Phylodynamic inference of contact network parameters through approximate Bayesian computation

Rosemary M. McCloskey Richard H. Liang Art F.Y. Poon

March 23, 2016

Background

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 (Klovdahl 1985; Morris 1993). 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 (O'Dea and Wilke 2010) and epidemic threshold (Barthélemy et al. 2005), in turn affecting the estimates of quantities such as effective population size (Goodreau 2006). 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 (Wang et al. 2015). True contact networks are a challenging type of data to collect, requiring extensive epidemiological investigation (Welch, Bansal, and Hunter 2011).

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 intra-host viral phylogeny (Drummond et al. 2003). 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 (Grenfell et al. 2004). These methods have revealed diverse properties of local viral outbreaks, from basic reproductive number (Stadler et al. 2011), to the degree of clustering (Hughes et al. 2009), to the elevated transmission risk during acute infection (Volz et al. 2012). On the other hand, although sophisticated methods have been developed for fitting complex population genetic models to phylogenies (Rasmussen, Volz, and Koelle 2014), 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 (Leventhal et al. 2012; Colijn and Gardy 2014; Goodreau 2006; Robinson et al. 2013), suggesting that such statistical inference might be possible (Welch, Bansal, and Hunter 2011).

Survey-based studies of sexual networks (Liljeros et al. 2001; Schneeberger et al. 2004; Colgate et al. 1989) have found that these networks have a degree distribution which follows a power law (although there has been some disagreement, see Handcock and Jones 2004). That is, number of nodes of degree k is proportional to $k^{-}\gamma$ for some constant γ. These networks are also referred to as "scale-free". 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 (1999), 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 here.

Previous work offers precedent for the possibility of statistical inference of structural network parameters. Britton and O'Neill (2002) develop a Bayesian approach to estimate the edge density in an Erdős-Rényi network (Erdős and Rényi 1960) given observed infection dates, and optionally recovery dates. Their approach was later extended by Groendyke, Welch, and Hunter (2011) and applied to a much larger data set of 188 individuals. Volz and Meyers (2007) and Volz (2008) 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. (2011) analysed the degree distribution of an approximate transmission network, estimated based on genetic similarity and estimated times of infection, relating 60% of human immunodeficiency virus (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). 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 (Sunnåker et al. 2013). In the specific case when ρ is a kernel function, the approach is known as kernel-ABC (Nakagome, Fukumizu, and Mano 2013; Poon 2015).

Here, we apply kernel-ABC to the problem of statistical inference of contact network parameters from an estimated transmission tree, using the tree kernel developed by Poon et al. (2013). We then estimate the parameters of the BA model on a variety of simulated and real data sets. Our results show that the attachment power parameter α can be inferred with reasonable accuracy, and can vary considerably between epidemics from different settings.

Methods

We implemented a Gillespie simulation algorithm (Gillespie 1976) for simulating epidemics and transmission trees over static contact networks, in the same fashion as several previous studies (*e.g.* O'Dea and Wilke 2010; Robinson et al. 2013; Leventhal et al. 2012; Groendyke, Welch, and Hunter 2011; Goodreau 2006). To check that our implementation was correct, we reproduced

Figure 1A of Leventhal et al. (2012) (our fig. S1), which plots the unbalancedness of transmission trees simulated over four network models at various levels of pathogen transmissibility. Our program is freely available at `https://github.com/rmcclosk/netabc`.

We chose to study the BA network model (Barabási and Albert 1999). In addition to m and α, we investigated the parameters 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. 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 (Csardi and Nepusz 2006) to generate the graphs. The analyses were run on Westgrid (`https://www.westgrid.ca/`) and a local computer cluster.

Kernel classifiers

We used the phylogenetic kernel developed by Poon et al. (2013) to test whether the parameters of the BA model had an effect on tree shape. We simulated 100 networks 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 Poon et al. 2013), were set to 0.3 and 4 respectively. We constructed a kernel support vector machine (kSVM) classifier for α using the *kernlab* package (Karatzoglou et al. 2004), and evaluated its accuracy 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 kSVM cross-validations, 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) and in epidemics of differing size (500, 1000, and 2000). 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 univariate classifier based on Sackin's in-

dex (Shao 1990) and an ordinary SVM based on the normalized lineages-through-time (nLTT) statistic (Janzen, Höhna, and Etienne 2015).

ABC simulations

We implemented the adaptive sequential Monte-Carlo (SMC) algorithm for ABC developed by Del Moral, Doucet, and Jasra (2012). To check that our implementation was correct, we applied 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 (2007)). 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.

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: $m \sim \text{Uniform}(1, 5)$, $\alpha \sim \text{Uniform}(0, 2)$, and (N, I) jointly uniform on the region $\{500 \leq N \leq 15000, 500 \leq I \leq 5000, I \leq N\}$. Following Del Moral, Doucet, and Jasra (2012) and Beaumont et al. (2009), all proposals were Gaussian, with variance equal to twice the empirical variance of the particles. The algorithm was run with 1000 particles, 5 simulated datasets per particle, and the "quality" parameter controlling the decay rate of the tolerance ε set to 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, using the *density* function with the default parameters in *R*. Highest posterior density (HPD) intervals were calculated with the *HPDinterval* function from the *R* package *coda* (Plummer et al. 2006).

Two further analyses were performed to address potential sources of error. 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 $\alpha = 0.5$.

Investigation of published data

We applied our kernel-ABC method to several 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.

Reference	Sequences (n)	Location	Risk group	Gene
(Wang et al. 2015)	173	Beijing, China	MSM	*pol*
(Cuevas et al. 2009)	287	Basque Country, Spain	mixed	*pol*
(Novitsky et al. 2013) (Novitsky et al. 2014)	180	Mochudi, Botswana	mixed	*env*
(Li et al. 2015)	280	Shanghai, China	MSM	*pol*
(Niculescu et al. 2015)	136	Romaina	IDU	*pol*

Table 1: Characteristics of published datasets investigated with kernel-ABC. Acronyms: MSM, men who have sex with men; IDU, injection drug users.

We downloaded all sequences associated with each study from GenBank. For the Novitsky et al. (2014) data, each sequence was aligned pairwise to the HXB2 reference sequence (Genbank accession number HIVHXB2CG) and the hypervariable regions were clipped out with *BioPython* version 1.66+ (Cock et al. 2009). Sequences were multiply aligned using *MUSCLE* version 3.8.31 (Edgar 2004), and alignments were manually inspected with *Seaview* version 4.4.2 (Gouy, Guindon, and Gascuel 2010). Phylogenies were constructed from the nucleotide alignments by approximate maximum likelihood using *FastTree2* version 2.1.7 with the general time-reversible (GTR) model. 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 (Drummond and Rambaut 2007)) as described previously (Poon 2015).

Two of the datasets (Li et al. 2015; Novitsky et al. 2014) were initially much larger than the others, containing 1265 and 1299 sequences respectively. To ensure that the analyses were comparable, we reduced these to a number of sequences similar to the smaller datasets. For the Li et al. (2015) data, we detected a cluster of size 280 using a patristic distance cutoff of 0.02 as described previously (Poon et al. 2014). Only sequences within this cluster were carried forward. For the Novitsky et al. (2014) data, no large clusters were detected using the same cutoff, so we analysed a subtree of size 180 chosen arbitrarily.

Results

Kernel classifiers

Accuracy of the kSVM classifiers varied based on the parameter being tested (fig. 1, left). 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). Figure 1 shows the cross-validation accuracy of each of the tree classifiers. 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 (figs. S3 to S6).

When classifying α, the kernel-SVM classifier had an average R^2 of 0.92, compared to 0.56 for the nLTT-based SVM, 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 kSVM, 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 kSVM was slightly higher on 1000-tip trees.

The accuracy of classifiers 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 kSVM 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 SVM 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

Figure 2 shows maximum *a posteriori* (MAP) point estimates of the BA model parameters obtained with kernel-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 (figs. S7 and S8). Average boundaries of 95% HPD intervals are given in table 2.

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

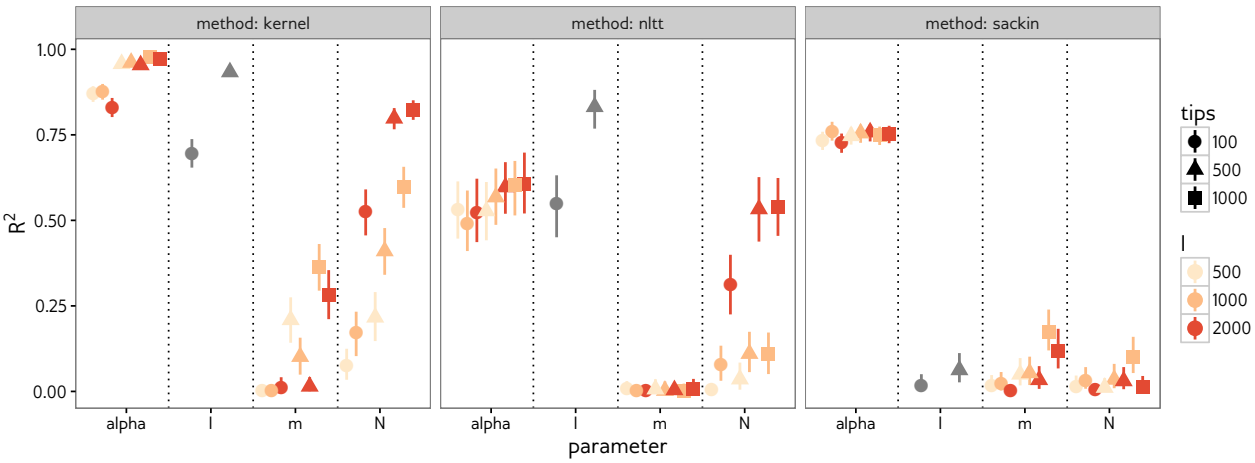

Figure 1: Cross-validation accuracy of kernel-SVM classifier (left), SVM classifier using nLTT (centre), and linear regression using Sackin's index (right) for BA model parameters. Kernel meta-parameters were set to $\lambda = 0.3$ and $\sigma = 4$. Each point was calculated based on 300 simulated transmission trees over networks with three different values of the parameter being tested. Vertical lines are empirical 95% confidence intervals based on 1000 two-fold cross-validations.

point estimates having a median [IQR] absolute error of 0.11 [0.05 - 0.18]. The errors when the true value of α was zero were significantly greater than those for the other values (Wilcoxon rank-sum test, $p = 0.0006$), but did not vary across the true values of the other parameters (one-way ANOVA). Estimates for I were also relatively accurate, with point estimate errors of 306 [108 - 607]. These errors were significantly higher when the true value of α was at least 1 (Wilcoxon rank-sum test, $p = 0.0006$) 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. Oddly, the error in the estimated m was higher for integer values of α (*i.e.* 0 and 1) than non-integer values (Wilcoxon rask-sum test, $p = 0.007$). 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 6588 [4214 - 8284]). No parameters influenced the accuracy of the N estimates (all one-way ANOVA).

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). Figure 3 shows the distributions for for one simulation (equivalent plots for all the simulations can be found in data S1). HPD intervals around α and I were narrow relative to the region of nonzero prior density, whereas the intervals for m and N were more widely dispersed.

To test the effect of model misspecification, we simulated one network where the nodes

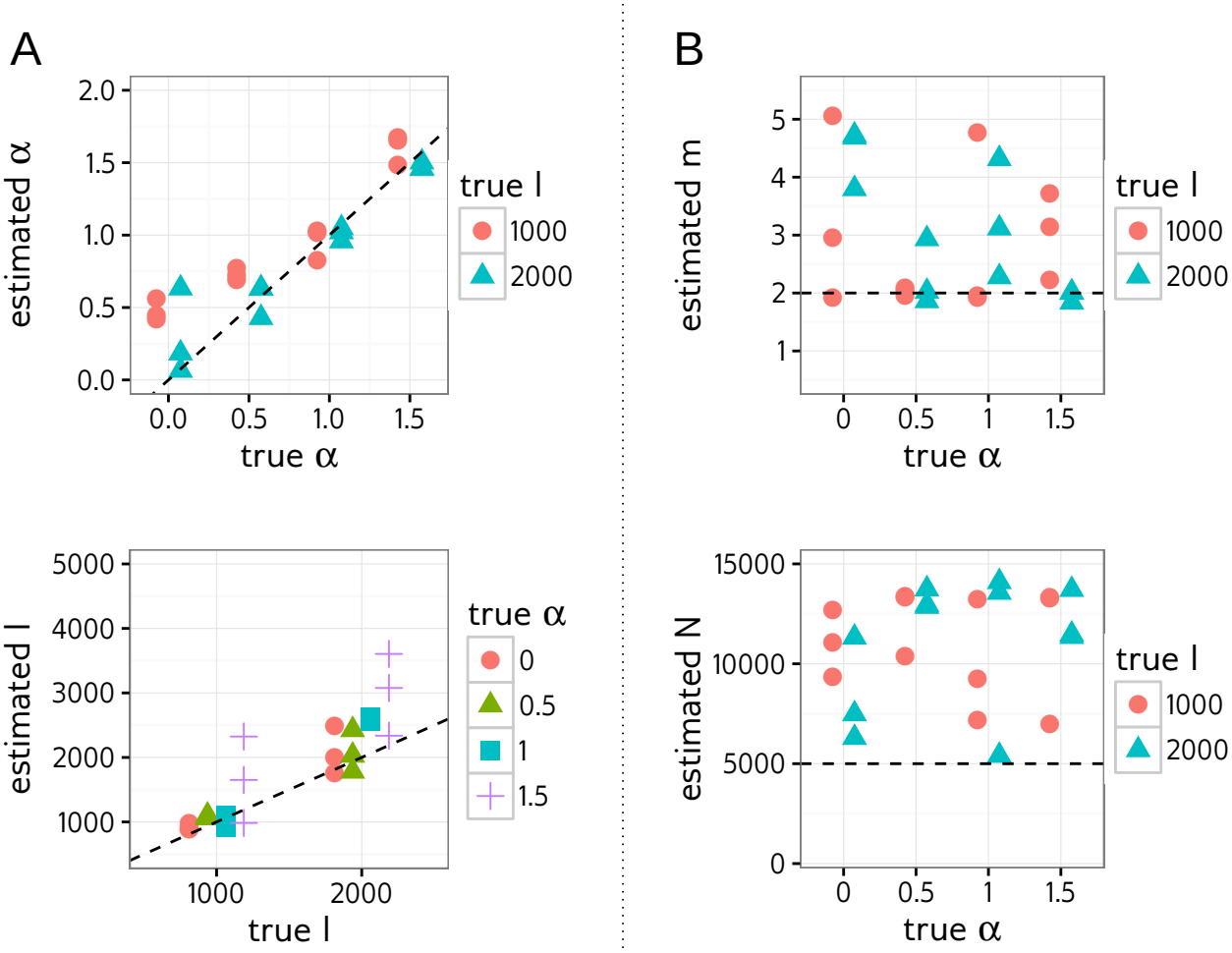

Figure 2: Point estimates of BA model parameters obtained by running kernel-ABC on simulated phylogenies without training, for simulations with $m = 2$. Dotted lines indicate true values, and limits of the y-axes are regions of uniform prior density.

exhibited heterogeneous preferential attachment power (half 0.5, the other half 1.5), with $m = 2$, $N = 5000$, and $I = 1000$. The MAP [95% HPD] estimates for each parameter were: α, 1 [0.75-1.12]; I, 1184 [506 - 1615]; m, 5 [2 - 5]; N, 10244 [3233- 14974]. 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 it's peers had already been sampled. The parameters for this experiment were $N = 5000$, $m = 2$, $\alpha = 0.5$, and $I = 2000$. The estimated values were α, 0.48 [0.03 - 0.82]; I, 2516 [1281 - 3938]; m, 3 [2 - 5]; N, 10930 [3106 - 14783]. Both of these results were in line with estimates obtained on other simulated datasets (table 2).

Parameter	True value	Mean point estimate	Mean HPD lower bound	Mean HPD upper bound
α	0.0	0.24	0.02	0.73
	0.5	0.42	0.02	0.81
	1.0	0.97	0.61	1.11
	1.5	1.48	1.26	1.83
I	1000	1155.68	598.68	2402.84
	2000	2646.07	1182.31	4058.13
m	2	2.92	1.75	4.92
	3	3.33	1.96	4.92
	4	3.62	1.88	5.00
N	5000	10962.61	2732.55	14701.87

Table 2: Average maximum *a posteriori* point estimates and 95% highest posterior density (HPD) interval widths for BA model parameter estimates obtained with kernel-ABC. Three transmission trees were simulated under each combination of the listed parameter values, and the parameters were estimated with kernel-ABC without training.

Real data

There was substantial heterogeneity among the parameter estimates for the five published HIV datasets we analysed. Two of the datasets (Niculescu et al. 2015; Wang et al. 2015) had estimated α values near unity (MAP estimate [95% HPD] 1.06 [0.63 - 1.27] and 1 [0.41 - 1.16]). Another two datasets (Li et al. 2015; Cuevas et al. 2009) had lower estimated values and wider HPD intervals (0.77 [0.01 - 1.03] and 0.66 [0.03 - 0.84]). The Novitsky et al. (2014) data had an extremely low estimated α and a very wide HPD interval (0.17 [0.04 - 1.39]). For all the datasets except Novitsky et al., estimated values of I were below 2000, with narrow HPD intervals around two of the datasets (Cuevas et al., 880 [290 - 1511]; Niculescu et al., 175 [138 - 454]) and wider intervals around the other two (Li et al., 1590 [284 - 3807]; Wang et al., 651 [268 - 4235]). The Novitsky et al. data was again the outlier, with a very high estimated I, and HPD interval spanning almost the entire prior region (7547 [228 - 8921]). No information was gleaned about the m parameter, with the HPD interval occupying the entire prior region for all datasets. The estimates of N were similarly uninformative, with the exception that the point estimate for the Wang et al. data was smaller (5839) than the estimates for other datasets (average 8927).

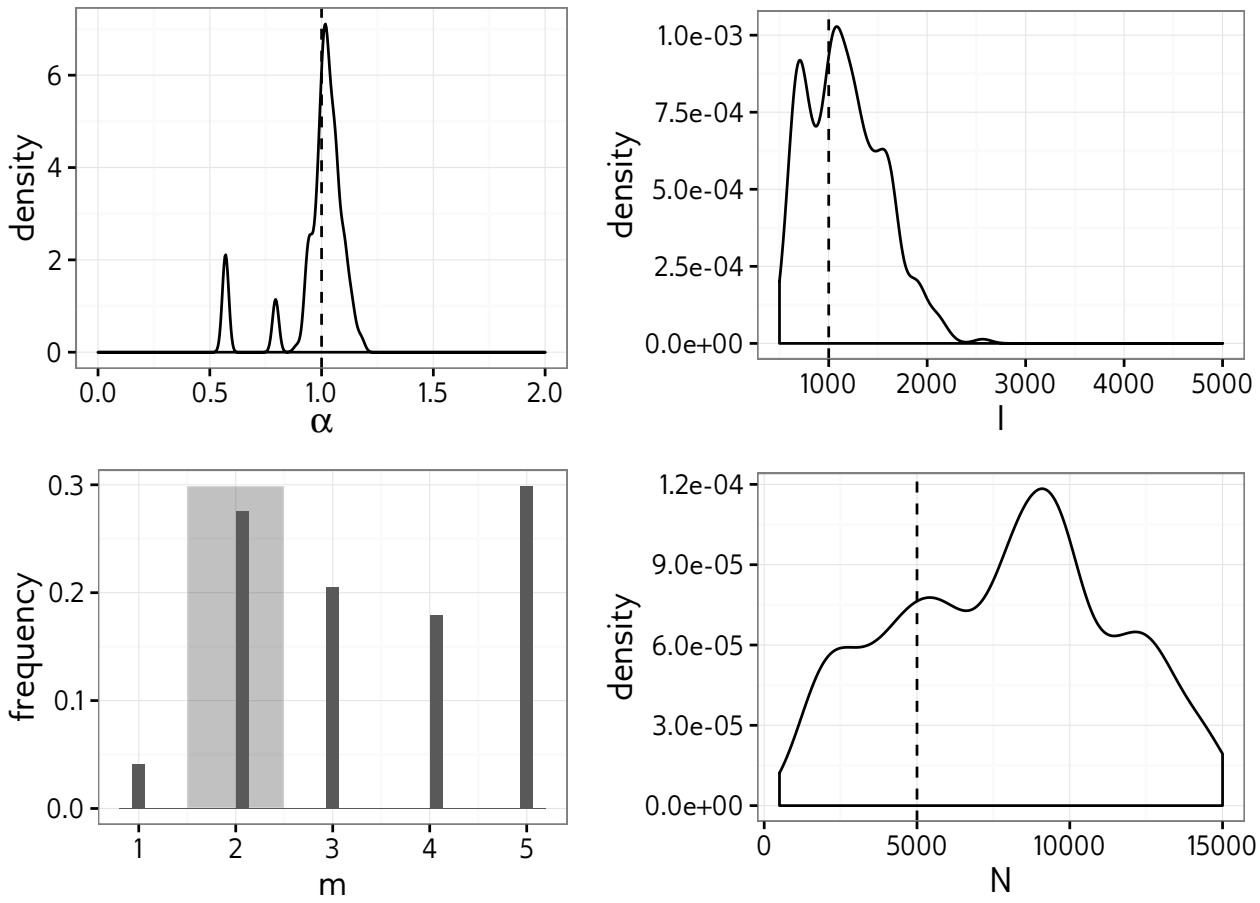

Figure 3: Marginal posterior distributions of BA model parameters estimated with kernel-ABC for a single simulated transmission tree. Dotted lines and shaded polygon indicate true values.

Discussion

Contact networks can have a strong influence on epidemic progression, and are potentially useful as a public health tool (Wang et al. 2015; Little et al. 2014). Despite this, few methods exist for investigating contact network parameters in a phylodynamic framework (Groendyke, Welch, and Hunter 2011; Volz 2008; Leigh Brown et al. 2011). Kernel-ABC is a model-agnostic method which can be used to investigate any quantity that affects tree shape. In this work, we developed a kernel-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 model giving rise to the power law degree distributions commonly observed in real-world networks.

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.

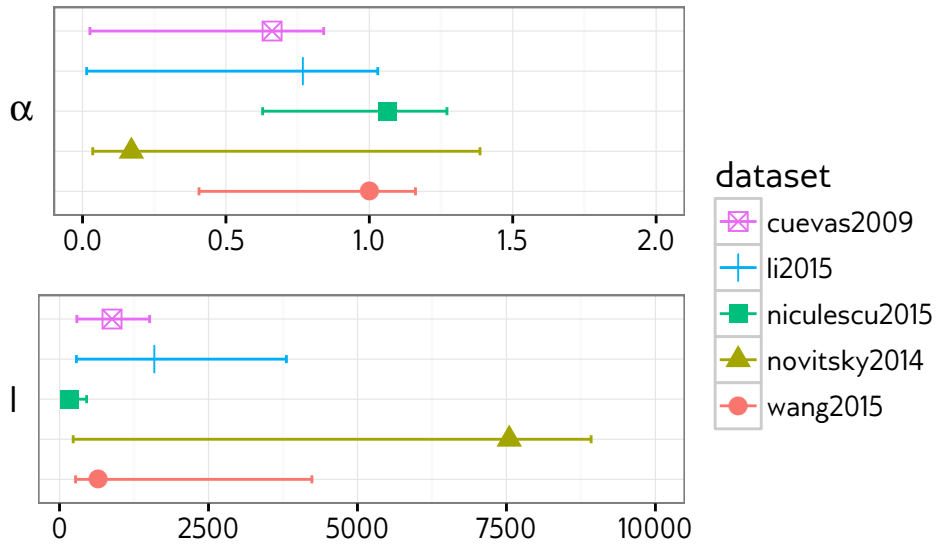

Figure 4: maximum *a posteriori* point estimates and 95% HPD intervals for parameters of the BA network model, fitted to five published HIV datasets with kernel-ABC.

This was reflected in the relative accuracy of the kernel-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 kernel-ABC estimates.

N was almost always significantly over-estimated using kernel-ABC. Since the prior on N and I is jointly uniform on a non-rectangular region ($I \leq N$), there is more prior mass on high N values. In retrospect, it is unreasonable to expect good estimation of N, because adding more nodes to a BA network does not change the edge density or overall shape. 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. 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.

As noted by Lintusaari et al. (2016), 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. S9). 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 similar to fig. S9 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 (Liljeros et al. 2001; Schneeberger et al. 2004; Colgate et al. 1989; Leigh Brown et al. 2011). We note however that Jones and Handcock (2003) estimated γ values greater than four, in one case as high as 17, for some datasets, indicating that a wider range of permitted γ values may be warranted.

Our investigation of published HIV datasets indicated heterogeneity in the contact network structures underlying several distinct local epidemics. The five datasets analysed fell into three categories (fig. 4). First, we estimated a preferential attachment power between 0.5 and 1 for the epidemics studied by Cuevas et al. (2009) and Li et al. (2015), with credible intervals occupying nearly the entire region from 0 to 1. Cuevas et al. studied a group of newly diagnosed individuals in the Basque Country, Spain. Although the individuals were of mixed risk groups, and therefore unlikely to comprise a single contact network, a high proportion of them (47%) grouped into local clusters based on genetic distance. The low estimated attachment power for these data is consistent with the sampled sequences comprising many distinct sub-networks rather than a single connected network. Li et al. sampled a large number of acutely infected MSM in Shanghai, China, in which we identified a large cluster from the phylogeny using a patristic distance cut-off (Poon et al. 2014). The low attachment power estimated for this dataset was surprising given the high phylogenetic relatedness of the sequences. It is possible that the number and diversity of circulating recombinant forms introduced errors into the estimated viral phylogeny.

For the outbreaks studied by Niculescu et al. (2015) and Wang et al. (2015), the estimated α was close to one, with a narrower credible interval than for the other studies. Niculescu et al. studied a recent outbreak among Romanian injection drug user (IDU), while Wang et al. sampled acutely infected MSM in Beijing, China. Both studies discovered a high degree of phylogenetic relatedness owing to the recent infection times and homogeneous risk groups of the studied populations. The estimated number of infections for these datasets were also quite low, although the HPD interval for Wang et al. was much wider than that for Niculescu et al.

The final studied dataset was an outlier in terms of estimated parameters. Novitsky et al. (2013) sampled approximately 44% of the HIV-infected individuals in the northern area of Mochudi, Botswana. Additional sampling in a later study (Novitsky et al. 2014) brought the genotyping coverage up to 70%. Even with such a high sampling coverage, we did not detect any large clusters using patristic distance, and therefore chose to analyze a subtree instead. Estimates of α and N both had very wide HPD intervals and were markedly different from the other datasets. The estimated number of infected nodes was also extremely high, much higher than the estimated HIV prevalence of the town. Several factors may have contributed to these

results. First, the authors note that the their sample was 75% female. In a primarily heterosexual risk environment, removal of a disproportionate number of males from the network could obfuscate the true network structure, for example if there were a disproportionate number of highly connected nodes of one gender. Second, the town in question was in close proximity to the country's capital, and the authors indicated that a high amount of migration takes place between the two locations. This suggests that the contact network may include a much larger group based in the capital city, which would explain the high estimate of I.

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 assigned it zero prior probability. This forces the average degree to be at least four, which may be unrealistically high for sexual networks. Additional modelling assumptions include the network being connected and static, all transmission rates being equal, no removal after infection, and identical behaviour of all nodes. This last is particularly problematic, as we showed by simulating a network where some nodes exhibited a higher attachment power than others. The estimated attachment power was simply the average of the two values, indicating that, although we could characterize the network in aggregate, the estimated parameters could not be said to apply to any individual node. Despite these issues, we felt it was best to demonstrate the method first on a simple model. It is possible 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. Leitner et al. 1996), and is frequently used in practice (e.g. Stadler and Bonhoeffer 2013), the topologies of a viral phylogeny and transmission tree can differ significantly (Ypma, Ballegooijen, and Wallinga 2013) due to within-host evolution and the sampling process. In addition, 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 (Poon 2015), has demonstrated that kernel-ABC is a broadly applicable and effective framework in which to perform phylodynamic inference.

References

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

Barthélemy, Marc et al. (2005). "Dynamical patterns of epidemic outbreaks in complex heterogeneous networks". In: *Journal of theoretical biology* 235.2, pp. 275–288.

Beaumont, Mark A et al. (2009). "Adaptive approximate Bayesian computation". In: *Biometrika*, asp052.

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

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

Colgate, Stirling A et al. (1989). "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, pp. 4793–4797.

Colijn, Caroline and Jennifer Gardy (2014). "Phylogenetic tree shapes resolve disease transmission patterns". In: *Evolution, medicine, and public health* 2014.1, pp. 96–108.

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

Cuevas, Maria Teresa et al. (2009). "HIV-1 transmission cluster with T215D revertant mutation among newly diagnosed patients from the Basque Country, Spain". In: *JAIDS Journal of Acquired Immune Deficiency Syndromes* 51.1, pp. 99–103.

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

Drummond, Alexei J and Andrew Rambaut (2007). "BEAST: Bayesian evolutionary analysis by sampling trees". In: *BMC evolutionary biology* 7.1, p. 214.

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

Edgar, Robert C (2004). "MUSCLE: multiple sequence alignment with high accuracy and high throughput". In: *Nucleic acids research* 32.5, pp. 1792–1797.

Erdős, Paul and Alfred Rényi (1960). "On the evolution of random graphs". In: *Publ. Math. Inst. Hungar. Acad. Sci* 5, pp. 17–61.

Gillespie, Daniel T (1976). "A general method for numerically simulating the stochastic time evolution of coupled chemical reactions". In: *Journal of computational physics* 22.4, pp. 403–434.

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

Gouy, Manolo, Stéphane Guindon, and Olivier Gascuel (2010). "SeaView version 4: a multiplatform graphical user interface for sequence alignment and phylogenetic tree building". In: *Molecular biology and evolution* 27.2, pp. 221–224.

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

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

Handcock, Mark S and James Holland Jones (2004). "Likelihood-based inference for stochastic models of sexual network formation". In: *Theoretical population biology* 65.4, pp. 413–422.

Hughes, Gareth J et al. (2009). "Molecular phylodynamics of the heterosexual HIV epidemic in the United Kingdom". In: *PLoS Pathog* 5.9, e1000590.

Janzen, Thijs, Sebastian Höhna, and Rampal S Etienne (2015). "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, pp. 566–575.

Jones, James Holland and Mark S Handcock (2003). "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, pp. 1123–1128.

Karatzoglou, Alexandros et al. (2004). "kernlab-an S4 package for kernel methods in R". In:

Klovdahl, Alden S (1985). "Social networks and the spread of infectious diseases: the AIDS example". In: *Social science & medicine* 21.11, pp. 1203–1216.

Leigh Brown, Andrew J et al. (2011). "Transmission network parameters estimated from HIV sequences for a nationwide epidemic". In: *Journal of Infectious Diseases*, jir550.

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

Leventhal, Gabriel E et al. (2012). "Inferring epidemic contact structure from phylogenetic trees". In: *PLoS Comput Biol* 8.3, e1002413–e1002413.

Li, Xiaoyan et al. (2015). "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, e0129559.

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

Lintusaari, Jarno et al. (2016). "On the Identifiability of Transmission Dynamic Models for Infectious Diseases". In: *Genetics*, genetics–115.

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

Morris, Martina (1993). "Epidemiology and social networks: Modeling structured diffusion". In: *Sociological Methods & Research* 22.1, pp. 99–126.

Nakagome, Shigeki, Kenji Fukumizu, and Shuhei Mano (2013). "Kernel approximate Bayesian computation in population genetic inferences". In: *Statistical applications in genetics and molecular biology* 12.6, pp. 667–678.

Niculescu, Iulia et al. (2015). "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, pp. 488–495.

Novitsky, Vladimir et al. (2013). "Phylogenetic relatedness of circulating HIV-1C variants in Mochudi, Botswana". In: *PloS one* 8.12, e80589.

Novitsky, Vlad et al. (2014). "Impact of sampling density on the extent of HIV clustering". In: *AIDS research and human retroviruses* 30.12, pp. 1226–1235.

O'Dea, Eamon B and Claus O Wilke (2010). "Contact heterogeneity and phylodynamics: how contact networks shape parasite evolutionary trees". In: *Interdisciplinary perspectives on infectious diseases* 2011.

Plummer, Martyn et al. (2006). "CODA: Convergence diagnosis and output analysis for MCMC". In: *R news* 6.1, pp. 7–11.

Poon, Art FY (2015). "Phylodynamic inference with kernel ABC and its application to HIV epidemiology". In: *Molecular biology and evolution*, msv123.

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

Poon, Art FY et al. (2014). "The impact of clinical, demographic and risk factors on rates of HIV transmission: a population-based phylogenetic analysis in British Columbia, Canada". In: *Journal of Infectious Diseases*, jiu560.

Rasmussen, David A, Erik M Volz, and Katia Koelle (2014). "Phylodynamic inference for structured epidemiological models". In: *PLoS Comput Biol* 10.4, e1003570.

Robinson, Katy et al. (2013). "How the dynamics and structure of sexual contact networks shape pathogen phylogenies". In: *PLoS computational biology* 9.6, e1003105.

Schneeberger, Anne et al. (2004). "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, pp. 380–387.

Shao, Kwang-Tsao (1990). "Tree balance". In: *Systematic Biology* 39.3, pp. 266–276.

Sisson, Scott A, Yanan Fan, and Mark M Tanaka (2007). "Sequential monte carlo without likelihoods". In: *Proceedings of the National Academy of Sciences* 104.6, pp. 1760–1765.

Stadler, Tanja and Sebastian Bonhoeffer (2013). "Uncovering epidemiological dynamics in heterogeneous host populations using phylogenetic methods". In: *Philosophical Transactions of the Royal Society of London B: Biological Sciences* 368.1614, p. 20120198.

Stadler, Tanja et al. (2011). "Estimating the basic reproductive number from viral sequence data". In: *Molecular biology and evolution*, msr217.

Sunnåker, Mikael et al. (2013). "Approximate bayesian computation". In: *PLoS Comput Biol* 9.1, e1002803.

Volz, Erik (2008). "SIR dynamics in random networks with heterogeneous connectivity". In: *Journal of mathematical biology* 56.3, pp. 293–310.

Volz, Erik M et al. (2012). "Simple epidemiological dynamics explain phylogenetic clustering of HIV from patients with recent infection". In: *PLoS Comput Biol* 8.6, e1002552–e1002552.

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

Wang, Xicheng et al. (2015). "Targeting HIV Prevention Based on Molecular Epidemiology Among Deeply Sampled Subnetworks of Men Who Have Sex With Men". In: *Clinical Infectious Diseases*, p. civ526.

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

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

Supplemental Materials

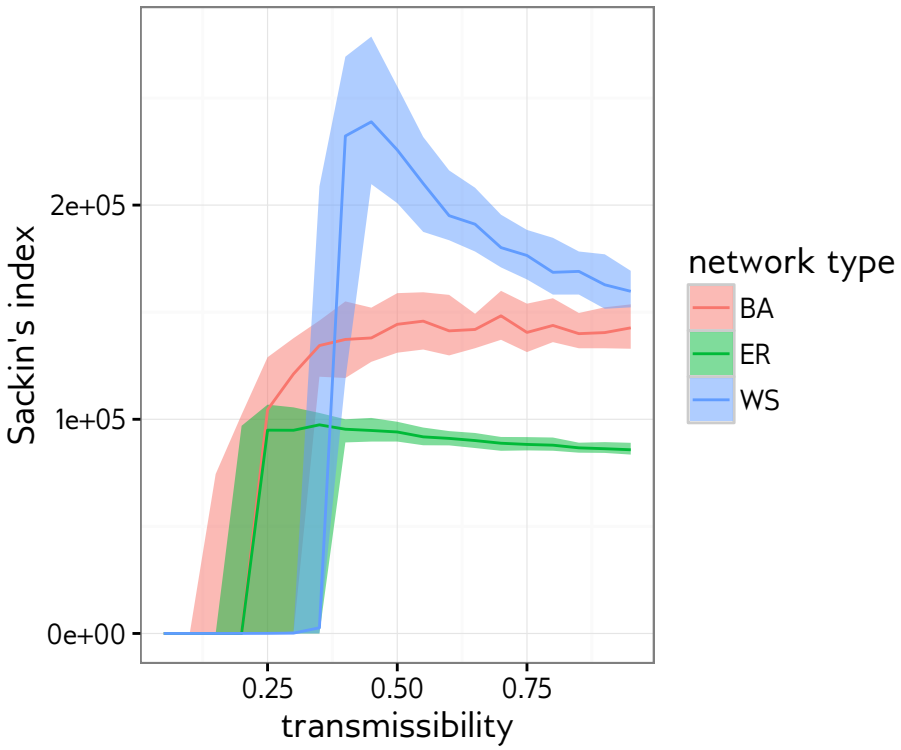

Figure S1: Reproduction of Figure 1A from Leventhal et al. (2012) used to check the accuracy of our implementation of Gillespie simulation. Transmission trees were simulated over three types of network, with pathogen transmissimility varying from 0 to 1. Sackin's index was calculated for each simulated transmission tree.

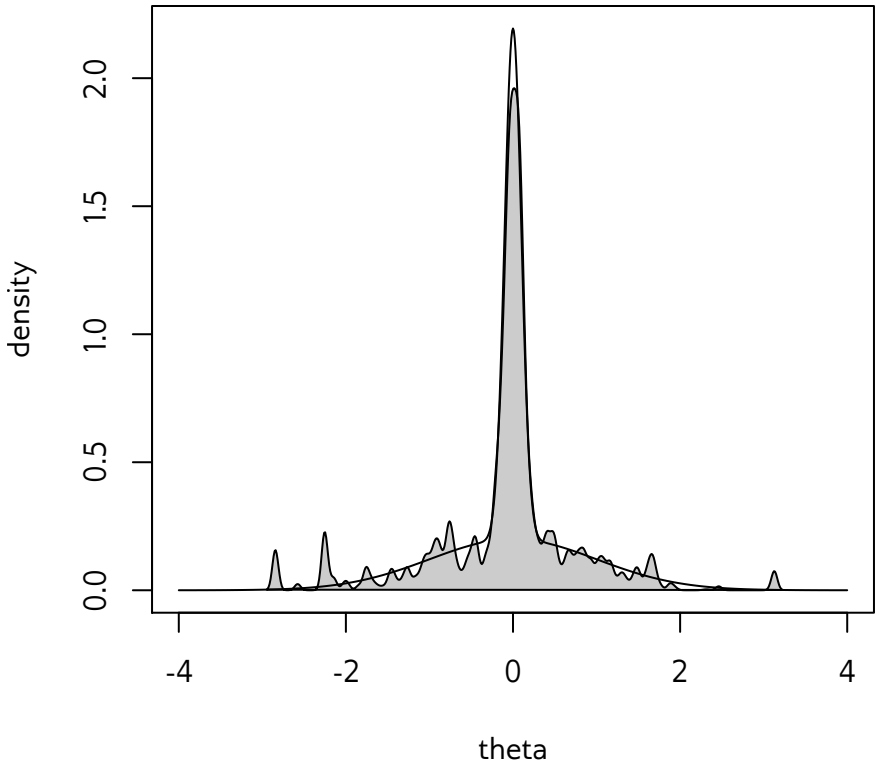

Figure S2: Approximation of mixture of Gaussians used by Del Moral, Doucet, and Jasra (2012) and Sisson, Fan, and Tanaka (2007) to test SMC. Solid black line indicates true distribution. Grey shaded area shows SMC approximation obtained with our implementation, using 10000 particles with one simulated data point per particle.

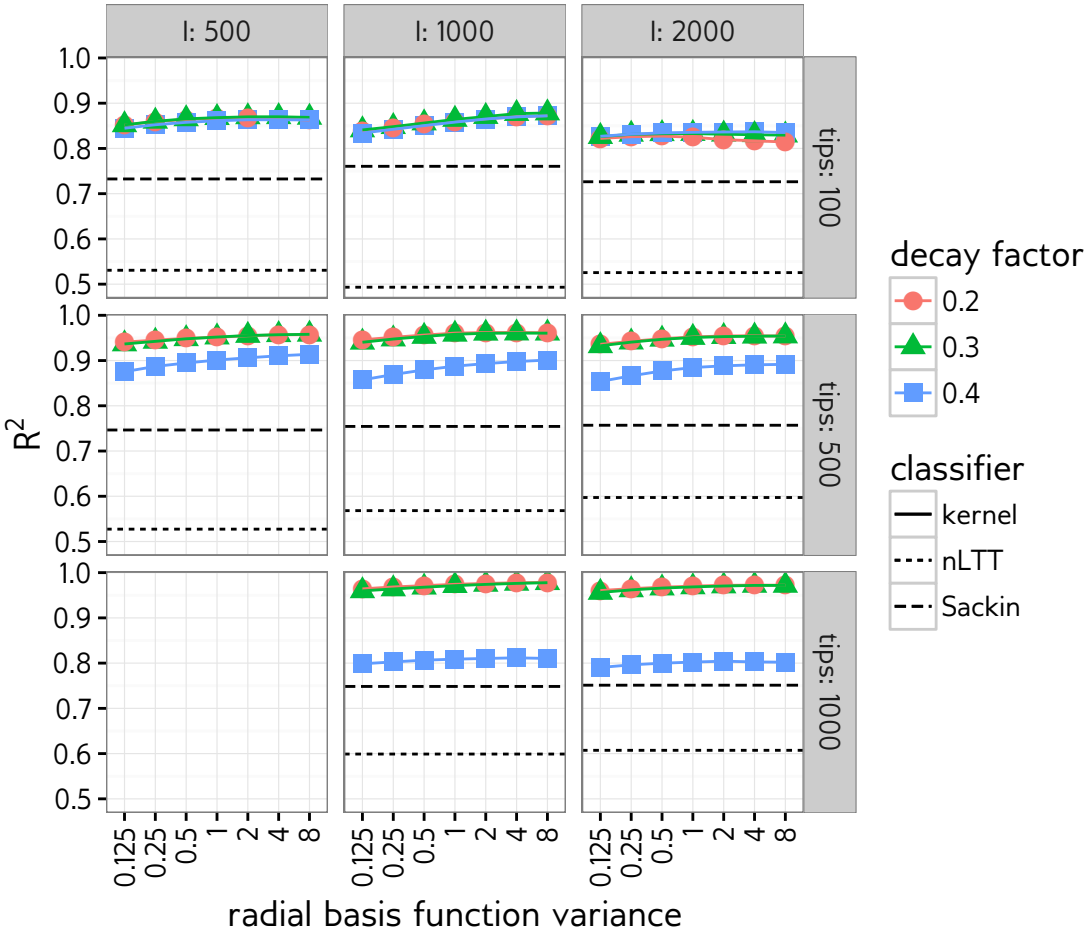

Figure S3: Cross-validation accuracy of kernel-SVM classifiers for α parameter of BA network model, for various tree kernel meta-parameters and epidemic scenarios. Each point was calculated based on 300 simulated transmission trees over networks with $\alpha = 0.5$, 1.0, or 1.5. Dotted and and dashed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

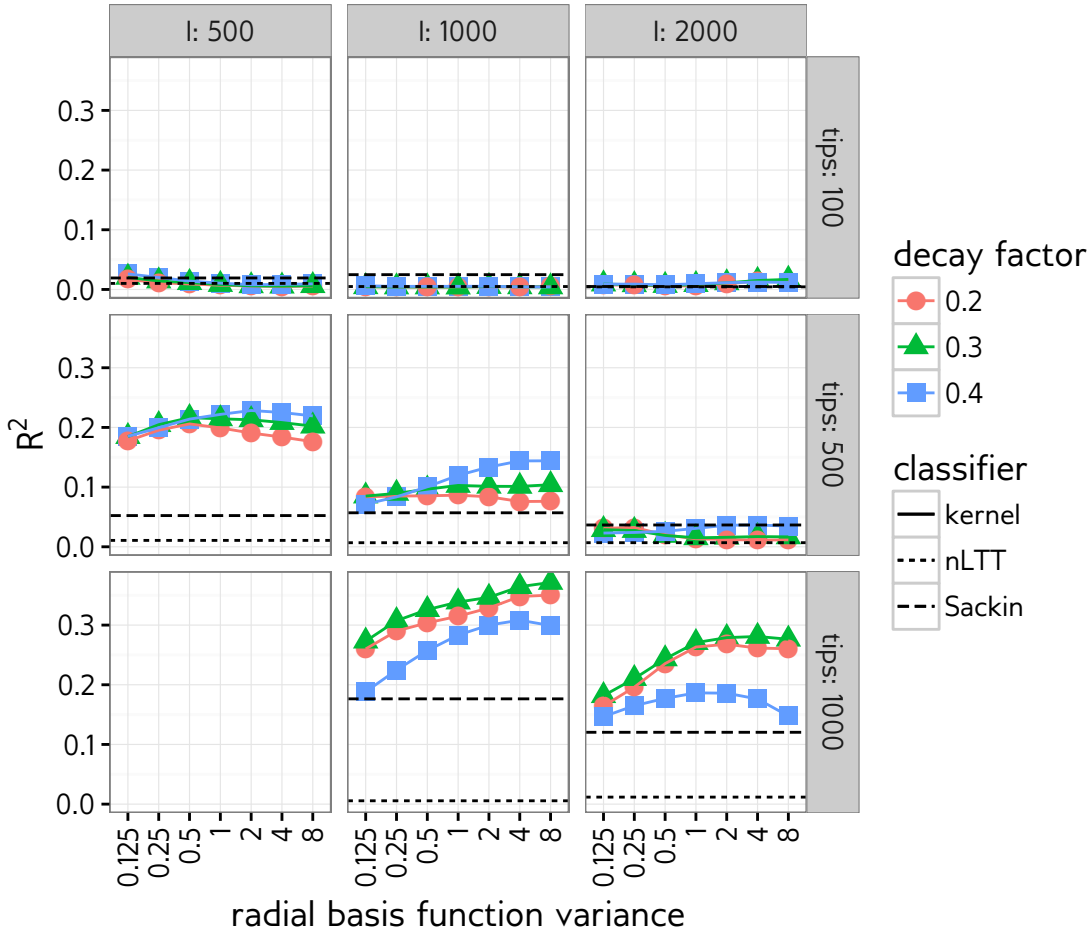

Figure S4: Cross-validation accuracy of kernel-SVM classifiers for m parameter of BA network model, for various tree kernel meta-parameters and epidemic scenarios. Each point was calculated based on 300 simulated transmission trees over networks with $m = 2$, 3, or 4. Dotted and and dashed lines indicate,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

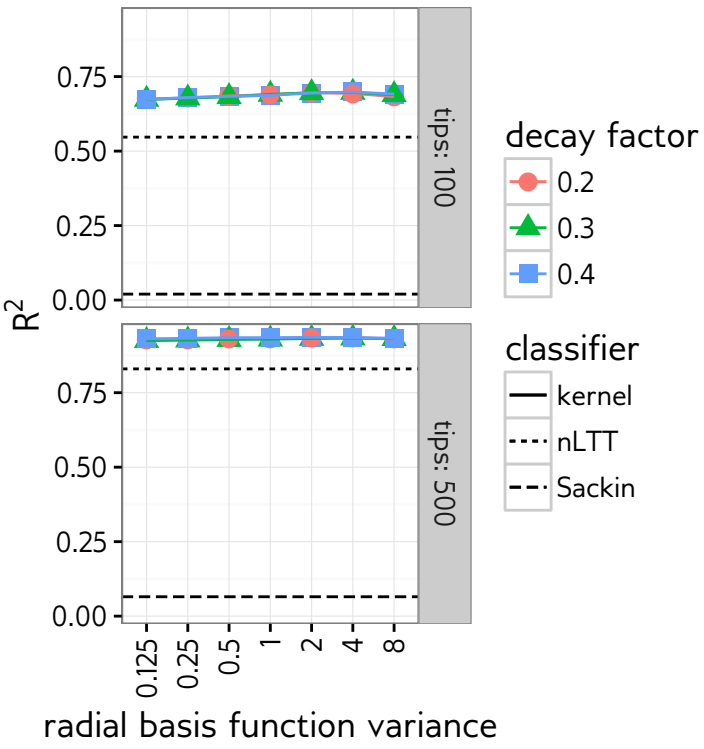

Figure S5: Cross-validation accuracy of kernel-SVM classifiers for number of infected nodes (I) under BA network model, for various tree kernel meta-parameters and two tree sizes. Each point was calculated based on 300 simulated transmission trees over networks with $I = 500$, 1000, or 2000. Dotted and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

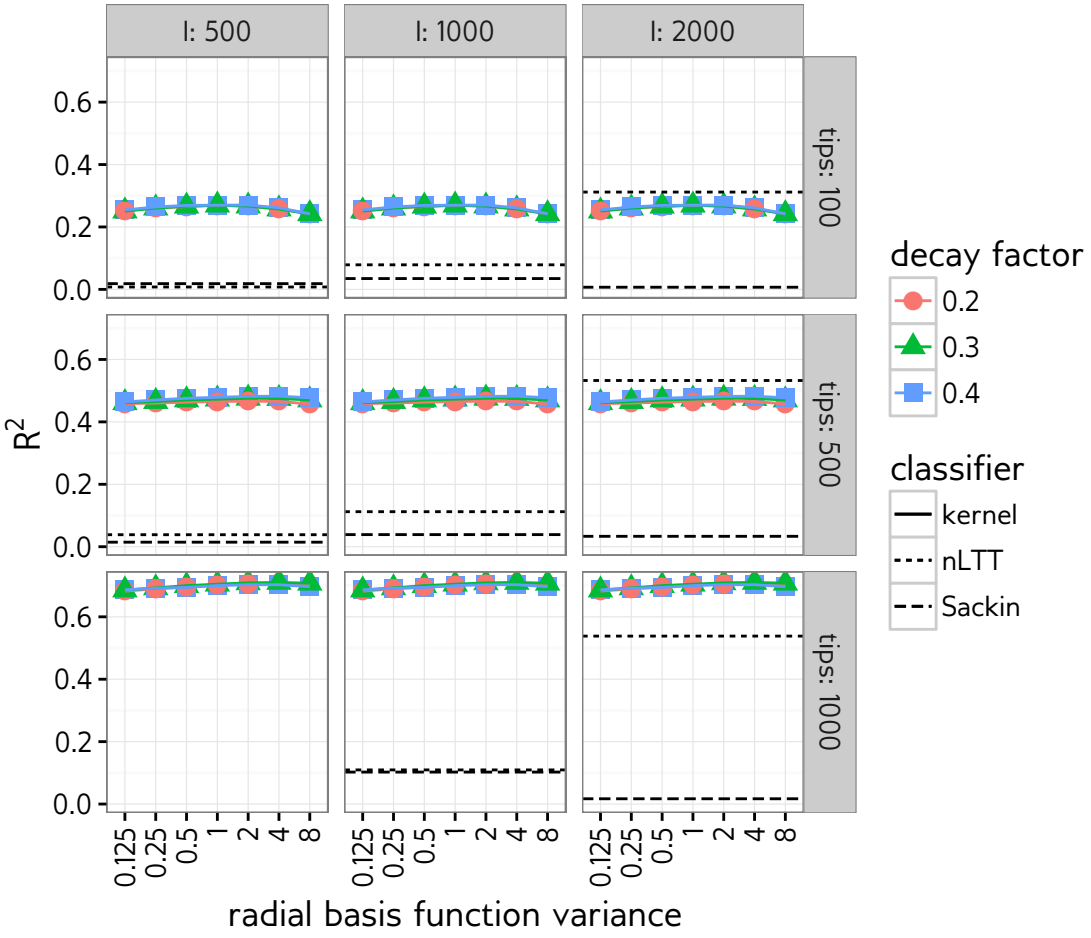

Figure S6: Cross-validation accuracy of kernel-SVM classifiers for total number of nodes (N) under BA network model, for various tree kernel meta-parameters and epidemic scenarios sizes. Each point was calculated based on 300 simulated transmission trees over networks with N = 3000, 5000, or 8000. Dotted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

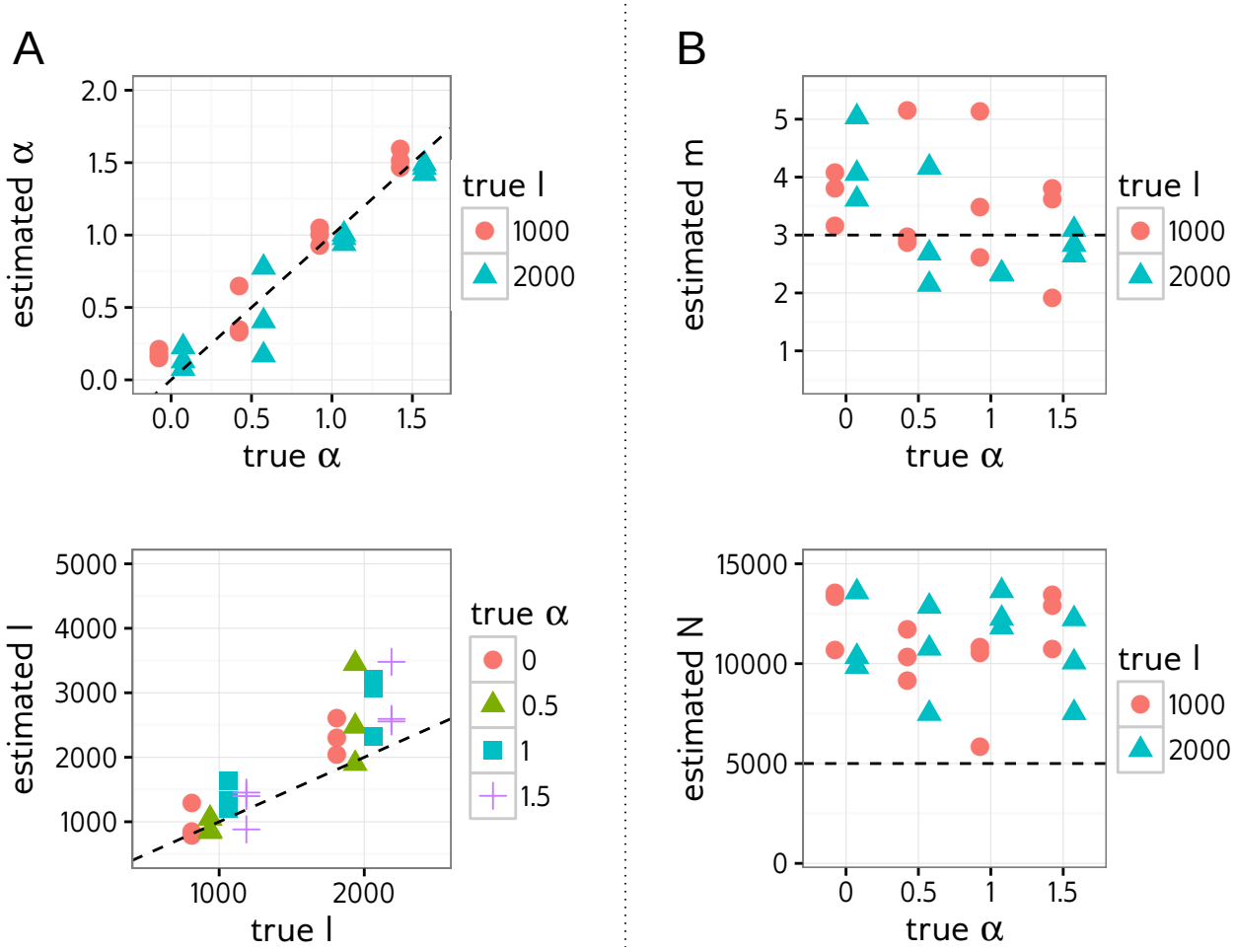

Figure S7: Point estimates of BA model parameters obtained by running kernel-ABC on simulated phylogenies without training, for simulations with $m = 3$. Dotted lines indicate true values, and limits of the y-axes are regions of uniform prior density.

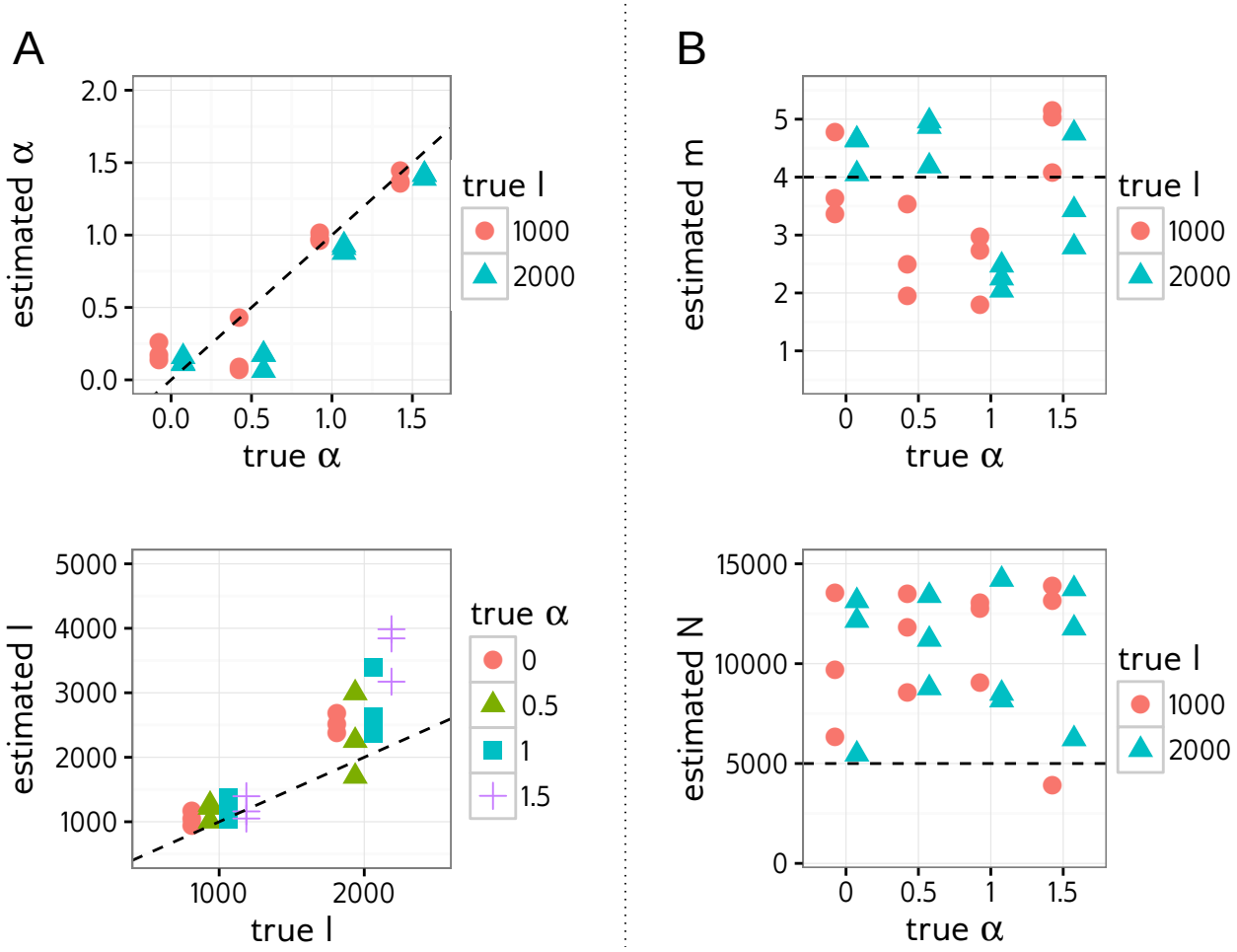

Figure S8: Point estimates of BA model parameters obtained by running kernel-ABC on simulated phylogenies without training, for simulations with $m = 4$. Dotted lines indicate true values, and limits of the y-axes are regions of uniform prior density.

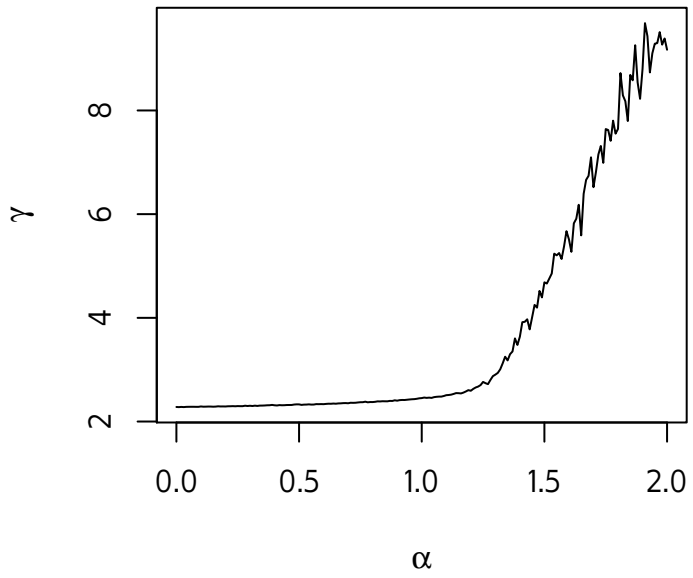

Figure S9: Relationship between preferential attachment power parameter α and power law exponent γ for networks simulated under the BA network model with $N = 5000$ and $m = 2$.

Data S1: Plots of marginal posterior distributions estimated with kernel-ABC for all simulated transmission trees.