Phylodynamic inference of contact network parameters through approximate Bayesian computation

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

March 16, 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, but for sexually transmitted diseases, either sexual or blood-to-blood contact is required. 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) have found that sexual contact networks are best described by a preferential attachment model (although there has been some disagreement, see Jones and Handcock 2003). Under these models, nodes with a high number of contacts attract new connections at an elevated rate. Networks produced by preferential attachment have a power-law degree distribution, meaning that the number of nodes of degree k is proportional to k^{γ} for some constant γ. These networks are also referred to as "scale-free". 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. These new edges are joined to existing nodes of degree k with probability proportional to k^{α}, so that nodes of high degree tend to attract more connections (in the original paper, only $\alpha = 1$ was investigated).

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. 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 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 different 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). Rather than calculating explicit likeilihoods, MCMC uses likelihood *ratios* in which the normalizing constant 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 impossible, 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 susceptible-infected (SI) dynamics, meaning that they became infected at a rate proportional to their numbers 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 ($2^{-3}, 2^{-2}, \ldots, 2^{3}$), as well as on trees with differing numbers of tips (100, 500, and 1000) and in epidemics of differing size I (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 index (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 α = {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: $m \sim$ Uniform(1, 5), $\alpha \sim$ Uniform(0, 2), and (N, I) jointly uniform on the triangular 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 α = 0.5, and the other half with power α = 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 α = 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.

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* ver-

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.

sion 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. 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. Figure 1 shows the cross-validation accuracy with $\lambda = 0.3$, $\sigma = 4$, $I = 1000$, and trees of size 500. The results were largely robust to variations in λ and σ, although accuracy varied between different epidemic and sampling scenarios (figs. S3 to S6).

The kSVM classifier for α had an average R^2 of 0.92 (fig. S3). The R^2 of the for the SVM classifer for α using the nLTT varied between 0.49 and 0.61. The R^2 of the linear regression against Sackin's index varied between 0.73 and 0.76. The I parameter was also classified accurately by the tree kernel (fig. S5), with an average R^2 of 0.93 for 500-tip trees or 0.7 for 100-tip trees. The nLTT performed almost as well on this parameter as the tree kernel, with R^2 values of 0.83 for 500-tip trees or 0.55 for 100-tip trees. Sackin's index was an extremely poor classifier for this parameter, with both R^2 values below 0.08.

The m parameter was much harder to classify (fig. S4), with R^2 for the kSVM varying between 0.004 and 0.36 depending on I and the tree size. The nLTT was also universally poor, with R^2 below 0.02 for all scenarios. Sackin's index fared slightly better than the nLTT, especially on trees with a higher number of tips, with R^2 between 0 and 0.18. Finally, the accuracy of the kSVM classifier for N varied widely (fig. S6), from 0.08 to 0.82. The nLTT performed well in classifying N when the epidemic was large, improving upon the tree kernel in the scenarios with $I = 2000$ and 100- or 500-tip transmission trees ($R^2 = 0.31$ and 0.53), although it performed poorly when I was 500 or 1000 (average $R^2 = 0.07$). Sackin's index could not classify N, with all R^2 below 0.11. Based on inspection of the cross-validation results, we chose to use the meta-parameters $\lambda = 0.3$ and $\sigma = 4$ in further analyses, and not to use Sackin's index or the nLTT.

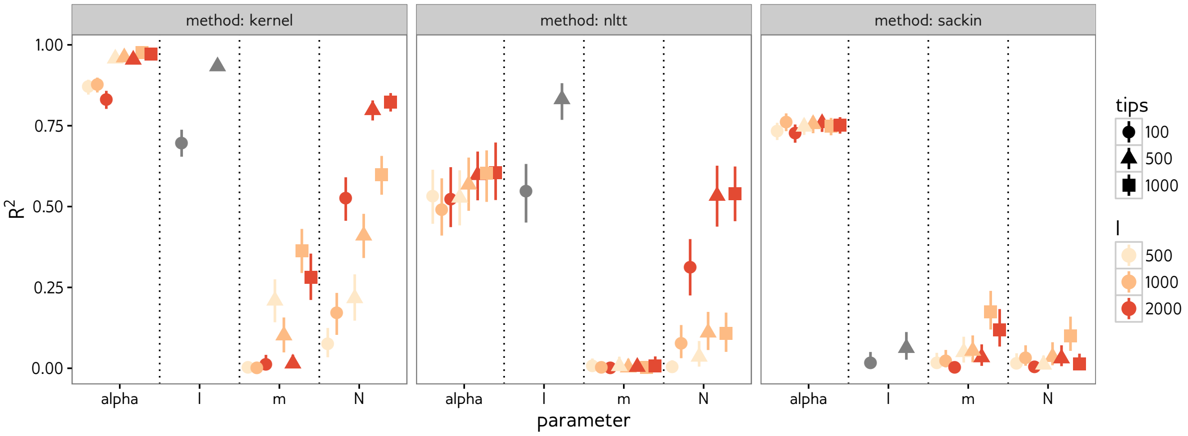

Figure 1: Cross-validation accuracy of kernel-SVM classifier (left), SVM classifier using nLTT (centre), and linear regression using Sackin's index (right) for BA model parameters. Kernel meta-parameters were set to $\lambda = 0.3$ and $\sigma = 4$. Each point was calculated based on 300 simulated transmission trees over networks with three different values of the parameter being tested. Vertical lines are empirical 95% confidence intervals based on 1000 2-fold cross-validations.

ABC simulations

Of the four parameters, α was the most accurately estimated, with a median [IQR] absolute error of 0.11 [0.05-0.18]. Figure 2 shows point estimates for all simulations with $m = 2$. The results $m = 3$ and $m = 4$ were similar (figs. S7 and S8). Average boundaries of 95% HPD intervals are given in table 2. The accuracy of the estimates was not significantly different between values of m or I (both one-way ANOVA, $p = 0.1$ and 0.25), although the errors when the true value of α was zero were significantly greater than the other values (Wilcoxon rank-sum test, $p = 6.41 \times 10^{-4}$). The error in the estimated value of I was 305.66

[107.76-606.59]. Errors were significantly higher for $\alpha \geq 1$ (Wilcoxon rank-sum test, $p = 6.12 \times 10^{-4}$) and for $I = 2000$ ($p = 1.58 \times 10^{-6}$), but not for any values of m (one-way ANOVA, $p = 0.33$). The m parameter was estimated correctly in 37 % of simulations, with an error of one in 40 % and of two or more in 22 % (the only possible m values were 1, 2, 3, 4, or 5). The true values of m and I did not significantly affect the error (one-way ANOVA, $p = 0.5$ and 0.68). Finally, the total number of nodes N was consistently over-estimated by about a factor of two (error 6.59×10^3 [4.21×10^3-8.28×10^3]). No other parameters influenced the accuracy of the N estimates (one-way ANOVA, $p \geq 0.72$).

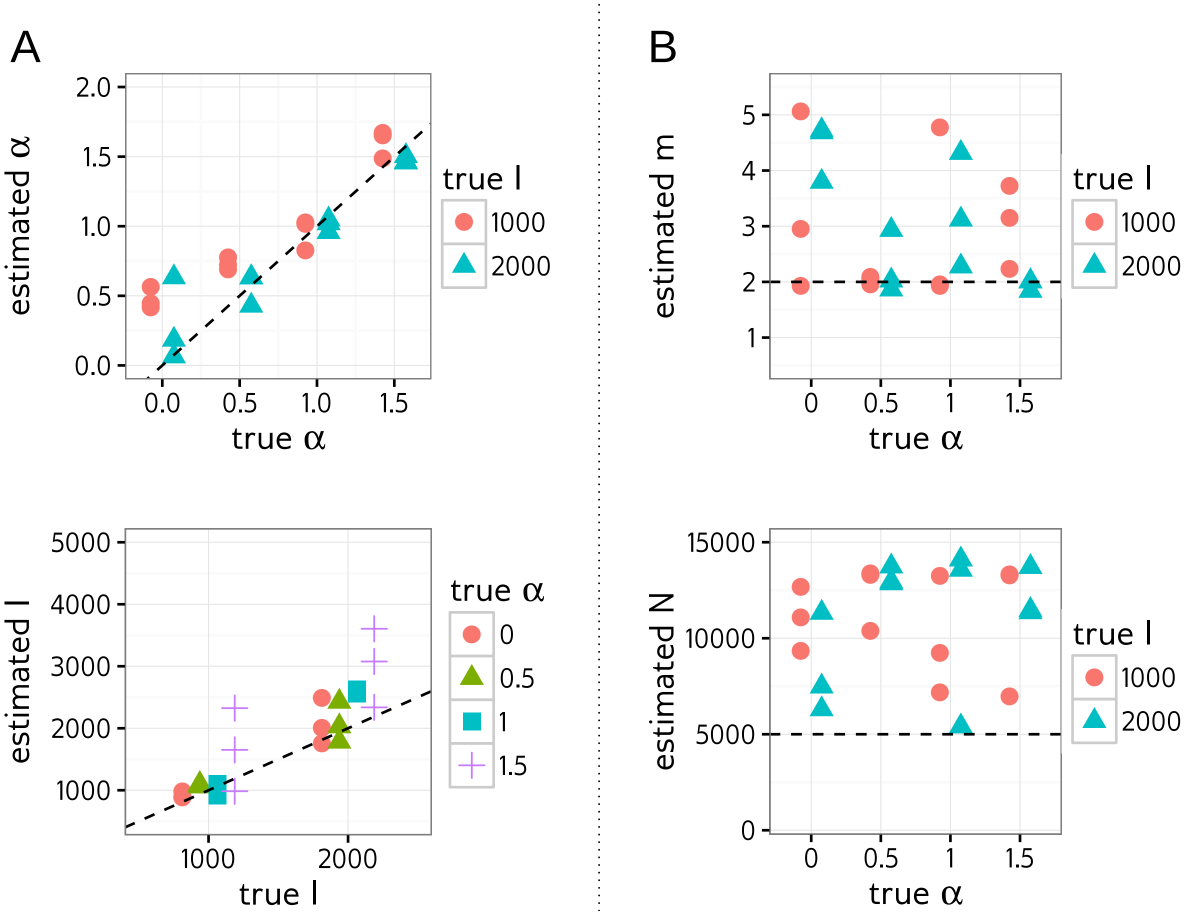

Figure 2: Point estimates of BA model parameters obtained by running kernel-ABC on simulated phylogenies without training, for simulations with $m = 2$. Dotted lines indicate true values, and limits of the y-axes are regions of uniform prior density.

The dispersion of the ABC approximation to the posterior also varied between the parameters, with narrower HPD intervals for those with the most accurate point estimates (table 2). Figure 3 shows the distributions for for one simulation (equivalent plots for all the simulations can be found in supplemental data). 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.

Parameter	True value	Mean point	Mean HPD	Mean HPD
		estimate	upper bound	lower 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 HPD interval widths for ABC model parameter estimates.

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 maximum *a posteriori* (MAP) [95% HPD] estimates for each parameter were: α, 1 [0.07- 1.07]; I, 5.37 [2.05- 5.82]; m, 1.01×10^4 [3.06×10^3- 1.5×10^4]; N, 1.18×10^3 [505.71- 1.62×10^3].

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, 0.46 [0.03- 0.83]; I, 3.53 [2.18- 5.9]; m, 1.07×10^4 [3.11×10^3- 1.47×10^4]; N, 2.46×10^3 [1.28×10^3- 3.88×10^3].

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. (2014), estimated values of I were below 2000, with narrow HPD intervals around two of the datasets (Cuevas et al. (2009), 880.52 [290.7-1.51×10^3]; Niculescu et al. (2015), 175.05 [138.74-454.51]) and wider intervals around the other two (Li et al. (2015), 1.59×10^3 [284.01-3.81×10^3]; Wang et al. (2015), 651.72 [268.94-4.24×10^3]). The Novitsky et al. (2014) data was again the outlier, with a very high estimated I, and HPD interval spanning almost the entire prior region (7.55×10^3 [228.96-8.92×10^3]). Little information was gleaned about the m parameter, with the HPD interval occuping

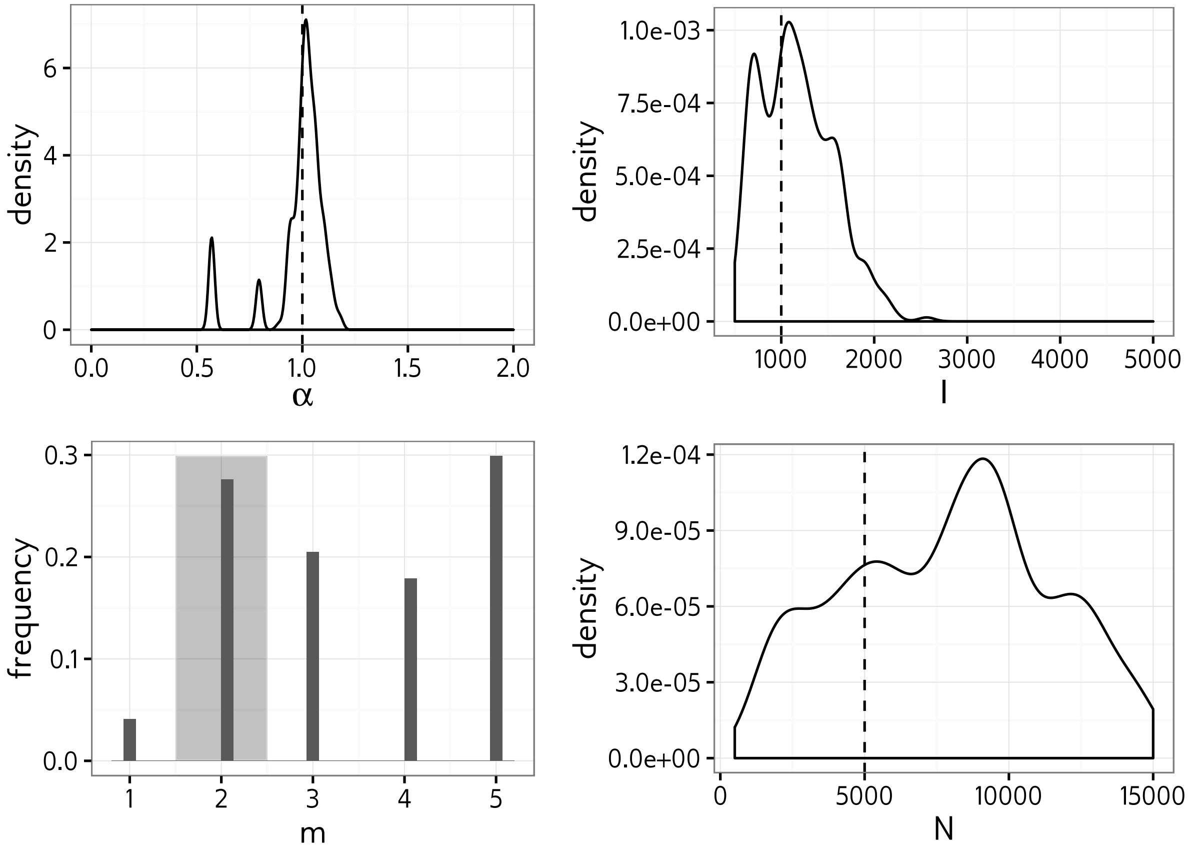

Figure 3: Marginal estimated posterior distributions obtained with ABC on a single simulated dataset. Dotted lines and shaded polygon indicate true values.

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. (2015) data was smaller (5.84×10^3) than the estimates for other datasets.

Discussion

Contact networks can have a strong influence on epidemic progression, and are potentially useful as a public health tool. Despite this, few methods exist for investigating contact network parameters in a phylodynamic framework. 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 incorprorating the preferential attachment feature commonly seen 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

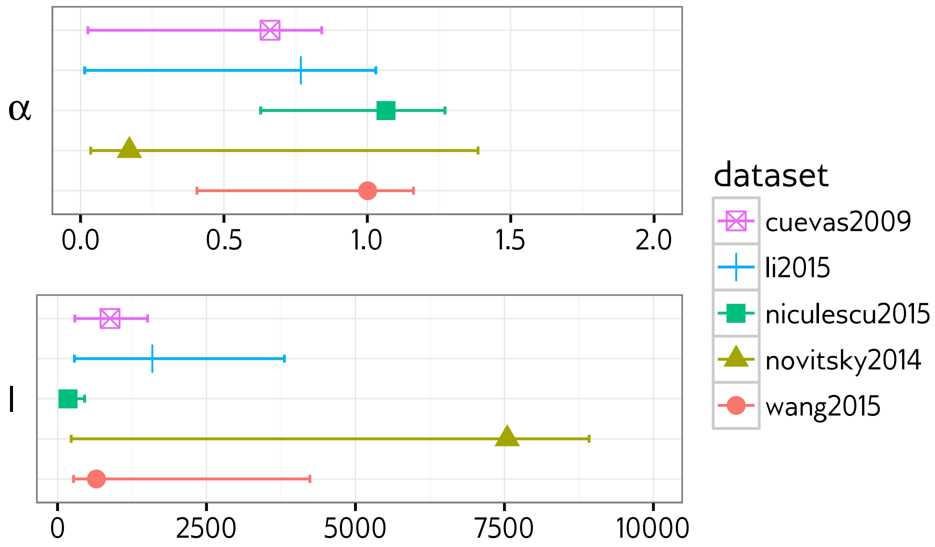

Figure 4: Point estimates and 95% HPD intervals for BA network parameters of published HIV datasets, obtained with kernel-ABC.

on tree shape. This was reflected in the relative accuracy of the kernel-ABC estimates of these parameters. On the other hand, the m and N parameters did not produce much variation in tree shape, resulting in both poorly performing classifiers and uninformative kernel-ABC estimates. The total number of nodes N was almost always significantly overestimated. Since the prior on N and I is jointly uniform on a triangular 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 imaging 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 an exponential 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. Cuevas et al. (2009) studied a group of newly diagnosed individuals in the Basque Country, Spain. Although the individuals were of mixed risk group, and therefore unlikely to comprise a single contact network, a high proportion of them (47%) grouped into local transmission clusters, allowing for the possibility of weak preferential attachment dynamics. We estimated a relatively low attachment power for these data, which is consistent with the sampled sequences comprising many distinct sub-networks. Li et al. (2015) sampled a large number of acutely infected MSM in Shanghai, China. By using a patristic distance cutoff (Poon et al. 2014), we identified and analysed a large cluster. Surprisingly, the estimated attachment power was fairly low with a large credible interval.

Niculescu et al. (2015) studied a recent outbreak among Romainian injection drug user (IDU). The estimated attachment power for this dataset was the highest among all we considered, at slightly above one. The dataset constituted a fairly high sampling prevalence of 33% of newly diagnosed IDU infections over the study period. Wang et al. (2015) sampled acutely infected MSM in Beijing, China, and discovered a high degree of clustering which indicated that the local MSM subnetwork had been deeply sampled. In both these datasets, the estimated attachment power was close to one.

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. The estimated attachment power was extremely low, with a very wide credible interval. The estimated number of infected nodes was also extremely high, much higher than the HIV-infected population of the town, with a similarly wide credible interval. Several factors may have contributed to this result. First, the authors note that the their sample was 75% female. In a primarily heterosexual risk environmet, removal of a disproportionate nmuber of males from the network could obfuscate the true network structure. 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 a reason for the high estimate of I: the contact network includes a much larger group based in the capital city.

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 implausable 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, the epidemic following simple SI dynamics, and the underlying behaviour of all nodes is the same. 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. However, we note that, given the model-agnostic nature of our method, it is possible to fit a model with heterogeneous node behaviour, and indeed this may prove a fruitful avenue for future investigations.

Our method itself has 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). In addition, it is computationally intensive, taking about a day when run on 20 cores in parallel with the setings we described in the methods. However, there are no other methods available to perform the types of estimation we do here, and we therefore believe our method will be useful to epidemiological researchers interested in the network structure underlying disease outbreaks.

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.

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.

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 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.

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

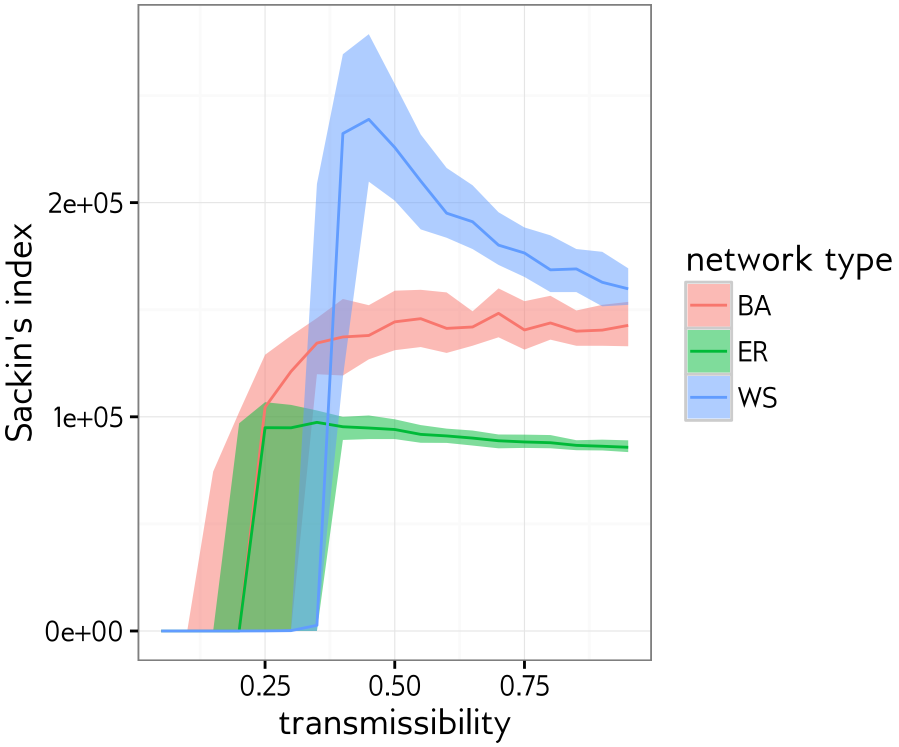

Figure S1: Reproduction of Figure 1A from Leventhal et al. (2012).

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.

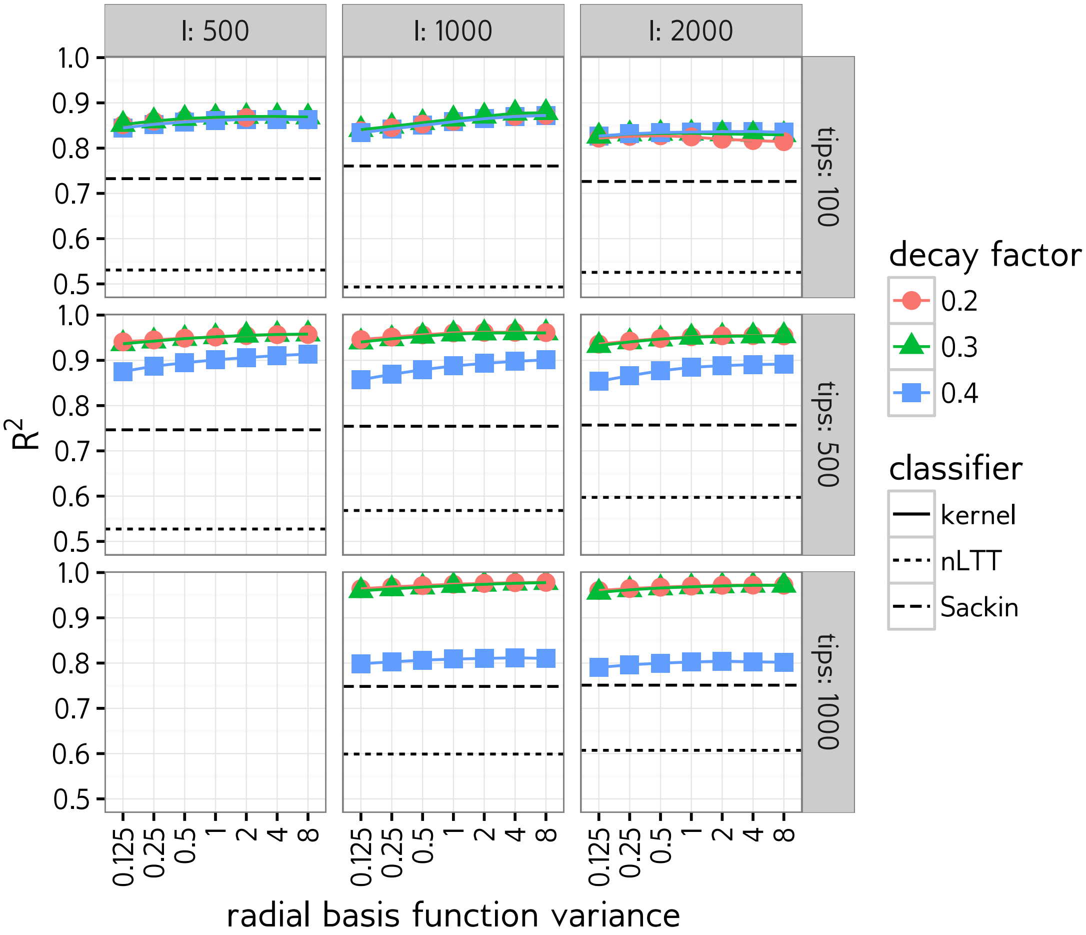

Figure S3: Cross-validation accuracy of kernel-SVM classifiers for α parameter of BA network model, for various tree kernel meta-parameters and epidemic scenarios. Each point was calculated based on 300 simulated transmission trees over networks with α = 0.5, 1.0, or 1.5. Dotted and and dashed lines 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.

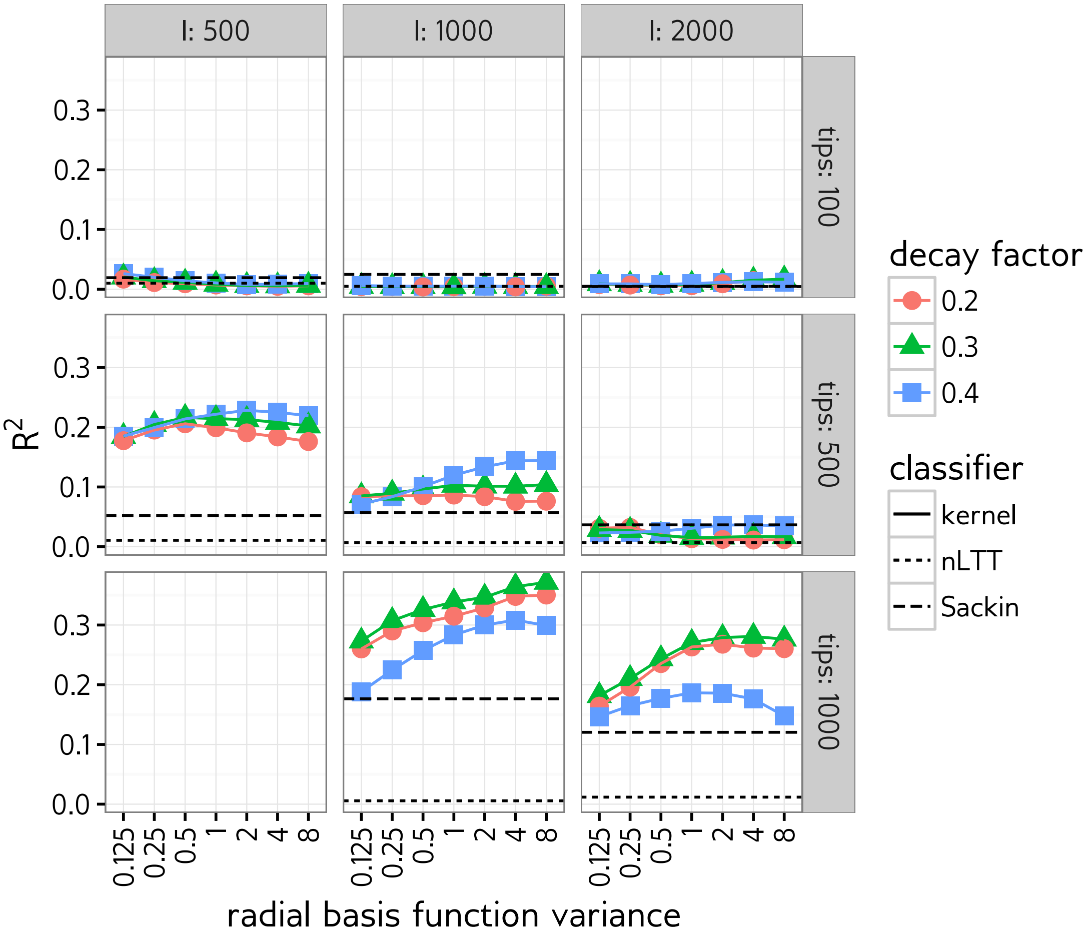

Figure S4: Cross-validation accuracy of kernel-SVM classifiers for m parameter of BA network model, for various tree kernel meta-parameters and epidemic scenarios. Each point was calculated based on 300 simulated transmission trees over networks with $m = 2, 3,$ or 4. Dotted and and dashed lines indicate, 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.

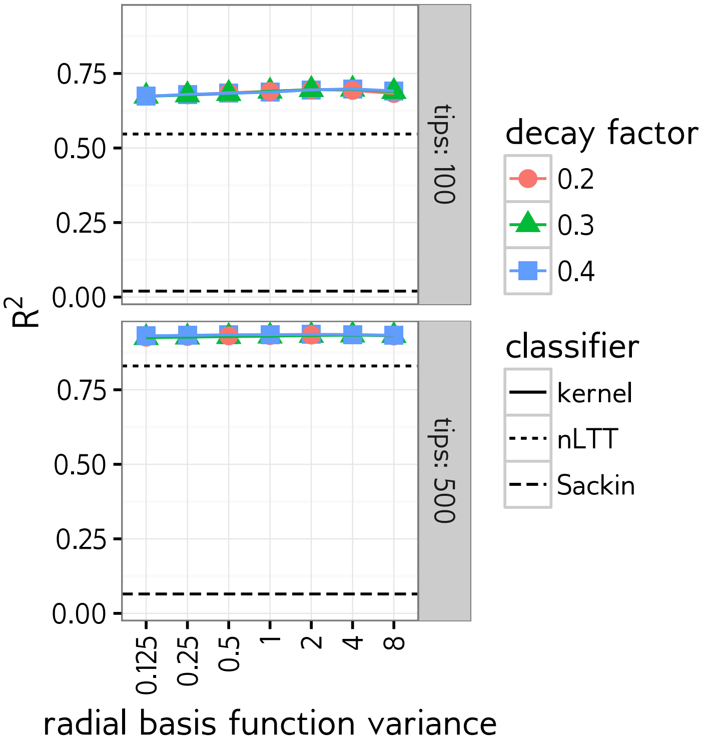

Figure S5: Cross-validation accuracy of kernel-SVM classifiers for number of infected nodes (I) under BA network model, for various tree kernel meta-parameters and two tree sizes. Each point was calculated based on 300 simulated transmission trees over networks with I = 500, 1000, or 2000. Dotted and 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.

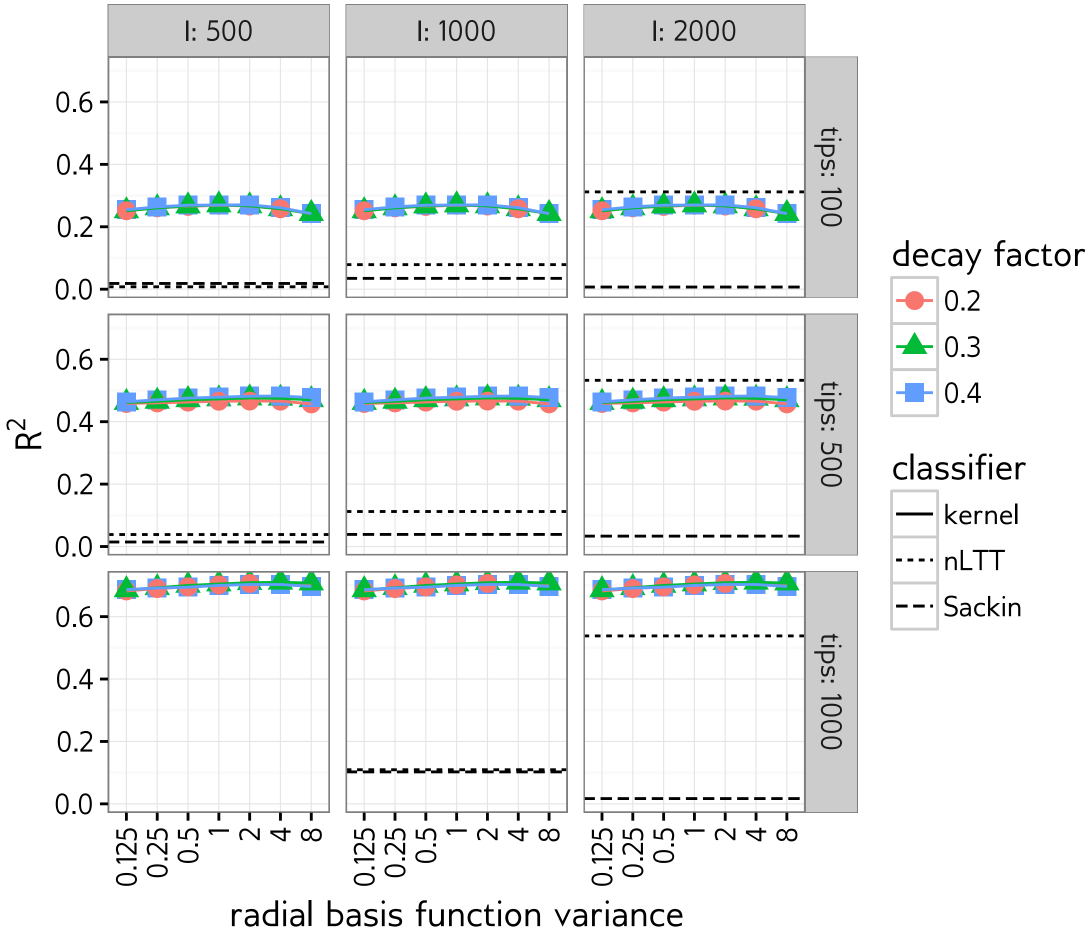

Figure S6: Cross-validation accuracy of kernel-SVM classifiers for total number of nodes (N) under BA network model, for various tree kernel meta-parameters and epidemic scenarios sizes. Each point was calculated based on 300 simulated transmission trees over networks with $N = 3000$, 5000, or 8000. 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.

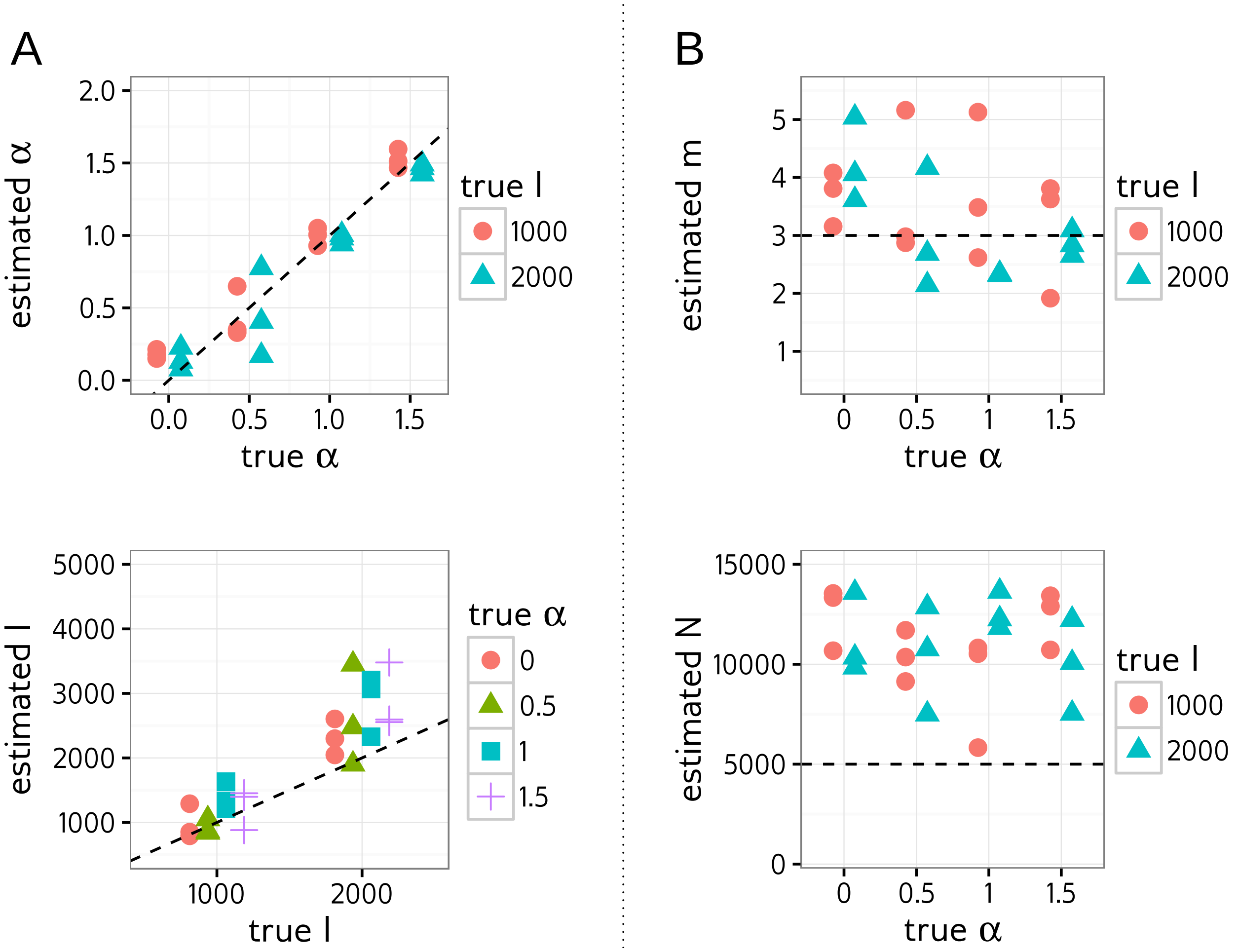

Figure S7: Point estimates of BA model parameters obtained by running kernel-ABC on simulated phylogenies without training, for simulations with $m = 3$. Dotted lines indicate true values, and limits of the y-axes are regions of uniform prior density.

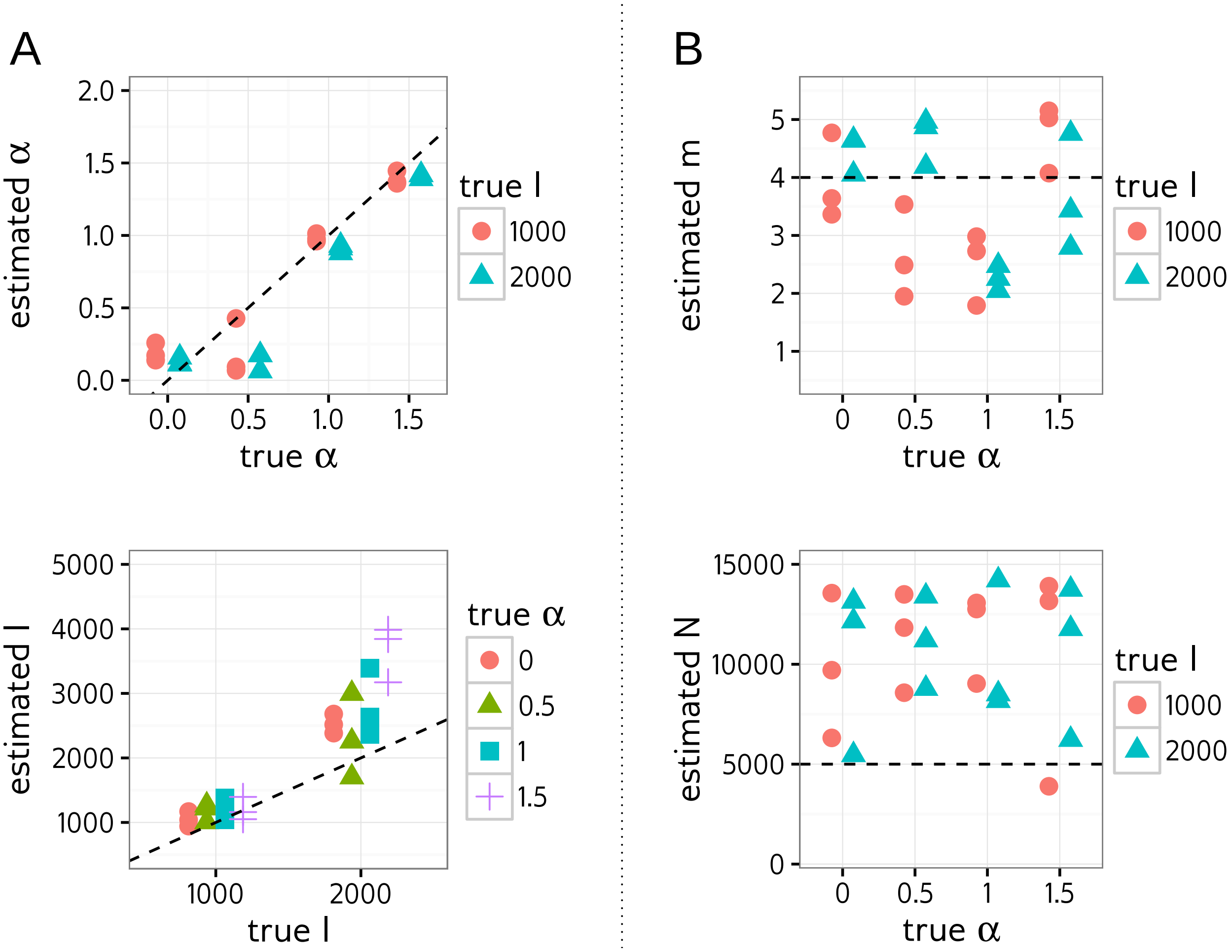

Figure S8: Point estimates of BA model parameters obtained by running kernel-ABC on simulated phylogenies without training, for simulations with $m = 4$. Dotted lines indicate true values, and limits of the y-axes are regions of uniform prior density.

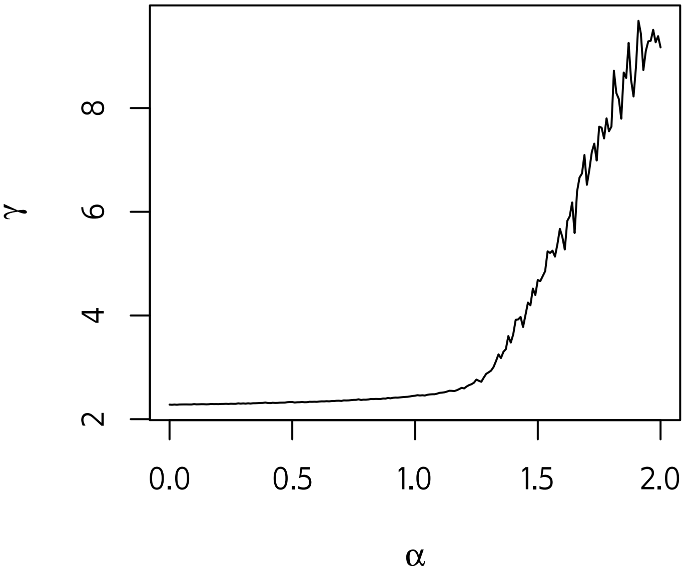

Figure S9: Relationship between preferential attachment power parameter α and power-law exponent γ for networks simulated under the BA network model with $N = 5000$ and $m = 2$.