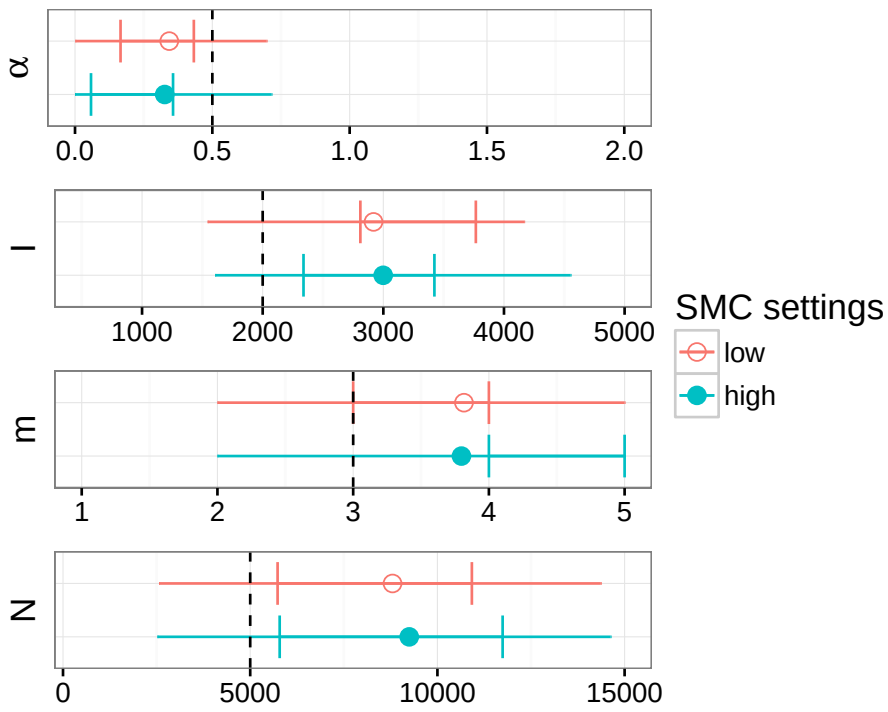

Figure S9: Posterior means (points), 50% HPD intervals (notches), and 95% HPD intervals (lines) for BA model parameter estimated with ABC using two sets of SMC settings. "Low" settings are 1000 particles, 5 simulated datasets per particle, and $\alpha_{\text{ESS}} = 0.95$. "High" settings are 2000 particles, 10 simulated datasets per particle, and $\alpha_{\text{ESS}} = 0.97$.

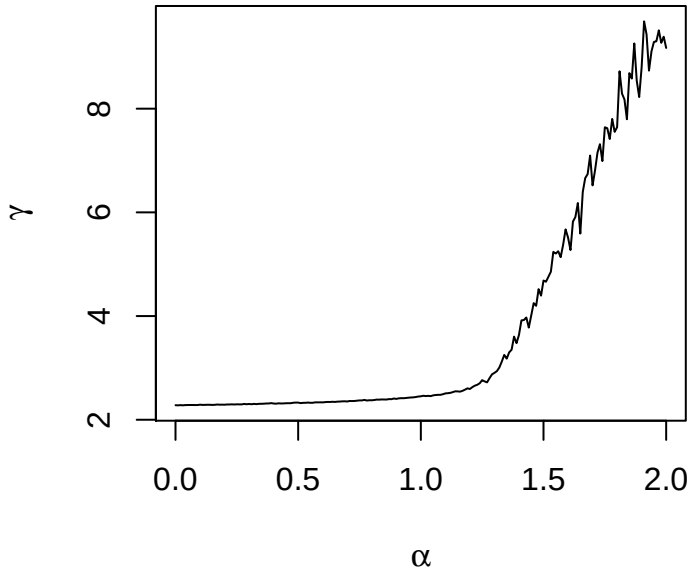

Figure S10: Relationship between preferential attachment power parameter α and power law exponent γ for networks simulated under the BA network model with $N = 5000$ and $m = 2$.

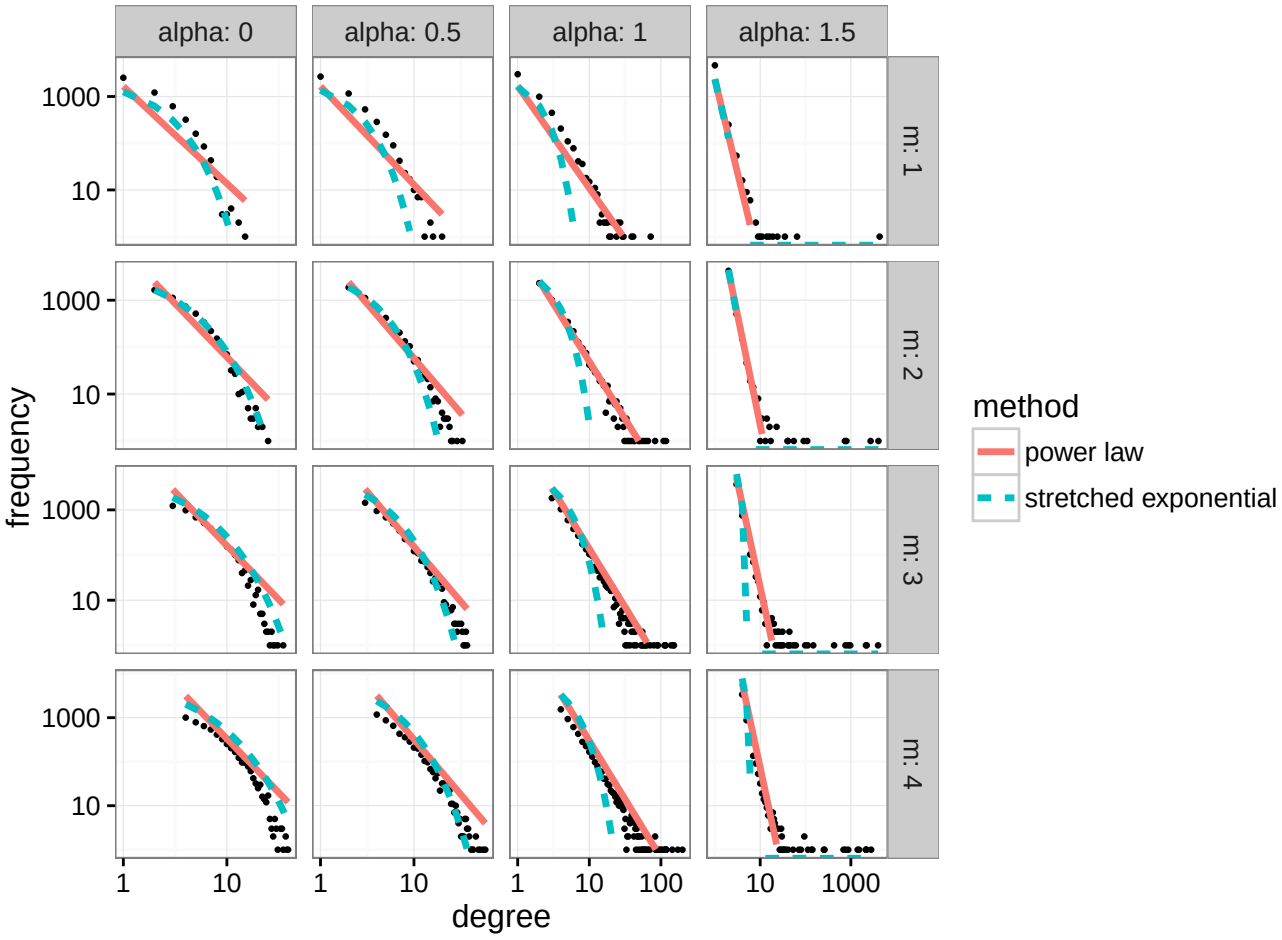

Figure S11: Best fit power law and stretched exponential curves for degree distributions of simulated BA networks for several values of α and m.

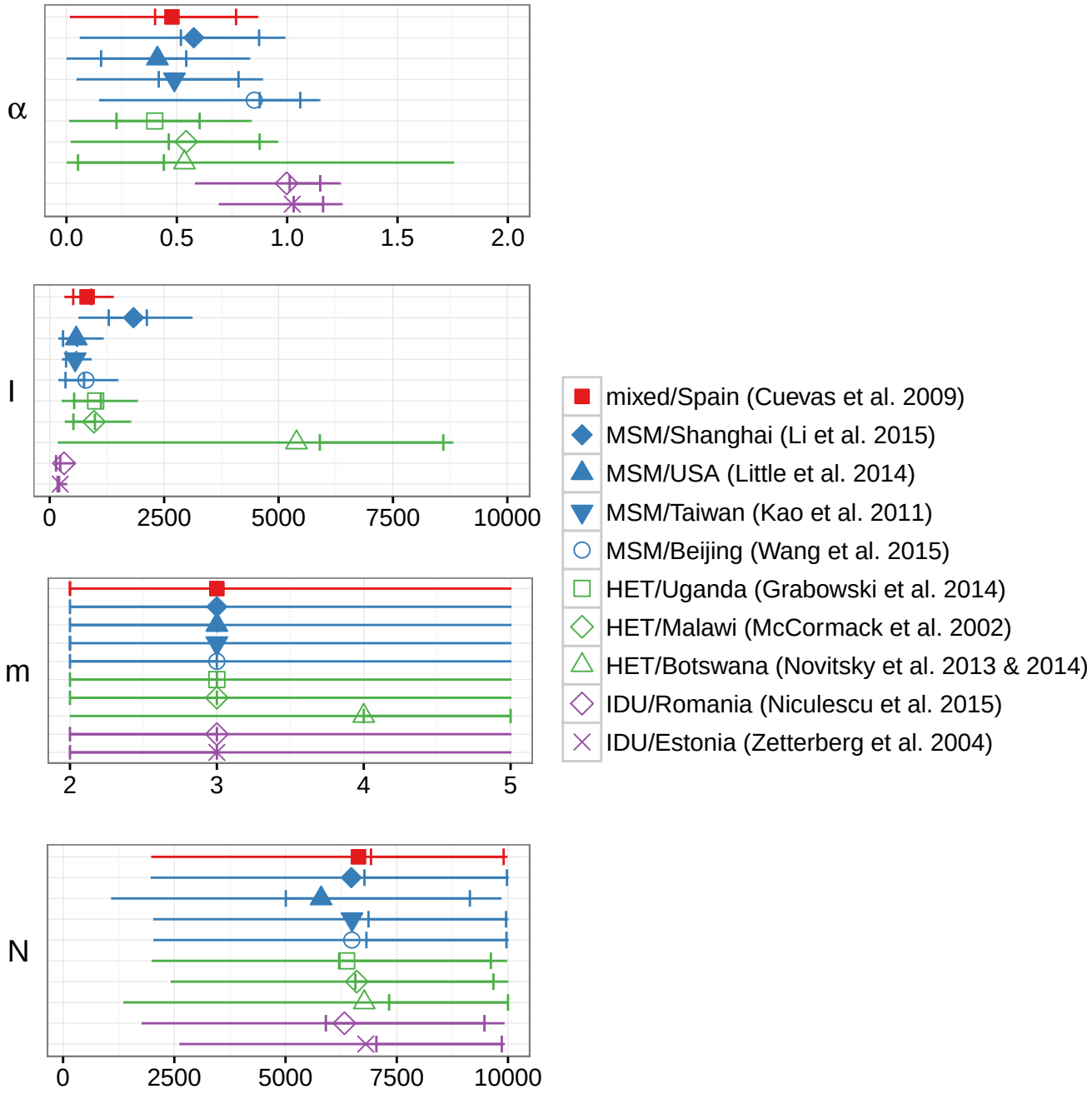

Figure S12: Maximum *a posteriori* point estimates and 95% HPD intervals for parameters of the BA network model, fitted to five published HIV datasets with ABC. x-axes indicate regions of nonzero prior density. In particular, the prior on m was DiscreteUniform(2, 5).

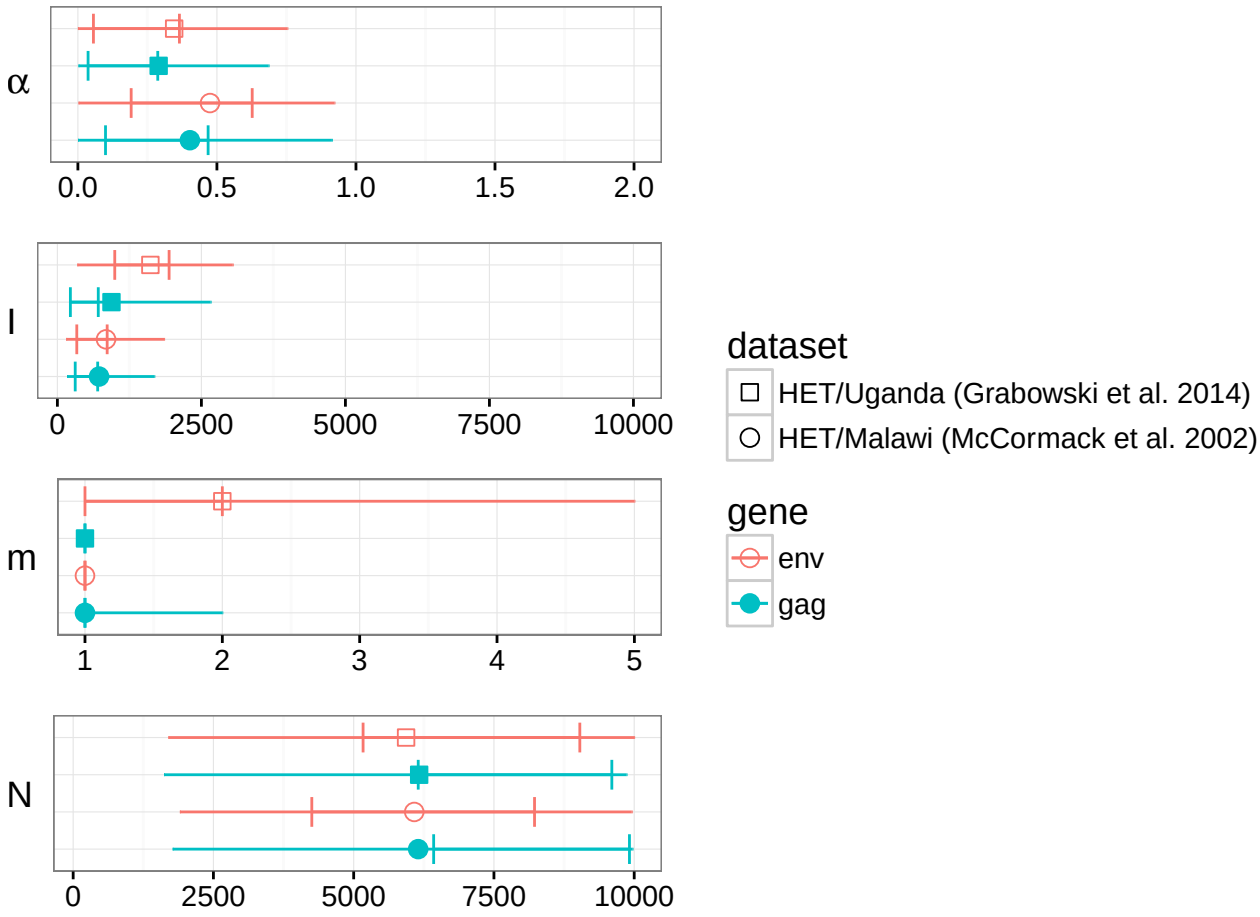

Figure S13: Posterior means (points), 50% HPD intervals (notches), and 95% HPD intervals (lines) for parameters of the BA network model, fitted to two HIV datasets where both *gag* and *env* genes were sequenced.

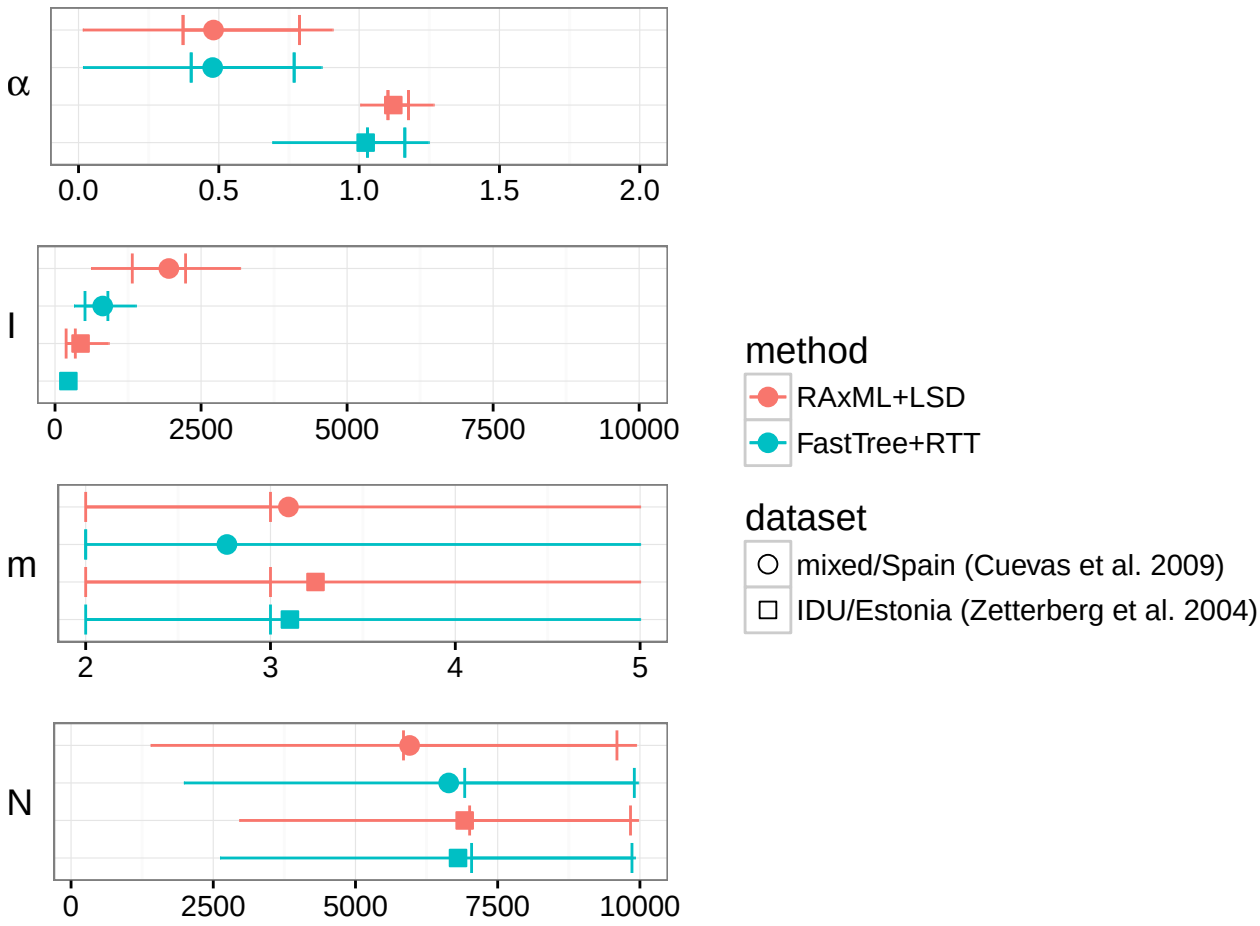

Figure S14: Posterior means (points), 50% HPD intervals (notches), and 95% HPD intervals (lines) for parameters of the BA network model, fitted to two HIV datasets using two phylogenetic reconstruction methods.

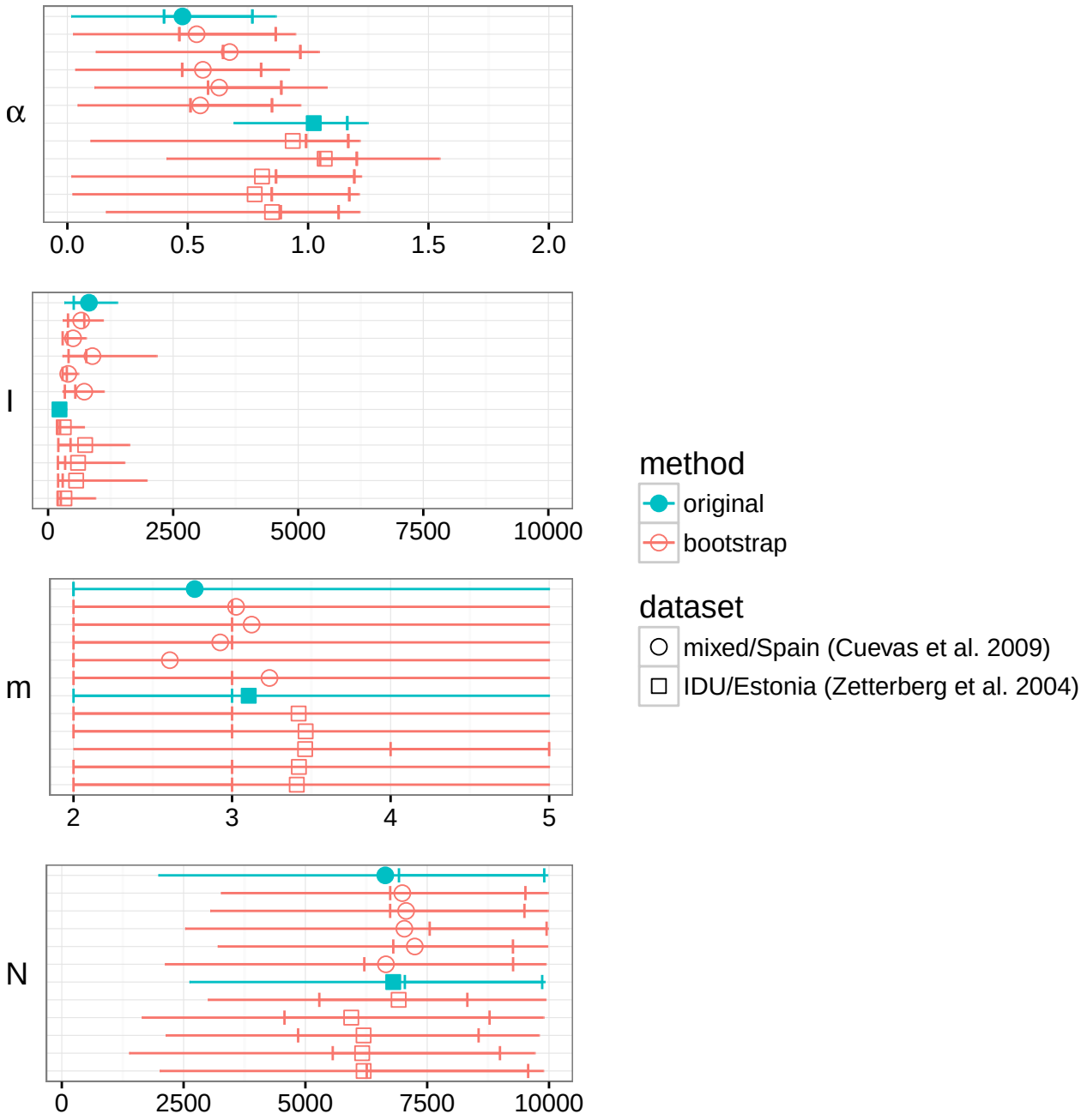

Figure S15: Posterior means (points), 50% HPD intervals (notches), and 95% HPD intervals (lines) for parameters of the BA network model, fitted to original and bootstrap replicate alignments of two HIV datasets.

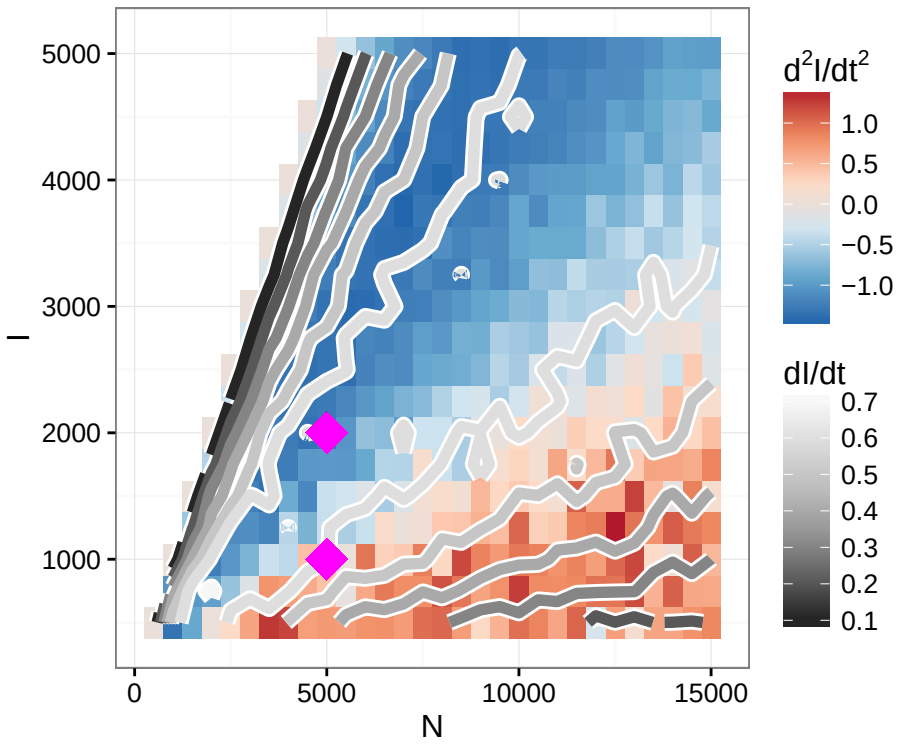

Figure S16: First and second time derivatives of epidemic growth curves at time of sampling for various values of I and I. Networks were simulated under the BA model with $\alpha = 1.0$, $m = 2$, and N varied along the values shown on the x-axis. Transmission trees were sampled at the time when I nodes were infected (y-axis). Logistic growth curves were fit to epidemic trajectories derived from the transmission trees, and their first and second derivatives were calculated at the time of sampling. Contours show first derivatives, while colours indicate second derivatives. Values of I and N used in simulation experiments with ABC are indicated by diamonds.