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

June 27, 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. 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 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. 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].

Survey-based studies of sexual networks [21–26] have found that these networks tend to have a degree distribution which follows a power law [although there has been some disagreement, see 27]. 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" [28]. 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 preferential attachment was introduced by Barabási and Albert [28], 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.

Previous work offers precedent for the possibility of statistical inference of structural network parameters. Britton and O'Neill [29] develop a Bayesian approach to estimate the edge density in an Erdős-Rényi network [30] given observed infection dates, and optionally recovery dates. Their approach was later extended by Groendyke, Welch, and Hunter [31] and applied to a much larger data set of 188 individuals. Volz and Meyers [32] and Volz [33] 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. [34] analysed 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) [35, 36]. 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 [37]. ~~In the specific case when ρ is a kernel function, the approach is known as ABC [38, 39].~~ Our group [39] and others [40] have demonstrated that taking ρ to be kernel function TODO

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. [41], 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. TODO: something about sub-linear PA and IDU

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 [42]. 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 [43], in the same fashion as several previous studies [*e.g.* 17, 19]. 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. [41]. 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 [37, 42].

In the original formulation of ABC-SMC [44, 45], 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 [42], 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. A computer program implementing our method is freely available at `https://github.com/rmcclosk/netabc` (last accessed June 27, 2016).

To check that our implementation of Gillespie simulation was correct, we reproduced Figure 1A of Leventhal et al. [17] (our **??**), which plots the unbalancedness 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 [44]). We were able to obtain a close approximation to the function (see **??**), 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 [46] to generate the graphs. The analyses were run on Westgrid (`https://www.westgrid.ca/`) and a local computer cluster.

~~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 two experiments designed to assess the identifiability of the BA parameters. These experiments only investigated one parameter of the BA model 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 [47] is a measure of tree imbalance which not take branch lengths into account, considering only the topology. The normalized lineages-through-time [nLTT, 48] compares two trees based on normalized distributions of their branching times, and does not explicitly consider the topology. In addition, 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. [41]). The results of this experiment were used to select values for these meta-parameters to carry forward based on their accuracy in distinguishing the different trees.

~~We used the phylogenetic kernel developed by Poon et al. [41]~~ 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 41],~~ 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 (kSVR) from the kernel matrix with the *kernlab* package [49], an ordinary SVR from the nLTT matrix with the e1071 package [50], and a linear regression from the Sackin's index values. The accuracy of each classifier was evaluated with 1000 two-fold cross validations.

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}, . . . ,

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 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 [47] and an ordinary SVR based on the normalized lineages-through-time (nLTT) statistic [48].~~

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 × 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 [42] and Beaumont et al. [45], 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* [51].

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 [52] 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 $\alpha_{\mathrm{ESS}} = 0.95$. The second run was performed with 2000 particles, 10 simulated transmission trees per particle, and $\alpha_{\mathrm{ESS}} = 0.99$. 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 (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. [53] data

Reference	Sequences (*n*)	Location	Risk group	Gene
Zetterberg et al. [53]	171/188	Estonia	IDU	*env*/*gag*
Niculescu et al. [54]	136	Romania	IDU	*pol*
Novitsky et al. [55]	180	Mochudi, Botswana	HET	*env*
Novitsky et al. [56]				
McCormack et al. [57]	141/154	Karonga District, Malawi	HET	*env*/*gag*
Grabowski et al. [58]	225	Rakai District, Uganda	HET	*env*/*gag*
Wang et al. [7]	173	Beijing, China	MSM	*pol*
Kao et al. [59]	275	Taiwan	MSM	*pol*
Little et al. [60]	180	San Fransisco, USA	MSM	*pol*
Li et al. [61]	280	Shanghai, China	MSM	*pol*
Cuevas et al. [62]	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).

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. [59] (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. Three datasets (IDU/Estonia, 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.

~~For the Novitsky et al. [56] 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+ [63]. Sequences were multiply aligned using *MUSCLE* version 3.8.31 [64], and alignments were manually inspected with *Seaview* version 4.4.2 [65]. Phylogenies were constructed from the nucleotide alignments by approximate maximum likelihood using *FastTree2* version 2.1.7 [66] with the generalized time-reversible (GTR) model [67]. 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 [68]) as described previously [39].

~~Two~~ Four of the datasets ~~[56, 61]~~ (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 [69]. 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 analysed ~~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 monopolistic 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 $\alpha \sim$ Uniform(0, 2) and N and I jointly uniform on the region $\{n \le N \le 10000, n \le I \le 10000, I \le N\}$, 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. 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 [28]: 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 (kSVR). We also tested classifiers based on Sackin's index [47] and the normalized lineages-through-time (nLTT) statistic [48]. The accuracy of each classifier on all BA parameters is shown in ~~varied based on the parameter being tested~~ 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 (**????????**). For all parameters except *m*, the absolute number of tips in the tree had a much greater impact on accuracy than the proportion of infected individuals these tips represented. However, for *m*, both the number and proportion of sampled tips had a strong impact or the accuracy of the kSVR (**??**).

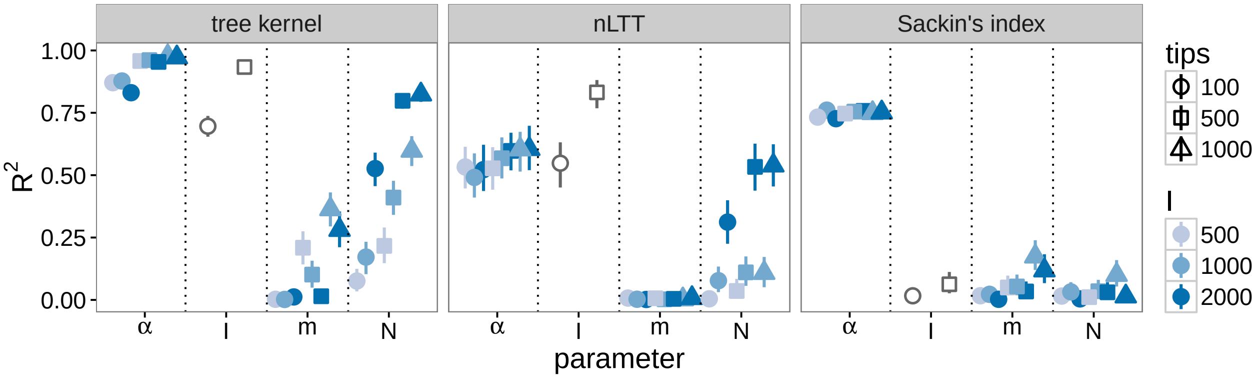

Figure 1: Cross-validation accuracy of kernel-SVR classifier (left), SVR classifier using nLTT (centre), and linear regression using Sackin's index (right) for BA model parameters. Kernel meta-parameters were set to $\lambda = 0.3$ and $\sigma = 4$. Each point was calculated based on 300 simulated transmission 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 accuracy of each classifier on all BA parameters is shown in fig. 1. Classifiers based on 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 (**????????**). For all parameters except *m*, the absolute number of tips in the tree had a much greater impact on accuracy than the proportion of infected individuals these tips represented. However, for *m*, both the number and proportion of sampled tips had a strong impact or the accuracy of the kSVR (**??**).

The kSVR classifier for α had an average R^2 of 0.92, compared to 0.56 for the nLTT-based SVR, and 0.75 for the linear regression against Sackin's index. There was little variation about the mean for different tree and epidemic sizes. No classifier could accurately identify the m parameter in any epidemic scenario, with average R^2 values of 0.12 for kSVR, 0.01 for the nLTT, and 0.06 for Sackin's index. Again, there was little variation in accuracy between epidemic scenarios, although the accuracy of the kSVR 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 R^2 values were 0.7, 0.55, and 0.02 for the tree kernel, nLTT, and Sackin's index respectively. For 500-tip trees, the values increased to 0.93, 0.83, and 0.07. Finally, the performance of classifiers for N depended heavily on the epidemic scenario. The R^2 of the kSVR classifier ranged from 0.08 for the smallest epidemic and smallest sample size, to 0.82 for the largest. Likewise, R^2 for the nLTT-based SVR ranged from 0.01 to 0.54. Sackin's index did not accurately classify N in any scenario, with an average R^2 of 0.03 and little variation between scenarios.

ABC simulations

?? shows ~~maximum *a posteriori* (MAP)~~ posterior mean point estimates of the BA model parameters obtained with ABC on simulated data. 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 are given in table 2.

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.

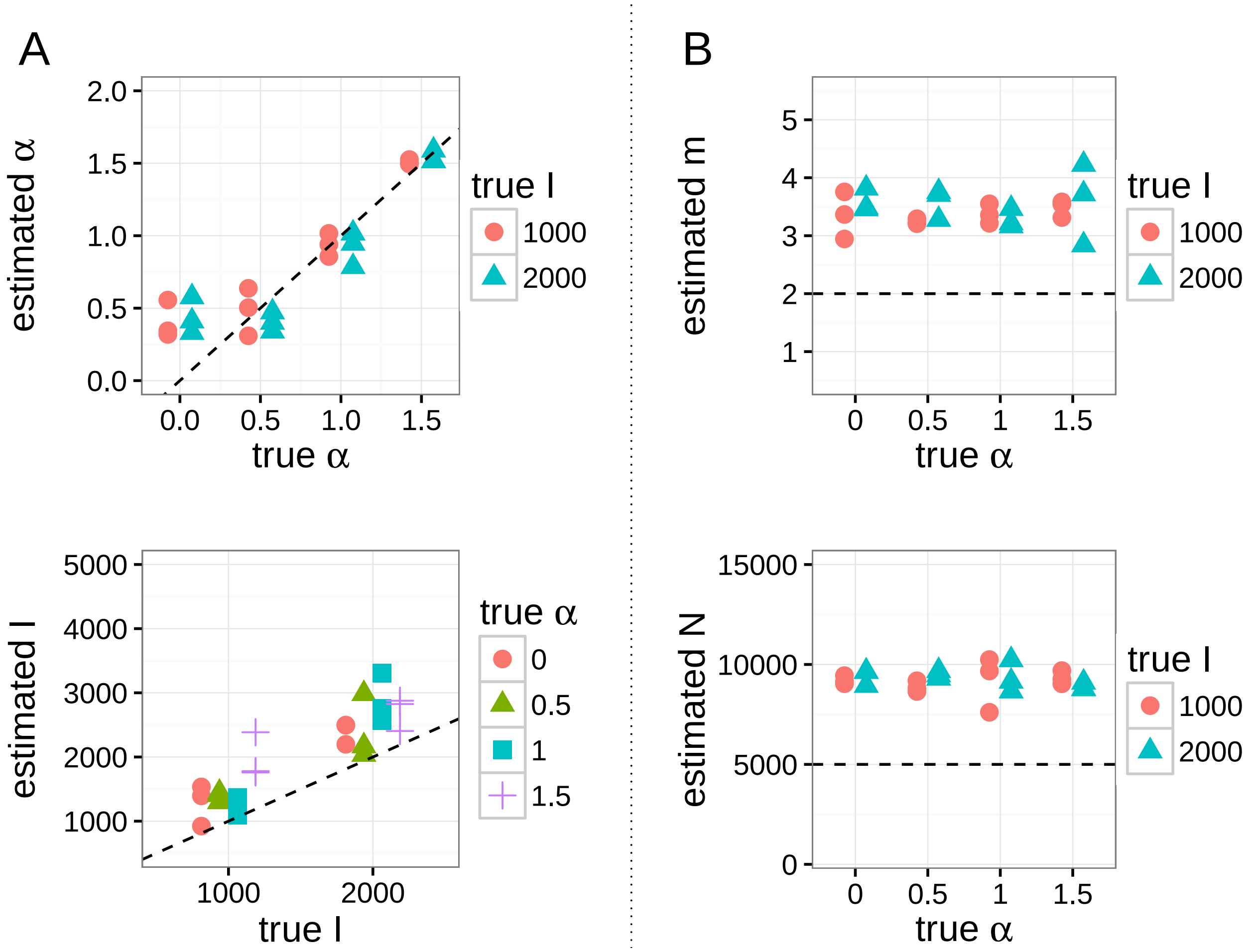

Figure 2: Posterior mean point estimates for BA model parameters obtained by running *netabc* on simulated data, for simulations with $m = 2$. Dashed lines indicate true values. (A) Estimates of α and I which were varied in these simulations against known values. (B) Estimates of m and N which were held fixed in these simulations at the values $m = 2$ and $N = 5000$.

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

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. 2 and **????**), GLMs were not fitted for these parameters. The estimated coefficients are shown in table 3.

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.

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 prevalence I 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 parameters~~, 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. In particular, the simulation shown is one of two with errors in both α and I between the 42nd and 58th percentiles. 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.

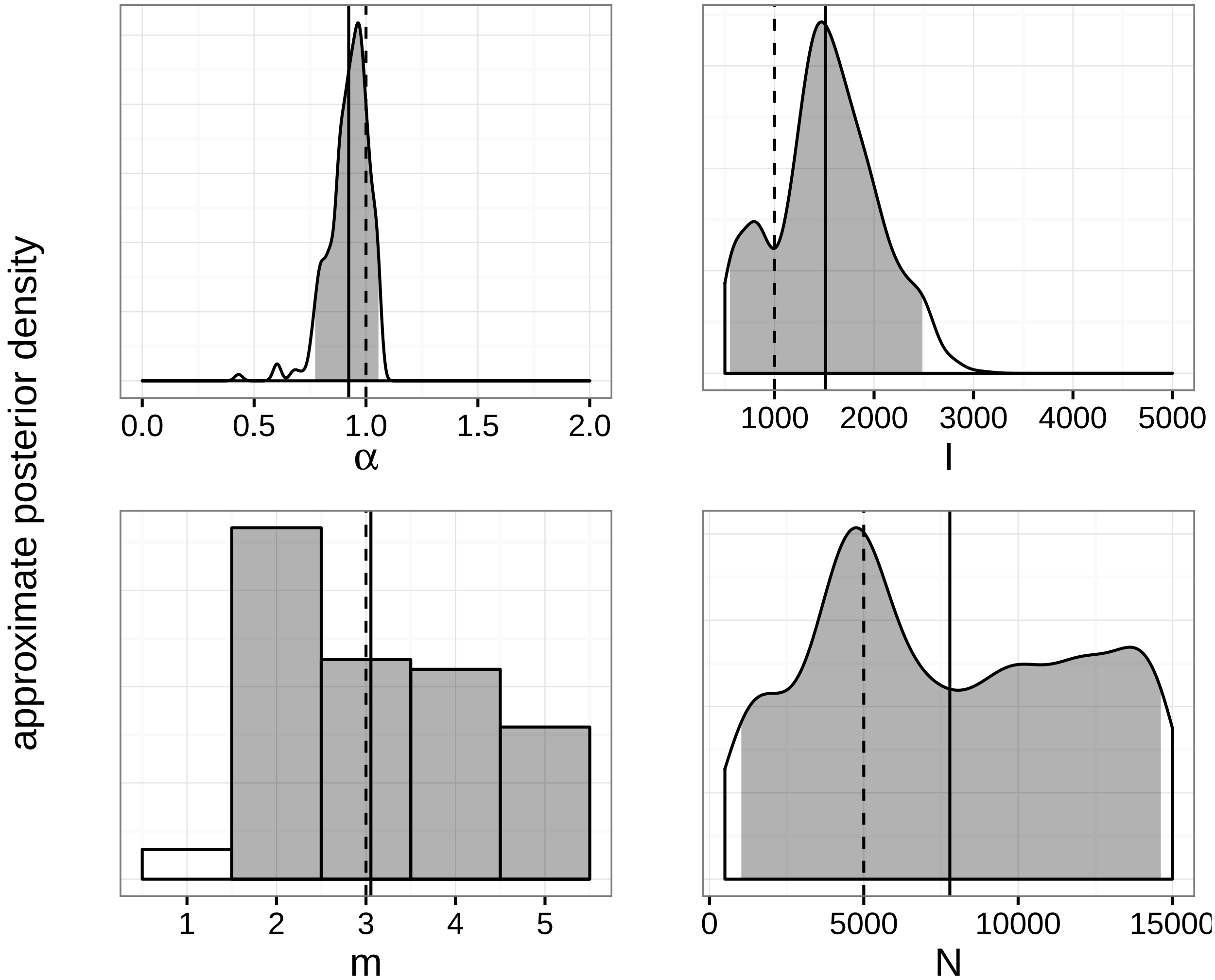

Figure 3: One-dimensional marginal posterior distributions of BA model parameters estimated by *netabc* from a simulated transmission tree. Dashed lines indicate true values, solid lines indicate posterior means, and shaded areas show 95% highest posterior density intervals.

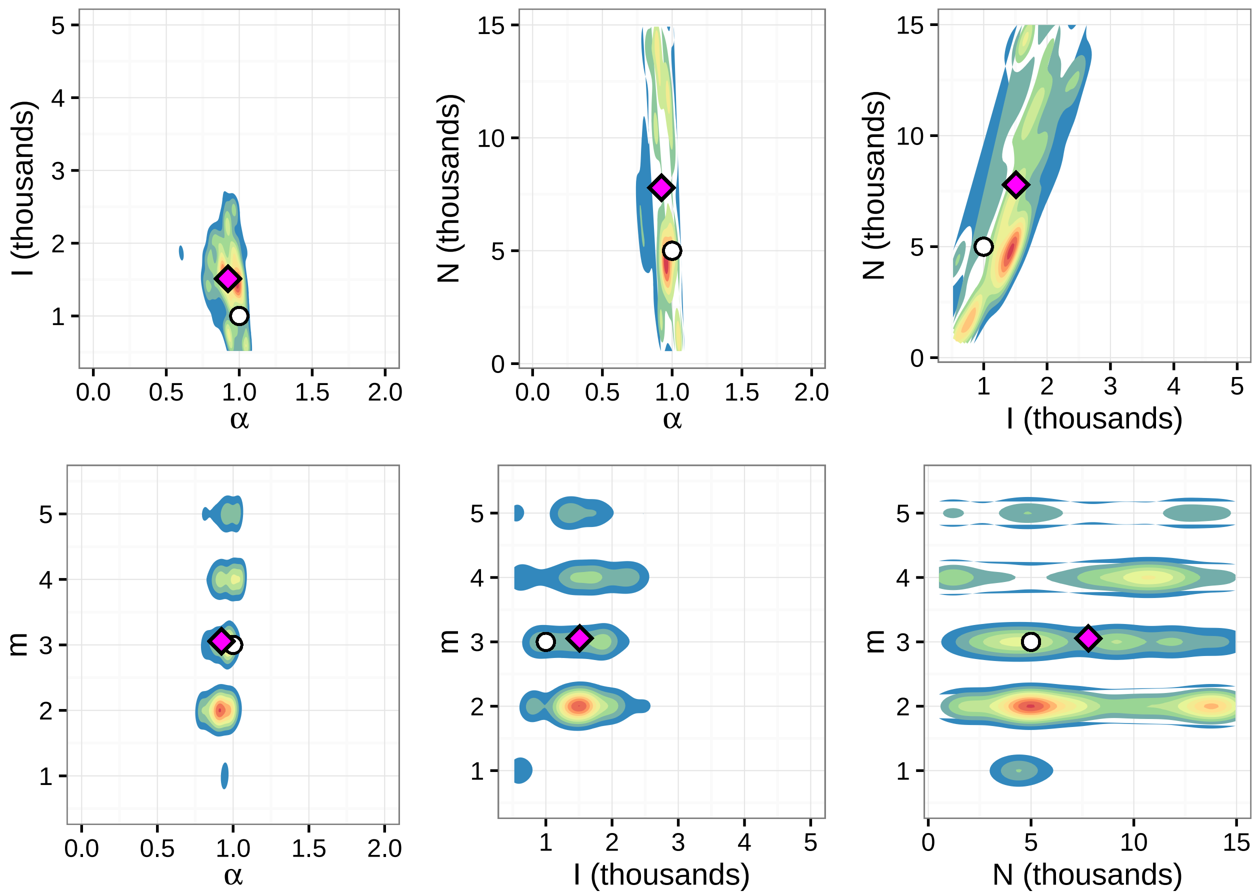

Figure 4: Two-dimensional marginal posterior distributions of BA model parameters estimated by *netabc* from a simulated transmission tree. White circles indicate true values, magenta diamonds indicate posterior means.

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 peer-driven sampling for α was somewhat lower than typical.

Real data

~~We applied ABC to five published HIV datasets (table 1), and found substantial heterogeneity among the parameter estimates (fig. 5 and **??**).~~ 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 **??????????**.~~ **??** shows point estimates and 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 are for the prior $m \sim \text{DiscreteUniform}(1,5)$ unless otherwise stated.

~~Two of the datasets [7, 54] 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. [54] data was narrower (0.05 - 1.18). When $m = 1$ was permitted, the Li et al. [61] and Cuevas et al. [62] 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. [56] 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 $m \geq 2$).~~

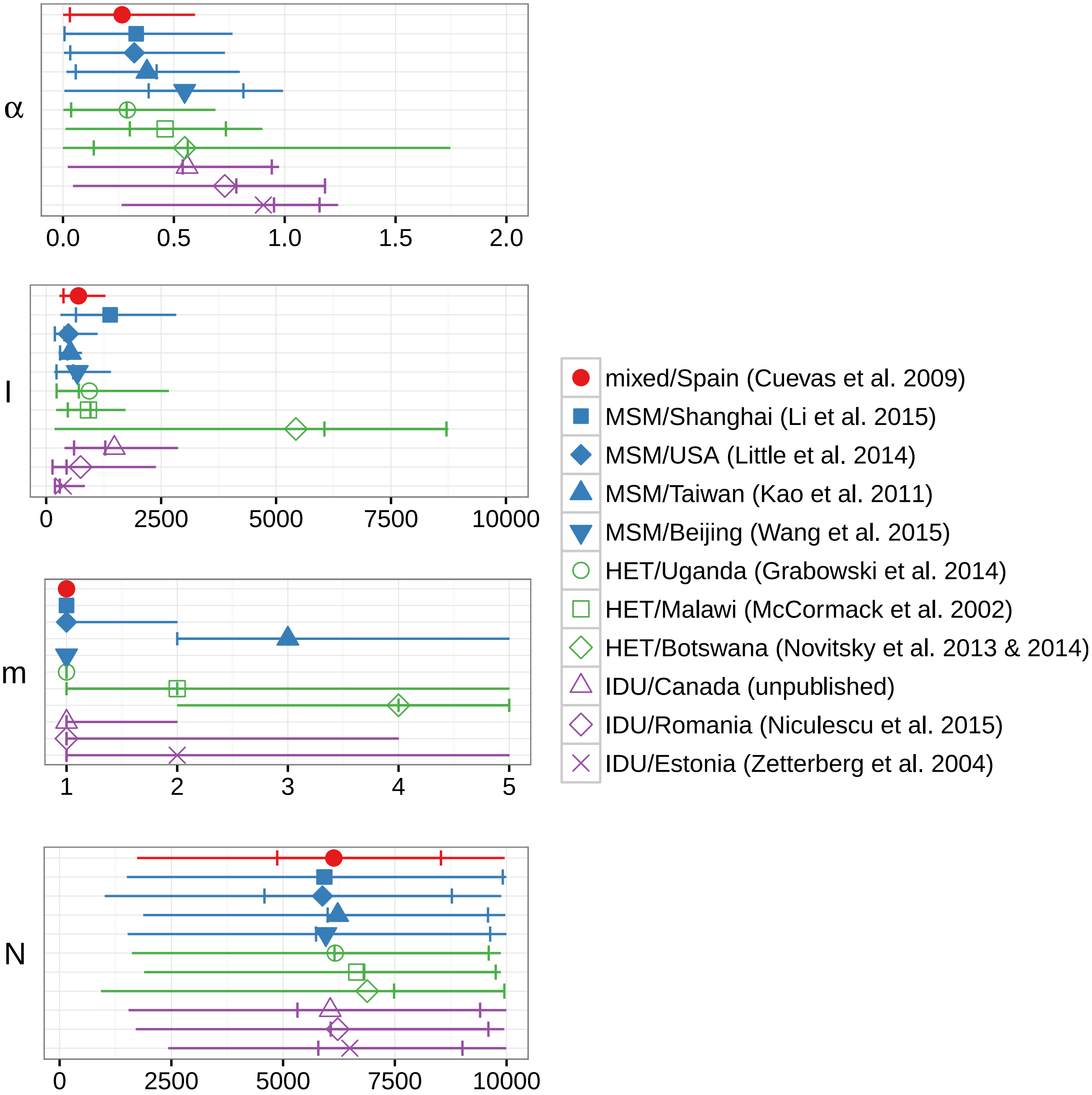

Figure 5: Posterior means (points), 50% HPD intervals (notches), and 95% HPD intervals (lines) for parameters of the BA network model, fitted to eleven HIV datasets with *netabc*. Legend labels indicate risk group and country of origin. Abbreviations: IDU, injection drug users; MSM, men who have sex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