Phylogenetic estimation of contact network parameters with kernel approximate Bayesian computation

by

Rosemary Martha McCloskey

B.Sc., Simon Fraser University, 2014

A thesis submitted in partial fulfillment of the requirements for the degree of

Master of Science

in

The Faculty of Graduate and Postdoctoral Studies

Bioinformatics

The University of British Columbia
(Vancouver)

August 2016

Abstract

Preface

The initial idea to use kernel approximate Bayesian computation to infer contact network model parameters was Dr. Poon's. The tree kernel was originally developed by Dr. Poon, but the version used here was implemented by me to improve computational efficiency. The idea to apply sequential Monte Carlo is credited to Dr. Alexandre Bouchard-Côté. Dr. Sarah Otto suggested the experiments involving a network with a heterogeneous α parameter and peer-driven sampling. Dr. Richard H. Liang provided guidance in the development of the Gillespie simulation algorithm. The *netabc* program, and all supplementary analysis programs, were written by me.

A version of chapter 2 has been submitted for publication in Virus Evolution with the title "Reconstructing network parameters from viral phylogenies." A presentation entitled "Phylodynamic inference of contact network parameters with kernel-ABC" was given based on chapter 2 at the 23rd HIV Dynamics and Evolution meeting on April 25, in Woods Hole, Massachussets, USA.

Throughout this work, the pronouns *we* and *our* refer to Rosemary M. McCloskey unless otherwise stated.

Contents

List of Tables

List of Figures

List of Symbols

I number of nodes which are eventually infected.

N number of nodes in the network.

α preferential attachment power parameter in Barabási-Albert networks.

λ decay factor meta-parameter for tree kernel.

σ radial basis function variance meta-parameter for tree kernel.

m number of edges added per vertex when constructing a Barabási-Albert network.

List of Abbreviations

ABC approximate Bayesian computation.

BA Barabási-Albert.

ER Erdős-Rényi.

ESS expected sample size.

GLM generalized linear model.

GSL GNU scientific library.

GTR generalized time-reversible.

HIV human immunodeficiency virus.

HMM hidden Markov model.

HPD highest posterior density.

IS importance sampling.

kPCA kernel principal components analysis.

kSVR kernel support vector regression.

LTT lineages-through-time.

MAP maximum *a posteriori*.

MCMC Markov chain Monte Carlo.

MH Metropolis-Hastings.

ML maximum likelihood.

nLTT normalized lineages-through-time.

ODE ordinary differential equation.

pdf probability density function.

SARS severe acute respiratory syndrome.

SI susceptible-infected.

SIR susceptible-infected-recovered.

SIS sequential importance sampling.

SMC sequential Monte Carlo.

SVM support vector machine.

SVR support vector regression.

TasP treatment as prevention.

WS Watts-Strogatz.

Acknowledgements

Chapter 1

Introduction

1.1 Objective

The spread of a disease is most often modelled by assuming either a homogeneously mixed population [1, 2], or a population divided into a small number of homogeneously mixed groups [3]. This assumption, also called the "law of mass action" or *panmixia*, implies that any two individuals in the same compartment are equally likely to come into contact causing transmission. Although this provides a reasonable approximation in many cases, the error introduced by assuming a panmictic population can be substantial when significant contact heterogeneity exists in the underlying population. Contact network models provide an alternative to compartmental models which do not require the assumption of panmixia. In addition to more accurate predictions, the parameters of the networks themselves may be of interest from a public health perspective. For example, certain vaccination strategies may be more or less effective in curtailing an epidemic depending on the underlying network's degree distribution. Phylodynamic methods have been used to fit many different types of model to phylogenetic data, but as far as we know, no methods have yet been developed to fit contact network models. The primary objective of this work is to develop such a method.

Calculating the likelihood of the parameters of a contact network models seems likely to be an intractable problem, which would imply that these models are amenable to neither maximum likelihood (ML) nor Bayesian inference. We not proven this is the case, but some intuition can be provided by examining the process involved in the likelihood calculation. Consider a contact network model with parameters θ, and an observed transmission tree T with n tips. In general, we do not know the labels of the internal nodes of T, only the labels of its tips. To fit this model using likelihood-based methods, we must calculate the likelihood of θ, that is, $\Pr(T \mid \theta)$. Let $\mathcal{G}$ be the set of all possible contact networks, and $\mathcal{N}$ be the set of all possible labellings of the internal nodes of T. We can write the likelihood as

$$\begin{aligned}\Pr(T \mid \theta) &= \sum_{v \in \mathcal{N}} \Pr(T, v \mid \theta) \\ &= \sum_{G \in \mathcal{G}} \sum_{v \in \mathcal{N}} \Pr(T, v \mid G, \theta) \Pr(G \mid \theta) \\ &= \sum_{G \in \mathcal{G}} \sum_{v \in \mathcal{N}} \Pr(T, v \mid G) \Pr(G \mid \theta),\end{aligned} \tag{1.1}$$

the last equality following from the fact that T and v depend only on G, not on θ. Although $\Pr(T, v \mid G)$ and $\Pr(G \mid \theta)$ may individually be straightforward to calculate, the number of possible directed graphs on N nodes is $2^{N(N-1)}$, larger if the nodes and edges in the graph may have different labels or attributes. Hence, the number of terms in the sum is at least exponential in n, as there must be at least n nodes in the network. In addition, eq. (1.1) assumes that T is complete, meaning that all infected individuals were sampled. This is rarely the case in practice - most often, the observed tree is a subsampled version of the true tree. In this case, the likelihood calculation becomes even more complex, because we must also sum over all possible complete trees.

Depending on the network model studied, it is possible that eq. (1.1) could be simplified into a tractable expression. However, a simpler alternative to likelihood-based methods, which would apply to any network model, is provided by ABC. All of the ingredients required to apply ABC to this problem are readily available. Simulating networks is straightforward under a variety of models. Epidemics on those networks, and the corresponding transmission trees, can also be easily simulated. As mentioned above, contact networks can profoundly affect transmission tree shape, and those shapes can be compared using a highly informative similarity measure. SMC has several advantages over other algorithms for ABC [4], including a recently-developed adaptive algorithm requiring minimal tuning on the part of the user [5]. In summary, our method to infer contact network parameters will combine the following: stochastic simulation of epidemics on networks, the tree kernel, and adaptive ABC-SMC. Since our distance measure is a kernel function, our method is a type of kernel-ABC. For ease of exposition, we will often use the term “kernel-ABC” to refer to our method specifically.

Empirical studies of sexual contact networks have found that these networks tend to be scale-free, meaning that their degree distributions follow a power law (although there has been some disagreement). Preferential attachment has been postulated as a mechanism by which scale-free networks could be generated. This makes the BA model, one of the simplest preferential attachment models, a natural choice to explore with our method. The second aim of this work is to use simulations to investigate the parameters of the BA model, including whether they have a detectable impact on tree shape, and whether they can be accurately recovered using kernel-ABC.

Due to its high global prevalence and fast mutation rate, human immunodeficiency virus (HIV) is one of the most commonly-studied viruses in a phylodynamic context. Consequently, a large volume of HIV sequence data is publicly available, more than for any other pathogen, and including sequences sampled from diverse geographic and demographic contexts. Since HIV is almost always spread through either sexual contact or sharing of injection drug supplies, the contact networks underlying HIV epidemics are highly structured. Moreover, since no cure yet exists, efforts to curtail the progression of an epidemic have relied on preventing further transmissions through measures such as treatment as prevention (TasP) and education leading to behaviour change. The effectiveness of this type of intervention can vary significantly based on the underlying structure of the network and the particular nodes to whom the intervention is targeted. Due to this combination of data availability and potential public health impact, HIV is an obvious context in which our method could be applied. Therefore, the third and final aim of this work is to apply kernel-ABC to fit the BA model to existing HIV outbreaks.

To summarize, this work has three objectives. First, we will develop a method which uses kernel-ABC to infer parameters of contact network models from observed transmission trees. Second, we will use simulations to characterize the parameters of the BA network model in terms of their effect on tree shape and how accurately they can be recovered with kernel-ABC. Finally, we will apply the method fit the BA model to several real-world HIV datasets.

1.2 Phylogenetics and phylodynamics

1.2.1 Phylogenetic trees

In evolutionary biology, a *phylogeny*, or *phylogenetic tree*, is a graphical representation of the the evolutionary relationships among a group of organisms or species (generally, *taxa*) [6]. The *tips* of a phylogeny, that is, the nodes without any descendants, correspond to *extant*, or observed, taxa. The *internal nodes* correspond to their (usually extinct) common ancestors. The edges or *branches* of the phylogeny connect ancestors to their descendants. Phylogenies may have a *root*, which is a node with no descendants distinguished as the most recent common ancestor of all the extant taxa [7]. When such a root exists, the tree is referred to as being *rooted*; otherwise, it is *unrooted*. The structural arrangement of nodes and edges in the tree is referred to as its *topology* [8].

The branches of the tree may have associated lengths, representing either evolutionary distance or calendar time between ancestors and their descendants. The term "evolutionary distance" is used here imprecisely to mean any sort of quantitative measure of evolution, such as the number of differences between the DNA sequences of an ancestor its descendant, or the difference in average body mass or height. A phylogeny with branch lengths in calendar time units is often referred to as *time-scaled*. In a time-scaled phylogeny, the internal nodes can be mapped onto a timeline by using the tips of the tree, which usually correspond to the present day, as a reference point [9]. The corresponding points on the timeline are called *branching times*, and the rate of their accumulation is referred to as the *branching rate*. Rooted trees whose tips are all the same distance from the root are called *ultrametric* trees [10]. These concepts are illustrated in fig. 1.1.

1.2.2 Transmission trees

In epidemiology, a *transmission tree* is a graphical representation of an epidemic's progress through a population. Like phylogenies, transmission trees have tips, nodes, edges, and branch lengths. How-

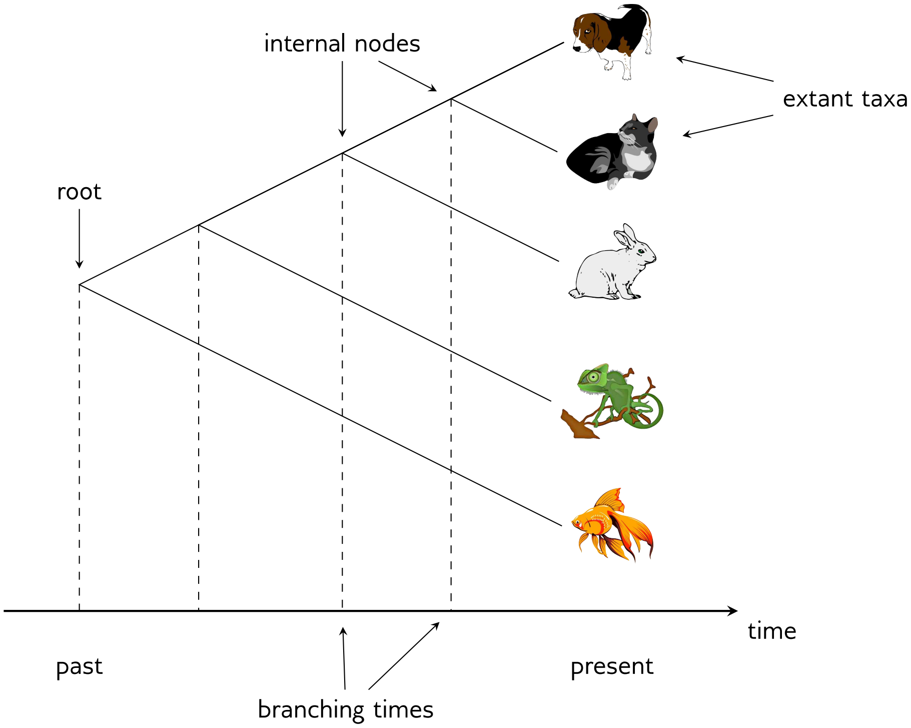

Figure 1.1: Illustration of a rooted, ultrametric, time-scaled phylogeny. The tips of the tree, which represent extant taxa, are placed at the present day on the time axis. Internal nodes, representing extinct common ancestors to the extant taxa, fall in the past. The topology of the tree indicates that cats and dogs are the most closely related pair of species, whereas fish is most distantly related to any other node in the tree.

ever, rather than recording an evolutionary process (speciation), they record an epidemiological process (transmission). The tips of a transmission tree represent the removal of infected hosts, while internal nodes correspond to transmissions from one host to another.Transmission trees generally have branch lengths in units of calendar time, with branching times indicating times of transmission. The root of a transmission tree corresponds to the initially infected patient who introduced the epidemic into the network, also known as the *index case*. The internal nodes may be labelled with the donor of the transmission pair, if this is known. The tips of the tree, rather than being fixed at the present day, are placed at the time at which the individual was removed from the epidemic, such as by death, recovery, isolation, behaviour change, or migration. Consequently, the transmission tree may not be ultrametric, but may have tips located at varying distances from the root. Such trees are said to have *heterochronous* taxa [11], in contrast to the *isochronous* taxa found in most phylogenies of macro-organisms. A transmission tree is illustrated in fig. 1.2 (right).

Each infected individual in an epidemic may appear in the transmission tree more than once. This is different from the transmission *network*, in which each infected individual appears exactly once, and edges are in one-to-one correspondence with transmissions [12, 13]. Transmission networks are discussed further in section 1.3, and the distinction between the two objects is illustrated in fig. 1.2. However, since transmission networks generally have no cycles (unless re-infection occurs), they are

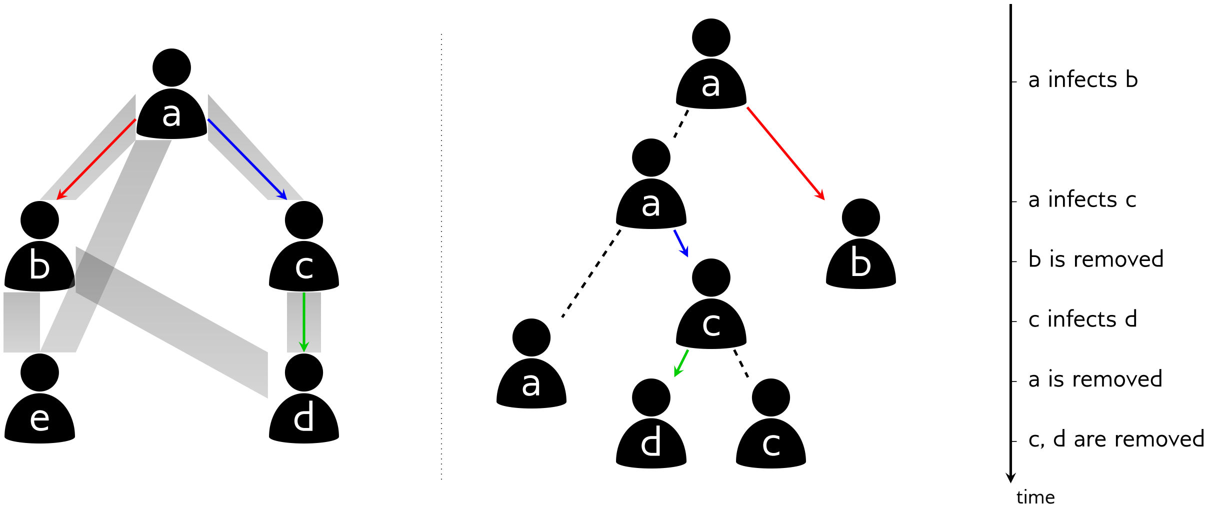

Figure 1.2: Illustration of epidemic spread over a contact network, and the corresponding transmission tree. (Left) A contact network with five hosts, labelled a through e. Thick shaded edges indicate symmetric contacts among the hosts. The transmission network is indicated by coloured arrows. The epidemic began with node a, who transmitted to nodes b and c. Node c further transmitted to node d. Node e was not infected. (Right) The transmission tree corresponding to this scenario, with a timeline of transmission and removal times.

trees in the graph theoretical sense, and hence are sometimes also referred to as transmission trees [*e.g.* 14]. In this work, we reserve the term "transmission tree" for the objects depicted on the right side of fig. 1.2, following *e.g.* [15]. The term "transmission network" is taken to mean the subgraph of the contact network along which transmissions occurred, following *e.g.* [12, 13].

Since transmission trees are essentially a detailed record of an epidemic's progress, they contain substantial epidemiological information. As a basic example, the lineages-through-time (LTT) plot [9], which plots the number of lineages in a phylogeny against time, can be used to quantify the incidence of new infections over the course of an epidemic [16]. However, in all but the most well-studied of epidemics, transmission trees are not possible to assemble through traditional epidemiological methods [12]. The time and effort to conduct detailed interviews and contact tracing of a sufficient number of infected individuals is usually prohibitive, and may be additionally be confounded by misreporting [17]. However, it turns out that for viral epidemics, some of the epidemiological information contained in the transmission tree leaves a mark on the viral genetic material circulating in the population. A family of methods called *phylodynamics* [18] addresses the challenge of estimating epidemiological parameters from viral sequence data [19].

1.2.3 Phylodynamics: linking evolution and epidemiology

The basis of phylodynamics is the fact that, for RNA viruses, epidemiological and evolutionary processes occur on similar time scales [11]. In fact, these two processes interact, such that it is possible to detect the influence of host epidemiology on the evolutionary history of the virus as recorded in an *inter-host viral phylogeny*. Phylodynamic methods aim to detect and quantify the signatures of epidemiological processes in these phylogenies [19, 20], which relate one representative viral genotype from each host in an infected population. These methods have been used to investigate parameters such as transmission rate, recovery rate, and basic reproductive number [19, 20]. The majority of phylodynamic studies attempt to infer the parameters of an epidemiological model for which the likelihood of an observed phylogeny can be calculated. Most often, this is some variation of the birth-death [21, 22] or coalescent [23, 24] models. These methods generally assume the viral phylogeny is known, and we assume the same in this work (see *e.g.*[25] for a review of phylogenetic inference).

Due to the relationship between the aforementioned processes, there is a degree of correspondence between viral phylogenies and transmission trees [14, 26, 27]. In particular, the transmission process is quite similar to *allopatric speciation* [28], where genetic divergence follows the geographic isolation of a sub-population of organisms. Thus, transmission, which is represented as branching in the transmission tree, causes branching in the viral phylogeny as well. Similarly, the removal of an individual from the transmission tree causes the extinction of their viral lineage in the phylogeny. Consequently, the topology of the viral phylogeny is sometimes used as a proxy for the topology of the transmission tree. Modern likelihood-based methods of phylogenetic reconstruction [*e.g.* 29, 30] produce unrooted trees whose branch lengths measure genetic distance in units of expected substitutions per site. On the other hand, transmission trees are rooted, and have branches measuring calendar time [20]. Therefore, estimating a transmission tree from a viral phylogeny requires the phylogeny to be rooted and time-scaled. Methods for performing this process include root-to-tip regression [31–33], which we apply in this work, and least-square dating [34]. Alternatively, the tree may be rooted separately with an outgroup [35] before time-scaling.

A caveat of estimating transmission trees in this manner is that the correspondence between the topologies of the viral phylogeny and transmission tree is not necessarily exact [26]. Due to intra-host diversity, the viral strain which is transmitted may have split from another lineage within the donor long before the transmission event occurred. Hence, the branching point in the viral phylogeny may be much earlier than that in the transmission tree. Another possibility is that one host transmitted to two or more recipients, but the lineages they each received originated within the donor host in a different order than that in which the transmissions occurred. In this case, the topology of the transmission tree and the viral phylogeny will be mismatched. In practice, this discordance has not proven an insurmountable problem: for example, Leitner et al. [36] were able to accurately recover a known transmission tree using a viral phylogeny. The problem of accurately estimating transmission trees is an ongoing area of research [37–40].

1.2.4 Tree shapes

To perform phylodynamic inference, we must be able to extract quantitative information from viral phylogenies. What is informative about a phylogeny, beyond the demographic characteristics of the individuals it relates, is its *shape*. The shape of a phylogeny has two components: the topology, and the distribution of branch lengths [41]. Methods of quantifying tree shape fall into two categories: summary statistics, and pairwise measures. Summary statistics assign a numeric value to each individual tree, while pairwise measures quantify the similarity between pairs of trees.

One of the most widely used tree summary statistics is Sackin's index [42], which measures the imbalance or asymmetry in a rooted tree. For the ith tip of the tree, we define N_i to be the number of branches between that tip and the root. The unnormalized Sackin's index is defined as the sum of all N_i. It is called unnormalized because it does not account for the number of tips in the tree. Among two trees having the same number of tips, the least-balanced tree will have the highest Sackin's index. However, among two equally balanced trees, the larger tree will have a higher Sackin's index. This makes it challenging to compare balances among trees of different sizes. To correct this, Kirkpatrick and Slatkin [43] derive the expected value of Sackin' index under the Yule model [44]. Dividing by this expected value normalizes Sackin's index, so that it can be used to compare trees of different sizes. An example of a pairwise measure is the nLTT [45], which compares the LTT [9] plots of two trees. Specifically, the two LTT plots are normalized so that they begin at $(0, 0)$ and end at $(1, 1)$, and the absolute difference between the two plots is integrated between 0 and 1. In the context of infectious diseases, the LTT is related to the prevalence [16], so large values may indicate that the trees being compared are the products of different epidemic trajectories [45].

Poon et al. [46] developed an alternative pairwise measure which applies the concept of a *kernel function* to phylogenies. Kernel functions, originally developed for support vector machines (SVMs) [47], compare objects in a space $\mathcal{X}$ by mapping them into a feature space $\mathcal{F}$ of high or infinite dimension via a function φ. The similarity between the objects is defined as

$$K(x, x') = \langle \varphi(x), \varphi(x') \rangle,$$

that is, the inner product of the objects' representations in the feature space. Computing $\varphi(x)$ may be computationally prohibitive due to the dimension of $\mathcal{F}$. The utility of a kernel function K is that it is constructed in such a way that it can compute the inner product without explicitly computing $\varphi(x)$. The kernel function developed in [46] will henceforth be referred to as the *tree kernel*. The feature space into which trees are mapped is the space of all possible *subset trees*, which are subtrees that do not necessarily extend all the way to the tips. The subset-tree kernel was originally developed for comparing parse trees in natural language processing [48] and did not incorporate branch length information. The version developed by Poon et al. [46] includes a radial basis function to compare the differences in branch lengths, thus incorporating both the trees' topologies and their branch lengths in a single similarity score.

The kernel score of a pair of trees, denoted $K(T_1, T_2)$, is defined as a sum over all pairs of nodes (n_1, n_2), where n_1 is a node in T_1 and n_2 is a node in T_2. Folloting Poon et al. [46], let $N(T)$ denote the set of all nodes in T, $\text{nc}(n)$ be the number of children of node n, c_n^j be the jth child of node n, and l_n be the vector of branch lengths connecting node n to its descendants. The *production rule* of n is its total number of children, and its number of leaf children. That is, if two nodes have the same number of children and among these, the same number of leaves, then they have the same production rule. Let $k_G(x, y)$ be a Gaussian radial basis function of the vectors x and y,

$$k_G(x, y) = \exp\left(-\frac{1}{2\sigma}\,\|x - y\|_2^2\right),$$

where $\|\cdot\|_2$ is the Euclidian norm and σ is a variance parameter. The tree kernel is defined as

$$K(T_1, T_2) = \sum_{n_1 \in N(T_1)} \sum n_2 \in N(T_2)\Delta(n_1, n_2),$$

where

$$\Delta(n_1, n_2) = \begin{cases} \lambda & n_1 \text{ and } n_2 \text{ are leaves} \\ \lambda k_G(l_{n_1}, l_{n_2}) \displaystyle\prod_{j=1}^{\mathrm{nc}(n_1)} \left(1 + \Delta(c_{n_1}^j, c_{n_2}^j)\right) & n_1 \text{ and } n_2 \text{ have the same production rule} \\ 0 & \text{otherwise.} \end{cases}$$

Here λ is a decay factor parameter, which penalizes large matches that tend to dominate the kernel score. In this work, we refer to the parameters λ and σ as *meta-parameters*, to avoid confusing them with model parameters we are trying to estimate.

1.3 Contact networks

1.3.1 Overview

Epidemics spread through populations of hosts through *contacts* between those hosts. The definition of contact depends on the mode of transmission of the pathogen in question. For an airborne pathogen like influenza, a contact may be simple physical proximity, while for HIV, contact could be via unprotected sexual relations or blood-to-blood contact (such as through needle sharing). A *contact network* is a graphical representation of a host population and the contacts among its members [13, 49, 50]. The *nodes* in the network represent hosts, and *edges* or *links* represent contacts between them. A contact network is shown in fig. 1.2 (left). Contact networks are a particular type of *social network* [51, 52], which is a network in which edges may represent any kind of social or economic relationship. Social networks are frequently used in the social sciences to study phenomena where relationships between people or entities are important [for a review see 53].

Edges in a contact networks may be *directed*, representing one-way transmission risk, or *undirected*, representing symmetric transmission risk. For example, a network for an airborne epidemic would use undirected edges, because the same physical proximity is required for a host to infect or to become infected. However, an infection which may be spread through blood-to-blood contact through transfusions transfusions would use directed edges, since the donor has no chance of transmitting to the recipient. Directed edges are also useful when the transmission risk is not equal between the hosts, such as with HIV transmission, where acting as the receptive partner carries a higher risk of infection than acting as the insertive partner. In this case, a contact could be represented by two directed edges, one in each direction between the two hosts, with the edges annotated by what kind of risk they imply [53]. An undirected contact network is equivalent to a directed network where each contact is represented by two symmetric directed edges. The *degree* of a node in the network is how many contacts it has. In directed networks, we may make the distinction between *out-degree* and *in-degree*, which count respectively the number incoming and outgoing edges. The *degree distribution* of a network denotes the probability that a node has any given number of links. The set of edges attached to a node are referred to as its *incident* edges.

Epidemiological models most often assume some form of contact homogeneity. The simplest models, such as the susceptible-infected-recovered (SIR) model, assume a completely homogeneously mixed population, where every pair of contacts is equally likely. More sophisticated models partition the population into groups with different contact rates between and among each group. However, these models still assume that every possible contact between a member of group i and a member of group j is equally likely. This assumption is clearly unrealistic for the majority of human communities, and can lead to significant errors in predicted epidemic trajectories when there is substantial heterogeneity present [54, 55]. Contact networks provide a way to relax this assumption by representing individuals and their contacts explicitly. It is important to note that, although panmixia is an unrealistic modelling assumption, it has not proven a substantial hurdle to epidemic modelling in practice [56]. Using this assumption, researchers have been able to derive estimates of the transmission rate and the basic reproductive number of various outbreaks, which have agreed with values obtained by on-the-ground data collection. Therefore, if one is interested only in these population-level variables, the additional complexity of contact network models may not be warranted. Rather, these models are most useful when we are interested in properties of the network itself, such as centrality, structural balance, and transitivity [53].

From a public health perspective, knowledge of contact networks has the potential to be extremely useful. On a population level, network structure can dramatically affect the speed and pattern of epidemic spread [*e.g.* 57, 58]. For example, epidemics are expected to spread more rapidly in networks having the "small world" property, where the average path length between two nodes in the network is relatively low [59]. Some sexually transmitted infections would not be expected to survive in a homogeneously mixed population, but their long-term persistence can be explained by contact heterogeneity [56, 60]. Hence, the contact network can provide an idea of what to expect as an epidemic unfolds. In terms of actionable information, vaccination strategies which would eradicate an epidemic in a random network might not work if the network is scale-free [13, see section 1.3.2]. On a local level, contact networks can be informative about the groups or individuals who are at highest risk of acquiring or transmitting infection, and would therefore benefit most from public health interventions [61, 62].

Contact networks are a challenging type of data to collect, requiring extensive epidemiological investigation in the form of contact tracing [12, 13, 17, 50]. Therefore, it has been necessary to explore less resource-intensive alternatives which still contain information about population structure. For instance, it is possible to obtain limited information about the contact network by individual interviews without contact tracing. Variables which can be estimated in this fashion are referred to as *node-level* measures [53]. One of the most well-studied of these is the degree distribution, which can be estimated by simply asking each person how many contacts they had in some interval of time [63–65].

An alternative approach has been the analysis of other networks, which can be estimated with phylogenetic methods from viral sequence data. Some work focuses on the *phylogenetic network*, in which two nodes are connected if the genetic distance between their viral sequences is below some threshold. Primarily, this work has focused on the detection of *phylogenetic clusters*, which are groups of individuals whose viral sequences are significantly more similar to each other's than to the general population's. The phylogenetic network is informative about "hotspots" of transmission and can be used to identify demographic groups to whom targeted interventions are likely to have the greatest effect [66]. However, this network may show little to no agreement with a contact data obtained through epidemiological methods [67–69], and therefore may be a poor proxy for the contact network. Other studies [70] have investigated the *transmission network*, which is the subgraph of the contact network consisting of infected nodes and the edges which led to their infections [12] (fig. 1.2, left). It is possible to estimate the transmission network phylogenetically, although the methods required for doing so are more sophisticated than for estimating the phylogenetic network [70]. These studies again mostly focusing on clustering, and also on degree distributions.

Other statistical methods have been developed to infer contact network parameters strictly from the timeline of an epidemic, using neither genetic data nor reported contacts. Britton and O'Neill [71] developed a Bayesian method to infer the p parameter of an Erdős-Rényi (ER) network, along with the transmission and removal rate parameters of the susceptible-infected (SI) model, using observed infection and optionally removal times. However, it was designed for only a small number of observations, and was unable to estimate p independently from the transmission rate. Groendyke, Welch, and Hunter [72] significantly updated and extended the methodology of Britton and O'Neill, and applied it to a measles outbreak affecting 188 individuals. They were able to obtain a much more informative estimate of p, although this data set included both symptom onset and recovery times for all individuals, and was unusual in that the entire contact network was presumed to be infected. Volz [58] developed differential equations describing the dynamics of the SIR model on a wide variety of random networks defined by their degree distributions. Although the topic of estimation was not addressed in the original paper, Volz's method could in principle be used to fit such models to observed epidemic trajectories, similar to what is done with the ordinary SIR model. Volz and Meyers [55] later extended the method to dynamic contact networks and applied it to a sexual network relating 99 individuals investigated during a syphilis outbreak.

1.3.2 Scale-free networks and preferential attachment

A *scale-free* network is one whose degree distribution follows a power law, meaning that the number of nodes in the network with degree k is proportional to $k^{-\gamma}$ for some constant γ [73]. Scale-free networks are characterized by a large number of nodes of low degree, with relatively few "hub" nodes of very high degree. Epidemiological surveys have indicated that human sexual networks tend to be scale-free [63–65]. Interestingly, many other types of network, including computer networks, biological neural networks, metabolic networks [74], and academic co-author networks, also have the scale-free property.

Several properties of scale-free networks are relevant in epidemiology. The high-degree hub nodes are known as *superspreaders* [75], which have been postulated to contribute in varying degree to the spread of diseases such as HIV [15] and severe acute respiratory syndrome (SARS) [76]. Scale-free networks have no epidemic threshold [60], meaning that diseases with arbitrarily low transmissibility can persist at low levels indefinitely. This is in contrast with homogeneously mixed populations, in which transmissibility below the epidemic threshold would result in exponential decay in the number of infected individuals and eventual extinction of the pathogen [56].

One mechanism which has been shown to lead to scale-free networks is *preferential attachment* [73, 77]. The simplest preferential attachment model is known as the Barabási-Albert (BA) model after its inventors [73]. Under this model, networks are formed by starting with a small number m_0 of nodes. New nodes are added one at a time until there are a total of N in the network. Each time a new node is added, $m \geq 1$ edges are added from it to other nodes in the graph. In the original formulation, the partners of the new node are chosen with probability linearly proportional to their degree. However, Barabási and Albert suggest extending the model such that the probability of choosing a partner of degree d is proportional to d^α for some constant α, and we use this extension here. When $m = 1$, the network takes on the distinctive shape of a tree, that is, it does not contain any cycles. Cycles are present in the network for all all other m values.

There has been some contention of the idea that contact networks are scale-free. Handcock and Jones [78] fit several stochastic models of partner formation to empirical degree distributions derived from population surveys of sexual behaviour. They found that a negative binomial distribution, rather than a power law, was the best fit to five out of six datasets, although the difference in goodness of fit was extremely small in four out of these five. Bansal, Grenfell, and Meyers [54] found that an exponential distribution, rather than a power law, was the best fit to degree distributions of six social and sexual networks.

1.3.3 Relationship between network structure and transmission trees

The contact network underlying an epidemic constrains the shape of the transmission network, which in turn determines the topology of the transmission tree relating the infected hosts (fig. 1.2). The index case who introduces the epidemic into the network becomes the root of the tree. Each time a transmission occurs, the lineage corresponding to the donor host in the tree splits into two, representing the recipient lineage and the continuation of the donor lineage. Figure 1.2 illustrates this correspondence. It must be emphasized that, although the order and timing of transmissions determines the tree topology uniquely, the converse does not hold. That is, for any given topology, there are in general many transmission networks which would lead to that topology. In other words, it impossible to distinguish who transmitted to whom from a transmission tree alone [79].

A number of studies have made progress in quantifying the relationship between contact networks and transmission trees. O'Dea and Wilke [80] simulated epidemics over networks with four types of degree distribution. They then estimated the Bayesian skyride [81] population size trajectory in two ways: from the phylogeny, using Markov chain Monte Carlo (MCMC); and from the incidence and prevalence trajectories, using the method developed by Volz et al. [82]. The concordance between the two skyrides, as well as the relationship between the skyride and prevalence curve, was qualitatively different for each degree distribution. Leventhal et al. [83] investigated the relationship between transmission tree imbalance and several epidemic parameters under four contact network models, and found that these relationships varied considerably depending on which model was being considered. Welch [84] simulated transmission trees over networks with varying degrees of community structure. They found that transmission trees simulated under networks with low clustering could not generally be distinguished from those simulated under highly clustered networks, and concluded that contact network clusters do not affect transmission tree shape. However, more recently, Villandre et al. [85] investigated the correspondence between contact network clusters and transmission tree clusters, and did find a moderate correspondence between the two in some cases.

1.4 Sequential Monte Carlo

1.4.1 Overview and notation

Sequential Monte Carlo (SMC) is the name for a family of statistical inference methods which rely on approximating probability distributions of interest with large collections of *particles*, here denoted $\left\{x^{(k)}\right\}$ [86, 87]. These collections or *populations* are constructed to form a *Monte Carlo approximation* to some distribution of interest π, meaning that the empirical distribution of the particles converges in distribution to π as the population size gets large [88]. The word *sequential* is used because the particle populations are modified in an iterative fashion over time, for example, to incorporate new evidence.

To fully describe SMC, we will introduce some notation and terminology. The definitions of these terms will become clearer as they are used. For a sequence $x_1, \dots, x_d$, we will write $\mathbf{x_i}$ to mean the partial sequence $x_1, \dots, x_i$. The subscript $^{(k)}$ will be used to indicate the kth particle in a population. To ease the notational burden we will omit the superscripts and subscripts on the weight functions w.

We define a *Markov kernel* as the continuous analogue of the transition matrix in a finite-state Markov model. For some spaces X and Y, $K : X \times Y \to [0, 1]$ such that

$$\int_Y K(x, y) \mathrm{d}y = 1 \tag{1.2}$$

for all $x \in X$. This is an "operational" definition of Markov kernel which will be suitable for our purposes. Note that a Markov kernels have nothing to do with the kernel functions defined in section 1.2.4, other than sharing a name (the word "kernel" is ubiquitous in mathematics).

1.4.2 Sequential importance sampling for sequential Monte Carlo

Sequential importance sampling (SIS) is one type of SMC method, whose aim is to sample from a distribution π on an high-dimensional space, say $\pi(\mathbf{x}) = \pi(x_1, \ldots, x_d)$. The basis of SIS is importance sampling (IS), which is a method of estimating summary statistics of distributions which are known only up to a normalizing constant, and therefore cannot be sampled from directly. That is, if π is such a distribution and f is any real-valued function, IS is concerned with estimating

$$\pi(f) = \int f(x)\pi(x)\mathrm{d}x = \int f(x)\frac{\gamma(x)}{Z}\mathrm{d}x,$$

where the integral is over the space on which π is defined, $\gamma(x)$ is known pointwise, and $Z = \int \gamma(x)\mathrm{d}x$ is the unknown normalizing constant. Suppose we have at hand another distribution η, called the *importance distribution*, from which we are able to sample. Define the *importance weight* as the ratio ratio $w(x) = \gamma(x)/\eta(x)$. We can express the normalizing constant Z in terms of the importance weight and distribution, $Z = \int w(x)\eta(x)\mathrm{d}x$, and in turn write the expectation of interest as

$$\int f(x)\pi(x)\mathrm{d}x = \frac{\int f(x)\gamma(x)\mathrm{d}x}{\int w(x)\eta(x)\mathrm{d}x}.$$

If we sample a large number of points from η, then $\eta(x)$ can be approximated by a Monte Carlo estimate. Since the remaining quantities f, γ, and w can all be evaluated pointwise, these are all the ingredients we need to obtain an estimate of $\pi(f)$. Although this is a simple and elegant approach, the drawback is that the variance of the estimate is proportional to the variance of the importance weights [87], which may be quite large if η and γ are very different. Therefore, the practical use of IS on its own is limited, since it depends on finding an importance distribution similar to π, which we usually know very little about *a priori*.

The objective of SIS is to build up an importance distribution η for π sequentially. By the general product rule, $\pi(\mathbf{x})$ can be decomposed as

$$\pi(\mathbf{x}) = \pi(x_1)\pi(x_2 \mid x_1)\cdots\pi(x_{d-1} \mid \mathbf{x_{d-2}})\pi(x_d \mid \mathbf{x_{d-1}}).$$

This decomposition is natural in many contexts, particularly for on-line estimation. For example, in a stateful model like an hidden Markov model (HMM), x_i may represent the state at time i, with $\pi(\mathbf{x})$ being the posterior distribution over possible paths. The importance distribution η for π will be constructed using a similar decomposition,

$$\eta(\mathbf{x}) = \eta(x_1)\eta(x_2 \mid x_1)\cdots\eta(x_{d-1} \mid \mathbf{x_{d-2}})\eta(x_d \mid \mathbf{x_{d-1}}).$$

The importance weights for η can be written recursively as

$$w(\mathbf{x_i}) = \frac{\pi(\mathbf{x_i})}{\eta(\mathbf{x_i})} = \frac{\pi(x_i \mid \mathbf{x_{i-1}})\pi(\mathbf{x_{i-1}})}{\eta(x_i \mid \mathbf{x_{i-1}})\eta(\mathbf{x_{i-1}})} = \frac{\pi(x_i \mid \mathbf{x_{i-1}})}{\eta(x_i \mid \mathbf{x_{i-1}})} \cdot w(\mathbf{x_{i-1}}). \tag{1.3}$$

Thus, we can choose $\eta(x_i \mid \mathbf{x_{i-1}})$ such that the variance of the importance weights is as small as possible at every step, eventually arriving at a full importance distribution. This choice is made on a problem-specific basis, taking any available information about $\pi(x_i \mid \mathbf{x_{i-1}})$ into account (see *e.g.* [87, 89] for many examples). One potential choice for $\eta(x_i \mid \mathbf{x_{i-1}})$ is simply $\pi(x_i \mid \mathbf{x_{i-1}})$, if it is possible to compute. In a Bayesian setting, the prior distribution may be used. The exact form of $\eta(x_i, \mathbf{x_{i-1}})$ which minimizes the variance of the weights is called the *optimal kernel* [90], the name deriving from the fact that $k(x_i, \mathbf{x_{i-1}}) = \eta(x_i, \mathbf{x_{i-1}})$ is a Markov kernel. In some applications, it is possible to approximate the optimal kernel or even compute it explicitly.

The recursive definition eq. (1.3) suggests an algorithm for obtaining a sample from π (algorithm 1). We begin with n "particles" which have been sampled from the importance distribution $\eta(x_0)$ for $\pi(x_0)$. The particles are updated and reweighted d times, corresponding to the d elements of the decomposition of π. At the ith step, each particle is extended to include x_i drawn according to the chosen $\eta(x_i \mid \mathbf{x_{i-1}})$, and the importance weights are recalculated and normalized.

Algorithm 1 Sequential importance sampling.

```
for $k = 1$ to $n$ do
    Sample $x_1^{(k)}$ from $\eta(x_1)$                                          ▷ Initialize the $k$th particle
    $w^{(k)} \leftarrow \frac{\pi\left(x_1^{(k)}\right)}{\eta\left(x_1^{(k)}\right)}$
end for
for $i = 2$ to $d$ do
    for $k = 1$ to $n$ do
        Sample $x_i^{(k)}$ from $\eta\left(x_i \mid \mathbf{x_{i-1}^{(k)}}\right)$          ▷ Extend the $k$th particle
        $w(\mathbf{x_i}^{(k)}) \leftarrow \frac{\pi\left(x_i^{(k)} \mid \mathbf{x_{i-1}^{(k)}}\right)}{\eta\left(x_i^{(k)} \mid \mathbf{x_{i-1}^{(k)}}\right)} \cdot w(\mathbf{x_{i-1}}^{(k)})$
    end for
    Normalize the weights so that $\sum w = 1$
end for
Sample $n$ particles with probabilities $w$
```

Of course, η is merely an approximation to π, and may be a fairly poor one depending on the application. Try as we might to keep the variances of the weights low, the cumulative errors at each sequential step tend to push many of the weights to very low values. This results in a poor approximation to π, since only a few particles retain high importance weights after all d sequential steps. To mitigate this problem, we periodically apply a resampling step when the variance in the importance weights becomes too high. Several different criteria have been proposed for when to resample, but we focus here on the one described by Liu [87], namely the decay of the expected sample size (ESS) below a prescribed threshold, conventionally $n/2$. The ESS of the population of particles is defined as

$$\mathrm{ESS}(w) = \frac{n}{1 + \mathrm{Var}(w)},$$

where n is the number of particles [87]. When the ESS drops below the threshold, we resample the particles according to their weights. This results in the removal of low-weight particles from the population, and also equalizes all the weights. Various resampling strategies beyond the basic sampling with replacement have been proposed, but we will not discuss those here.

1.4.3 The sequential Monte Carlo sampler

The SIS algorithm described above aims to sample from a high-dimensional distribution $\pi(x)$, by sequentially sampling from d distributions of lower but increasing dimension. Del Moral, Doucet, and Jasra [91] developed an *SMC sampler* with an alternative objective: to sample sequentially from d distributions $\pi_1, \ldots, \pi_d$, all of the same dimension and defined on the same space. The π_i are assumed to form a related sequence, such as posterior distributions attained by sequentially considering new evidence. As with SIS, we assume that $\pi_i(x) = \gamma_i(x)/Z_i$, where γ_i is known pointwise and the normalizing constant Z_i is unknown.

Both algorithms involve progression through a sequence of related distributions. For SIS, these distributions are lower-dimensional marginals of the target distribution, while for the SMC sampler, they are of the same dimension and constitute a smooth progression from an initial to a final distribution. In both cases, the neighbouring distributions in the sequence are related to each other in some way, and we can take advantage of that relationship to create a sequence of importance distributions alongside the sequence of targets. In SIS, the neighbouring marginals $\pi(\mathbf{x_i})$ and $\pi(\mathbf{x_{i+1}})$ were related by the conditional density $\pi(x_i \mid \mathbf{x_{i-1}})$, which we used to inform the importance distribution. In SMC, the relationship between subsequent distributions is less explicit, but it is assumed that they are related closely enough that an importance distribution for π_i can be easily transformed into one for π_{i+1}. In particular, the sequence of importance distributions η_i is constructed as

$$\eta_i(x') = \int \eta_{i-1}(x) K_i(x, x') \mathrm{d}\, x, \tag{1.4}$$

where K_i is a Markov kernel and the integral is over the space on which the π_i are defined. The choice of K_i should be based on the percieved relationship between π_{i-1} and π_i. Del Moral, Doucet, and Jasra [91] propose the use of a MCMC kernel with equilibrium distribution π_i. That is,

$$K_i(x, x') = \max\left(1, \frac{q(x', x)\pi_i(x)}{q(x, x')\pi_i(x')}\right),$$

where $q(\xi, x)$ is a proposal function such as a Gaussian distribution centered at ξ (see section 1.5.1).

Although this method of building up η appears straightforward, the drawback is that the importance distribution itself becomes intractible. In particular, evaluating $\eta_i(x)$ involves a i-dimensional integral of the type in eq. (1.4). As it is necessary to evaluate $\eta(x)$ pointwise to perform IS, this construction appears to have defeated the purpose of providing an importance distribution for each π_i. Del Moral, Doucet, and Jasra [91] overcome this problem with two "artificial" objects. First, they propose the existence of *backward* Markov kernels $L_{i-1}(x_i, x_{i-1})$. For now, these kernels are arbitrary, and will be precisely defined on a problem-specific basis. Second, they define an alternative sequence of target distributions

$$\tilde{\pi}_i(\mathbf{x_i}) = \pi_i(x_i) \prod_{k=1}^{i-1} L_k(x_{k+1}, x_k)$$

of increasing dimension. This brings us back to the setting described above in section 1.4.2, namely of building up an importance distribution of dimension d sequentially through lower-dimensional distributions. We can write $\tilde{\pi}_i$ in terms of $\tilde{\pi}_{i-1}$ by noticing that

$$\frac{\tilde{\pi}_i(\mathbf{x_i})}{\tilde{\pi}_{i-1}(\mathbf{x_{i-1}})} = \frac{\pi_i(x_i) \prod_{k=1}^{i-1} L(x_{k+1}, x_k)}{\pi_{i-1}(x_{i-1}) \prod_{k=1}^{i-2} L(x_{k+1}, x_k)} = \frac{\pi_i(x_i) L(x_i, x_{i-1})}{\pi_{i-1}(x_{i-1})},$$

and hence

$$\tilde{\pi}_i = \frac{\pi_i(x_i) L(x_i, x_{i-1})}{\pi_{i-1}(x_{i-1})} \cdot \tilde{\pi}_{i-1}.$$

Therefore, the importance weights for these new targets are defined recursively as

$$\begin{aligned} w(\mathbf{x_i}) &= \frac{\tilde{\pi}_i(\mathbf{x_i})}{\eta_i(\mathbf{x_i})} && (1.5) \\ &= \frac{\tilde{\pi}_{i-1}(\mathbf{x_{i-1}}) \pi_i(x_i) L(x_i, x_{i-1})}{\eta_{i-1}(\mathbf{x_{i-1}}) \pi_{i-1}(x_{i-1}) K_i(x_{i-1}, x_i)} && (1.6) \\ &= w(\mathbf{x_{i-1}}) \cdot \frac{\pi_i(x_i) L_{i-1}(x_i, x_{i-1})}{\pi_{i-1}(x_{i-1}) K_i(x_{i-1}, x_i)} && (1.7) \\ &\propto w(\mathbf{x_{i-1}}) \cdot \frac{\gamma_i(x_i) L_{i-1}(x_i, x_{i-1})}{\gamma_{i-1}(x_{i-1}) K_i(x_{i-1}, x_i)}. && (1.8) \end{aligned}$$

The final key piece of information is to notice that, because the L_i are Markov kernels, π_i is simply the marginal in $\mathbf{x_{i-1}}$ of $\tilde{\pi}$. Therefore, a sample from $\tilde{\pi}_i$ automatically gets us a sample from π_i, by considering only the ith component of $\mathbf{x_i}$. These are all the ingredients we need to apply SIS. The sequences of kernels L and K should be chosen based on the problem at hand to inimize the variance in the importance weights as well as possible. For a fixed choice of K_i, the backward kernels L_i which minimize this variance are called the *optimal* backward kernels. The full SMC sampler algorithm is presented as algorithm 2. A resampling step is applied whenever the ESS of the population drops too low, as discussed in the previous section.

1.5 Approximate Bayesian computation

1.5.1 Model fitting

A *mathematical model* is a formal description of a hypothesized relationship between some observed data, x and outcomes y. A *parametric* model defines a family of possible relationships between data and outcomes, indexed by one or more numeric parameters θ. A *statistical* model describes the re-

Algorithm 2 Sequential Monte Carlo sampler of Del Moral, Doucet, and Jasra [91].

```
for $k = 1$ to $n$ do
    Sample $x_1^{(k)}$ from $\eta_1(x_1)$                                ▷ Initialize the $k$th particle
    $w^{(k)} \leftarrow \frac{\gamma_1\left(x_1^{(k)}\right)}{\eta_1\left(x_1^{(k)}\right)}$
    Normalize the weights so that $\sum w = 1$
end for
for $i = 2$ to $d$ do
    for $k = 1$ to $n$ do
        Sample $x_i^{(k)}$ from $K(x_{i-1}^{(k)}, x_i)$                   ▷ Extend the $k$th particle
        $w^{(k)} \leftarrow w^{(k)} \cdot \frac{\gamma_i(x_i)L_{i-1}(x_i, x_{i-1})}{\gamma_{i-1}(x_{i-1})K_i(x_{i-1}, x_i)}$
    end for
    Normalize the weights so that $\sum w = 1$
    if ESS$(w) < T$ then
        Resample the particles according to $w$
        for $k = 1$ to $n$ do
            $w^{(k)} \leftarrow 1/n$
        end for
    end if
    Sample the $i$th component of $n$ particles with probabilities $w$
end for
```

lationship between data and outcomes in terms of probabilities. Statistical models define, either explicitly or implicitly, the probability of observing y given $\mathbf{x}$ and, if the model is parametric, θ. Note that it is entirely possible to have no data $\mathbf{x}$, only observed outcomes y. In this case, a model would describe the process by which y is generated.

To illustrate these concepts, consider the well-known linear model. For clarity, we will restrict our attention to the case of one-dimensional data and outcomes where $x = \{x_1, \ldots, x_n\}$ and $y = \{y_1, \ldots, y_n\}$ are vectors of real numbers. The linear model postulates that the outcomes are linearly related to the data, modulo some noise introduced by measurement error, environmental fluctuations, and other external factors. Formally, $y_i = \beta x_i + \varepsilon_i$, where β is the slope of the linear relationship, and ε_i is the error associated with measurement i. We can make this model a statistical one by hypothesizing a distribution for the error terms ε_i; most commonly, it is assumed that they are normally distributed with variance σ. In mathematical terms, $Y_i \sim \beta x_i + \mathcal{N}(0, \sigma^2)$, where "$\sim$" means "is distributed as". We can see from this formulation that the model is parametric, with parameters $\theta = (\beta, \sigma)$. Moreover, we can write down the probability density π of observing outcome y_i given the parameters,

$$\pi(y \mid \beta, \sigma) = \prod_{i=1}^{n} f_{\mathcal{N}(0,\sigma^2)}(y_i - \beta x_i),$$

where $f_{\mathcal{N}(0,\sigma^2)}$ is the probability density of the normal distribution with mean zero and variance σ^2. Note that we are treating the x_i as fixed quantities, and therefore have not conditioned the probability density on $\mathbf{x}$. Also, we have assumed that all the y_i are independent.

TODO

For a general model, the probability density of y given the parameters θ is also known as the *likelihood*, written $\mathcal{L}$, of θ. That is, $\mathcal{L}(\theta \mid y) = f(y \mid \theta)$ for the model's probability density function (pdf) f. The higher the value of the likelihood, the more likely the observations y are under the model. Thus, the likelihood provides a natural criterion for fitting the model parameters: we want to pick θ such that the probability density of our observed outcomes y is as high as possible. The parameters which optimize the likelihood are known as the *ML* estimates, denoted $\hat{\theta}$. That is,

$$\hat{\theta} = \arg\max_{\theta} \ \mathcal{L}(\theta \mid y).$$

ML estimation is usually performed with numerical optimization. In the simplest terms, many possible values for θ are examined, $\mathcal{L}(\theta \mid y)$ is calculated for each, and the parameters which produce the highest value are accepted. Many sophisticated numerical optimization methods exist, although they may not be guaranteed to find the true ML estimates if the likelihood function is complex.

ML estimation makes use only of the data and outcomes to estimate the model parameters θ. However, it is frequently the case that the investigator has some additional information or belief about what θ are likely to be. For example, in the linear regression case, the instrument used to measure the outcomes may have a well-known margin of error, or the sign of the slope may be obvious from previous experiments. The Bayesian approach to model fitting makes use of this information by codifying the investigator's beliefs as a *prior distribution* on the parameters, denoted $\pi(\theta)$. Instead of considering only the likelihood, Bayesian inference focuses on the product of the likelihood and the prior, $f(y \mid \theta)\pi(\theta)$. Bayes' theorem tells us that this product is related to the *posterior distribution* on θ,

$$f(\theta \mid y) = \frac{f(y \mid \theta)\pi(\theta)}{\int f(y \mid \theta)\pi(\theta) \mathrm{d}\theta}. \tag{1.9}$$

In principle, $f(y \mid \theta)\pi(\theta)$ can be optimized numerically just like $\mathcal{L}(\theta \mid y)$, which would also optimize the posterior distribution. The resulting optimal parameters are called the maximum *a posteriori* (MAP) estimates. However, from a Bayesian perspective, θ is not a fixed quantity to be estimated, but rather a random variable with an associated distribution (the posterior). Therefore, the MAP estimate by itself is of limited value without associated statistics about the posterior distribution, such as the mean or credible intervals. Unfortunately, to calculate such statistics, it is necessary to evaluate the normalizing constant in the denominator of eq. (1.9), which is almost always an intractable integral.

A popular method for circumventing the normalizing constant is the use of MCMC to obtain a sample from the posterior distribution. MCMC works by defining a Markov chain whose states are indexed by possible model parameters. The transition probability from state θ_1 to state θ_2 is taken to be

$$\max\left(1, \frac{f(y \mid \theta_2)\pi(\theta_2)q(\theta_2, \theta_1)}{f(y \mid \theta_1)\pi(\theta_2)q(\theta_1, \theta_2)}\right),$$

where $q(\theta, \theta')$ is a symmetric *proposal distribution* used in the algorithm to generate the chain. The stationary distribution of this Markov chain is equal to the posterior distribution on θ. Therefore, if a long enough random walk is performed on the chain, the distribution of states visited will be a Monte Carlo approximation of $f(\theta \mid y)$, from which we can calculate statistics of interest. Actually performing this random walk is straightforward and can be accomplished via the Metropolis-Hastings algorithm (algorithm 3).

Algorithm 3 Metropolis-Hastings algorithm for Markov chain Monte Carlo.

```
Draw $\theta$ according to the prior $\pi(\theta)$
loop
    Propose $\theta'$ according to $q(\theta, \theta')$
    Accept $\theta \leftarrow \theta'$ with probability $\max\left(1, \frac{f(y \mid \theta')\pi(\theta')q(\theta', \theta)}{f(y \mid \theta\ )\pi(\theta\ )q(\theta, \theta')}\right)$
end loop
```

1.5.2 Overview of ABC

Most mathematical models are amenable to fitting via one or both of the approaches, ML or Bayesian inference, discussed above. However, there are some, particularly in the domain of population genetics [92, 93], for which calculation of either the likelihood or the product of the likelihood and the prior may be infeasible. For example, one or both of these quantities may be expressible only as an intractable integral. Approximate Bayesian computation (ABC) is designed for such cases, where standard likelihood-based techniques for model fitting cannot be applied.

Ordinarily, Bayesian inference targets the posterior distribution $f(\theta \mid y)$. That is, in the Bayesian framework, model parameters with higher posterior density are "better" in the sense that they offer a more credible explanation for the observed data. Approximate Bayesian computation offers an alternative metric for parameter credibility, namely the similarity of simulated datasets to the observed data. If datasets simulated under the model closely resemble the real data, it follows that the model is a reasonable approximation to the real-world process generating the observed data. More formally, suppose we have a distance measure ρ defined on the space of all possible data our model could generate. ABC aims to sample from the joint posterior distribution of model parameters and simulated datasets z which are within some small distance ε of the observed data y,

$$\pi_\varepsilon(\theta, z \mid y) = \frac{\pi(\theta) f(z \mid \theta) \mathbb{I}_{A_{\varepsilon,y}}(z)}{\int_{A_{\varepsilon,y} \times \Theta} \pi(\theta) f(z \mid \theta) \mathrm{d}\theta}.$$

Here, $A_{\varepsilon,y}$ is an ε-ball around y with respect to ρ, Θ is the space of all possible model parameters, and $\mathbb{I}$ is the indicator function [94]. As we shall see in the next section, this distribution can be sampled from exactly. The word "approximate" derives from the assumption that, for a suitably chosen distance ρ and a small enough ε, the marginal in z of this distribution approximates the posterior of interest [94]. That is,

$$\int \pi_\varepsilon(\theta, z \mid y)\mathrm{d}z \approx f(\theta \mid y).$$

This distribution is variously referred to as the *ABC target distribution* or the ABC approximation to the posterior. Note that in many formulations, the distance function ρ is defined as $\rho(S(\cdot), S(\cdot))$ where S is a function which maps data points into a vector of summary statistics. This can be useful if the data are high-dimensional or of a complex type, but it is not strictly necessary. For instance, if the data are numeric and of low dimension, the distance function may simply be the Euclidian distance [95]. For more complex data, Nakagome, Fukumizu, and Mano [96] proposed the use of a kernel function (defined in section 1.2.4), an approach they dubbed *kernel-ABC*.

1.5.3 Algorithms for ABC

Algorithms for performing ABC fall into one of three categories: rejection, MCMC, and SMC. To simplify the math, we shall restrict the descriptions of these algorithms to the case of one simulated dataset per parameter particle (the meaning of this will become clear shortly). The extension to multiple datasets per particle is straightforward and will be given at the end of the section. We use the variable x to refer to the pair (θ, z), so that the ABC target distribution can be written $\pi_\varepsilon(x \mid y)$.

Rejection ABC is the simplest method, and also the one which was first proposed [97, 98]. The algorithm, outlined in algorithm 4, repeats the following steps until a desired number of samples from the target distribution are obtained. Parameter values θ are sampled according to the prior distribution $\pi(\theta)$. Then, a simulated dataset z is generated from the model with the sampled parameter values. By definition, the probability density of obtaining the particular dataset z is $f(z \mid \theta)$. Finally, the parameters are sampled if the distance of z from the observed data y is less than ε, that is, with probability $\mathbb{I}_{A_{\varepsilon,y}}(z)$. Putting this all together, the parameters θ are sampled with probability proportional to

$$\pi(\theta) f(z \mid \theta) \mathbb{I}_{A_{\varepsilon,y}}(z),$$

which is exactly the numerator of the ABC target distribution. Thus, θ represents an unbiased sample from the approximate posterior.

Algorithm 4 Rejection ABC.

```
loop
    Draw $\theta$ according to $\pi(\theta)$
    Simulate a dataset $z$ from the model with parameters $\theta$
    if $\rho(y, z) < \varepsilon$ then
        Sample $\theta$
    end if
end loop
```

Rejection ABC is easy to understand and implement, but it is not generally computationally feasible. If the posterior is very different from the prior, a very large number of samples may need to be taken in order to find a simulated dataset which is close to z. The inefficiency is compounded by the curse of dimensionality - the measure of the ε-ball around y decreases exponentially with the number of dimensions. ABC-MCMC (algorithm 5) was designed to overcome these hurdles [99]. The approach is similar to ordinary Bayesian MCMC (section 1.5.1), except that a distance cutoff replaces the likelihood ratio. That is, the transition probability between states x and x' is defined as

$$\max\left(1, \frac{f(z' \mid \theta')q(\theta', \theta)}{f(z \mid \theta)q(\theta, \theta')} \cdot \mathbb{I}_{A_{\varepsilon,y}}(z')\right).$$

Algorithm 5 ABC-MCMC.

```
Draw $\theta$ according to $\pi(\theta)$
loop
    Propose $\theta'$ according to $q(\theta, \theta')$
    Simulate a dataset $\mathbf{z}'$ according to the model with parameters $\theta$
    Accept $\theta \leftarrow \theta'$ with probability $\max\left(1, \frac{\pi(\theta')q(\theta', \theta)}{\pi(\theta\ )q(\theta, \theta')} \cdot \mathbb{I}_{A_{\varepsilon,y}}(z')\right)$
end loop
```

Some of the same computational inefficiencies arise with ABC-MCMC as with rejection. For example, in regions of low posterior density, the probability to simulate a dataset proximal to the observed data is low. Various strategies have been developed to mitigate this, including reducing the tolerance level ε as the chain progresses [100].

The most recently developed class of algorithm for ABC is ABC-SMC [95, 101]. As with ABC-MCMC, the algorithm is a straightforward modification of an existing Bayesian inference method, in this case the SMC sampler (section 1.4.3). The sequence of target distributions is defined as $\pi_i = \pi_{\varepsilon_i}(x \mid y)$ for a decreasing sequence of tolerances ε_i. The intention is for the algorithm to progress smoothly through a sequence of target distributions which ends at the ABC approximation to the posterior. As discussed in section 1.4.3, the choices of the kernels K and L is problem-specific, and so appropriate kernels must be chosen for ABC. Several options have been proposed [5, 95, 101].

All the algorithms discussed in this section can be straightforwardly extended to sample from the joint distribution

$$\pi_\varepsilon(\theta, z_1, \ldots, z_M \mid y),$$

which is equivalent to associating M simulated datasets to each parameter particle instead of just one. The simulated dataset z is replaced by $z = z_1, \ldots, z_M$, and the indicator function for the ε-ball around y is replaced by

$$\sum_{k=1}^{M} \mathbb{I}_{A_{\varepsilon,y}}(z_i).$$

For ABC-MCMC and ABC-SMC, the proposal distribution $q(\theta, \theta')f(z \mid \theta')$ is replaced by

$$q_i(\theta, \theta') \prod_{k=1}^{M} f(z_i \mid \theta').$$

Chapter 2

Reconstructing contact network parameters from viral phylogenies

2.1 Methods

2.1.1 Kernel-ABC method

Netabc is a computer program to perform statistical inference of contact network parameters from an estimated transmission tree using kernel-ABC. The program combines three major components: Gillespie simulation, to simulate transmission trees on contact networks; the phylogenetic kernel, to compare simulated to observed transmission trees; and adaptive ABC-SMC, to maintain a population of particles and advance it toward the ABC target distribution. We give a high-level overview of the program here, before describing these three components in detail.

As described in section 1.4, *netabc* keeps track of a population of particles $x^{(k)}$, each of which contains particular parameter values $\theta^{(k)}$ for the model we are trying to fit. A small number of contact networks $z^{(k)}$ are generated for each particle, in accordance with that particle's parameters. An epidemic is simulated over each of these networks using Gillespie simulation, and by keeping track of its progress, a transmission tree is obtained. Thus, each particle becomes associated with several simulated transmission trees. These trees are compared to the observed tree using the phylogenetic kernel. Particles are weighted according to the similarity of their associated simulated trees with the true tree, with more similar trees receiving higher weights. The particles are iteratively perturbed to explore the parameter space, and particles with simulated trees too distant from the true tree are periodically dropped and resampled. Once a convergence criterion is attained, the final set of particles is used as a Monte Carlo approximation to the target distribution of ABC, which is assumed to resemble the posterior distribution on model parameters (see section 1.5). A graphical schematic of this algorithm is given in fig. 2.1.

Netabc is written in the *C* programming language. The *igraph* library [102] is used to generate and store contact networks and phylogenies. Judy arrays [103] are used for hash tables and dynamic programming matrices. The GNU scientific library (GSL) [104] is used to generate random

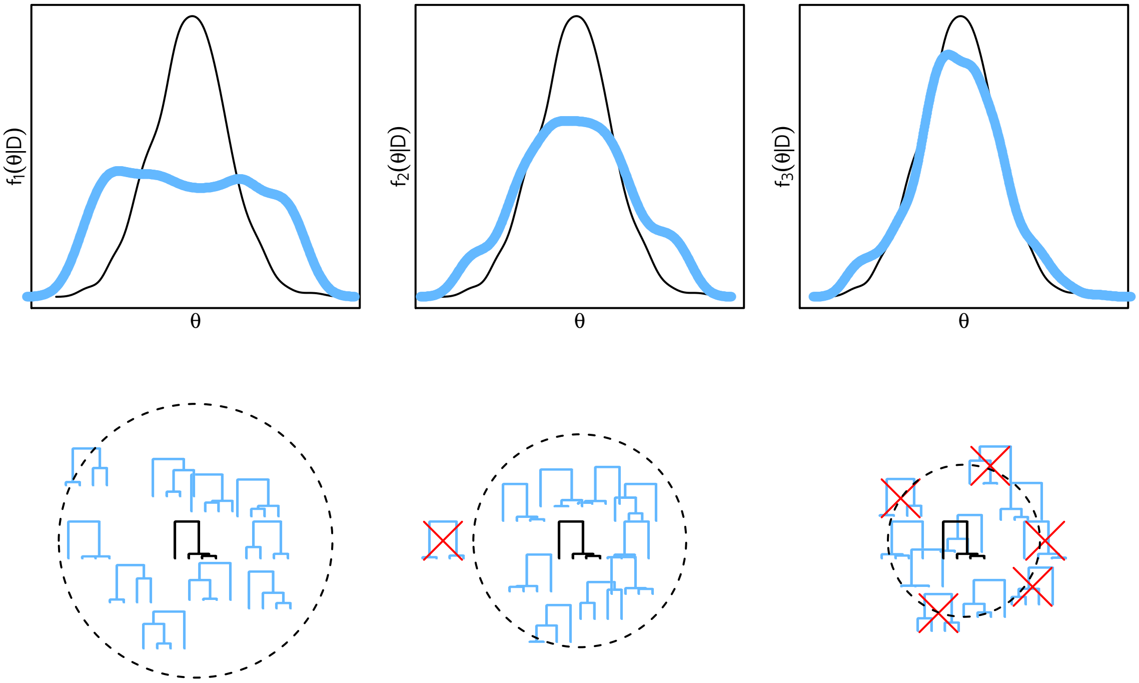

Figure 2.1: Graphical schematic of the ABC-SMC algorithm implemented in *netabc*. Particles are initially drawn from their prior distributions, making the initial population a Monte Carlo approximation to the prior. At each iteration, particles are perturbed, and a distance threshold around the true tree contracts. Particles are rejected, and eventually resampled, when all their associated simulated trees lie outside the threshold. As the algorithm progresses, the population smoothly approaches a Monte Carlo approximation of the ABC target distribution, which is assumed to resemble the posterior.

draws from probability distributions, and to perform the bisection step in the adaptive ABC-SMC algorithm. Parallelization is implemented with POSIX threads [105]. In addition to the *netabc* binary to perform kernel-ABC, we provide three additional stand-alone utilities: *treekernel*, to calculate the phylogenetic kernel; *nettree*, to simulate a transmission tree over a contact network; and *treestat*, to compute various summary statistics of phylogenies. The programs are freely available at https://github.com/rmcclosk/netabc.

Epidemic simulation

The simulation of epidemics, and the corresponding transmission, trees over contact networks is performed in *netabc* using the Gillespie simulation algorithm [106]. This method has been independently implemented and applied by several authors [*e.g.* 69, 72, 80, 83, 85]. Groendyke, Welch, and Hunter [72] published their implementation as an *R* package, but since the SMC algorithm is quite computationally intensive, we chose to implement our own version in *C*.

Let $G = (V, E)$ be a directed contact network. We assume the individual nodes and edges of G follow the dynamics of the SIR model [2]. Each directed edge $e = (u, v)$ in the network is associated with a transmission rate β_e, which indicates that, once u becomes infected, the waiting time until u infects v is distributed as Exponential(β_e). Note that v may become infected before this time has elapsed, if v has other incoming edges. v also has a removal rate γ_v, so that the waiting time until removal of v from the population is Exponential(γ_v). Removal may correspond to death or recovery with immunity, or a combination of both, but in our implementation recovered nodes never re-enter the susceptible population. We define a *discordant edge* as an edge (u, v) where u is infected and v has never been infected.

To describe the algorithm, we introduce some notation and variables. Let in(v) be the set of incoming edges to v, and out(v) be the set of outgoing edges from v. Let I be the set of infected nodes in the network, R be the set of removed nodes, and S be the remaining susceptible nodes, and D be the set of discordant edges in the network. Let β be the total transmission rate over all discordant edges, and γ be the total removal rate of all infected nodes,

$$\beta = \sum_{e \in D} \beta_e, \quad \gamma = \sum_{v \in I} \gamma_v.$$

The variables S, I, R, D, β, and γ are all updated as the simulation progresses. When a node v becomes infected, it is deleted from S and added to I. Any formerly discordant edges in $\in (v)$ are deleted from D, and edges in out(v) to nodes in S are added to D. If v is later removed, it is deleted from I and added to R, and any discordant edges in out(v) are deleted from D. At the time of either infection or removal, the variables β and γ are updated to reflect the changes in the network. Since these updates are straightforward, we do not write them explicitly in the algorithm.

The Gillespie simulation algorithm is given as Algorithm 2.1.1. The transmission tree T is simulated along with the epidemic. We keep a map called *tip*, which maps infected nodes in I to the tips of T. The simulation continues until either there are no discordant edges left in the network, or we reach a user-defined cutoff of time ($t_{\max}$) or number of infections ($I_{\max}$). We use the notation Uniform$(0, 1)$ to indicate a number drawn from a uniform distribution on $(0, 1)$, and likewise for Exponential(λ). The combined number of internal nodes and tips in T is denoted $|T|$.

Algorithm 6 Simulation of an epidemic and transmission tree over a contact network

```
infect a node $v$ at random, updating $S$, $I$, $D$, $\beta$ and $\gamma$
$T \leftarrow$ a single node with label 1
$tip[v] \leftarrow 1$
$t \leftarrow 0$
while $D \neq \emptyset$ and $|I| + |R| < I_{\max}$ and $t < t_{\max}$ do
    $s \leftarrow \min(t_{\max} - t, \text{Exponential}(\beta + \gamma))$
    for $v \in tip$ do
        extend the branch length of $tip[v]$ by $s$
    end for
    $t \leftarrow t + s$
    if $t < t_{\max}$ then
        if Uniform$(0, \beta + \gamma) < \beta$ then
            choose an edge $e = (u, v)$ from $D$ with probability $\beta_e/\beta$ and infect $v$
            add tips with labels $(|T| + 1)$ and $(|T| + 2)$ to $T$
            connect the new nodes to $tip[v]$ in $T$, with branch lengths $0$
            $tip[v] \leftarrow |T| - 1$
            $tip[u] \leftarrow |T|$
        else
            choose a node $v$ from $I$ with probability $\gamma_v/\gamma$ and remove $v$
            delete $v$ from $tip$
        end if
        update $S$, $I$, $R$, $D$, $\beta$, and $\gamma$
    end if
end while
```

Phylogenetic kernel

The tree kernel developed by Poon et al. [46] provides a comprehensive similarity score between two phylogenetic trees, via the dot-product of the two trees' feature vectors in the infinite-dimensional space of all possible subset trees with branch lengths (see section 1.2.4). The kernel was implemented using the fast algorithm developed by Moschitti [107]. First, the production rule of each node, which is the total number of children and the number of leaf children, is recorded. The nodes of both trees are ordered by production rule, and a list of pairs of nodes sharing the same production rule is created. These are the nodes for which the value of the tree kernel must be computed - all other pairs have a value of zero. The pairs to be compared are then re-ordered so that the child nodes are always evaluated before their parents. Due to its recursive definition, ordering the pairs in this way allows the tree kernel to be computed by dynamic programming. The complexity of this implementation is $O(|T_1||T_2|)$, where $|T|$ counts the number of nodes in the tree T.

The tree kernel cannot be used directly as a distance measure for ABC, since it is maximized, not minimized, when the two trees being compared are the same. Therefore, we defined the distance between two trees as

$$\rho(T_1, T_2) = 1 - \frac{K(T_1, T_2)}{\sqrt{K(T_1, T_1)K(T_2, T_2)}},$$

which is a number between 0 and 1 minimized when $T_1 = T_2$. This is similar to the normalization used by Poon et al. [46] and Collins and Duffy [48].

Adaptive sequential Monte Carlo for Approximate Bayesian computation

We implemented the adaptive SMC algorithm for ABC developed by Del Moral, Doucet, and Jasra [5]. This algorithm is similar to the reference ABC-SMC algorithm described in section 1.5.3, except that the sequence of tolerances ε_i is automatically determined rather than specified in advance. The tolerances are chosen such that the ESS of the particle population, which indicates the quality of the Monte Carlo approximation (see section 1.4.2), decays at a controlled rate. This prevents the Monte Carlo approxmiation from collapsing by the variance in weights growing too large. A single parameter α (not to be confused with the BA model parameter) controls the decay rate, with ε_i being chosen to satisfy

$$\text{ESS}(w_i) = \alpha\,\text{ESS}(w_{i-1}).$$

Here, w_i is the vector of weights at the ith step. Note that, since w_i depends on ε_i, this equation solves for the updated weights and the updated tolerance simultaneously. As pointed out by Del Moral, Doucet, and Jasra [5], the equation has no analytic solution, but can be solved numerically by bisection. The forward kernels K_i are taken to be MCMC kernels with stationary distributions π_{ε_i} and proposal distributions

$$q_i(\theta, \theta') \prod_{k=1}^{M} \Pr(z_i^{(k)'} \mid \theta'),$$

where θ is the vector of model parameters and z_k are M datasets simulated according to θ'. In our implementation, q is either a Gaussian proposal for continuous parameters, or a Poisson proposal for discrete parameters. For the Poisson proposals, the number of steps to move the particle is drawn from a Poisson distribution, and the direction in which to move the particle is chosen uniformly at random. For both proposals, the variance was set equal to twice the empirical variance of the particles, following [5, 101]. The backwards kernels are

$$L_{i-1}(x', x) = \frac{\pi_n(x)K(x, x')}{\pi_n(x')}.$$

When substituted into eq. (1.8), the forward kernels $K(x, x')$ and densities $\pi_n(x') = \pi_{\varepsilon_n}(x')$ cancel out, and we are left with the weight update

$$\begin{aligned} w_i(x) &\propto w_{i-1}(x) \frac{\pi_n(x \mid y)}{\pi_{i-1}(x \mid y)} \\ &= w_{i-1}(x) \frac{\pi(x)\pi_i(y \mid x)}{\pi(x)\pi_{i-1}(y \mid x)} \\ &= w_{i-1}(x) \frac{\sum_{k=i}^{M} \mathbb{I}_{A_{\varepsilon_i, y}}(z_k)}{\sum_{k=i}^{M} \mathbb{I}_{A_{\varepsilon_{i-1}, y}}(z_k)}. \end{aligned}$$

In other words, when the distance threshold ε_{i-1} is contracted to ε_i, the particles' weights are multiplied by the proportion of simulated datasets which are still inside the new threshold. The algorithm may be stopped when one of two termination conditions is reached. The user may specify a final tolerance ε, or a final acceptance rate of the MCMC kernel. The latter condition stops the algorithm when the particles are not moving around very much, implying little change in the estimated target.

2.1.2 Analysis of Barabási-Albert model

We investigated four parameters related to the BA model, denoted N, m, α, I. The first three of these are parameters of the model itself, while I is related to the simulation of transmission trees over the network. However, we will refer to all four as BA parameters. N denotes the total number of nodes in the network, or equivalently, susceptible individuals in the population. When a node is added to the network, m new undirected edges are added incident to it, and are attached to existing nodes of degree k with probability proportional to k^α (section 1.3.2). To simulate transmission trees over a BA network, we allowed an epidemic to spread until I nodes were infected, and sampled a transmission tree at that time. We assumed that all contacts had symmetric transmission risk, which was implemented by replacing each undirected edge in the network with two directed edges (one in each direction).

We did not consider the time scale of the transmission trees in these simulations, only their shape. Therefore, the transmission rate along each edge in the network was set to 1, and all transmission trees' branch lengths were scaled by their mean. The removal rate of each node was set to 0, implying no recovery or death in the population. These assumptions are similar to those made by Leventhal et al. [83].

Kernel classifiers

The experiments presented here involved a large number of variables which were varied combinatorially. For ease of exposition, we will describe a single experiment first, then enumerate the values of all variables for which the experiment was repeated. The parameters of the tree kernel, λ and σ (section 1.2.4) will be referred to as *meta-parameters* to distinguish them from the parameters of the BA model. With the exception of our own programs, all analyses were done in *R*, and all packages listed below are *R* packages.

The attachment power parameter α was varied among three values: 0.5, 1.0, and 1.5. For each value, the *sample_pa* function in the *igraph* package was used to simulate 100 networks, with the other parameters set to $N = 5000$ and $m = 2$. This step yielded a total of 300 networks. An epidemic was simulated on each network using our *nettree* binary until $I = 1000$ nodes were infected, at which point 500 of them were sampled to form a transmission tree. A total of 300 transmission trees were thus obtained, comprised of 100 trees for each of the three values of α. The trees were "ladderized" so that the subtree descending from the left child of each node was not smaller than that descending from the right child. Summary statistics, such as Sackin's index and the ratio of internal to terminal branch lengths, were computed for each simulated tree using our *treestat* binary. The trees were visualized using the *ape* package [108]. Our *treekernel* binary was used to calculate the value of the kernel for each pair of trees, with the meta-parameters set to $\lambda = 0.3$ and $\sigma = 4$. These values were stored in a symmetric 300 × 300 kernel matrix. Similarly, we computed the nLTT statistic between each pair of trees using our *treestat* binary, and stored them in a second 300×300 matrix.

To investigate the effect of α on tree shape, we constructed classifiers for α based on three statistics. First, we used the *kernlab* package [109] to create a kSVR classifier using the computed kernel matrix. Second, we used the *e1071* package [110] to create an ordinary SVR classifier using the pairwise nLTT matrix. Finally, we performed an ordinary linear regression of α against Sackin's index. Each of these classifiers was evaluated with 1000 two-fold cross-validations. We also performed a kPCA projection of the kernel matrix, and used it to visualize the separation of the different α values in the tree kernel's feature space. A schematic of this experiment is presented in fig. 2.2.

Similar experiments were performed with the values shown in table 2.1. The other three BA parameters, namely N, m, and I, were each varied while holding the others fixed. The experiments for α, m, and N were repeated with three different values of I. All experiments were repeated with trees having three different numbers of tips. Kernel matrices were computed for all pairs of the meta-parameters $\lambda = \{0.2, 0.3, 0.4\}$ and $\sigma = \{1/8, 1/4, 1/2, 1, 2, 4, 8\}$.

varied parameter	N	α	m	I	tips	λ	σ
N	3000, 5000, 8000	1.0	2	500, 1000, 2000	100, 500, 1000	0.2, 0.3, 0.4	1/8, 1/4, 1/2, 1, 2, 4, 8
α	5000	0.5, 1.0, 1.5	2	500, 1000, 2000	100, 500, 1000	0.2, 0.3, 0.4	1/8, 1/4, 1/2, 1, 2, 4, 8
m	5000	1.0	2, 3, 4	500, 1000, 2000	100, 500, 1000	0.2, 0.3, 0.4	1/8, 1/4, 1/2, 1, 2, 4, 8
I	5000	1.0	2	500, 1000, 2000	100, 500	0.2, 0.3, 0.4	1/8, 1/4, 1/2, 1, 2, 4, 8

Table 2.1: Values of parameters and other variables used in tree kernel simulation experiments. Each row corresponds to one of the BA model parameters. One kernel matrix was created for every combination of values except the one indicated in the "varied parameter" column, which was varied when producing simulated trees.

parameter	grid values	test values	N	α	m	I	tips
N	1050, 1125, . . . , 15000	1000, 3000, . . . , 15000	-	1.0	2	1000	100, 500, 1000
α	0, 0.01, . . . , 2	0, 0.25, . . . 2	5000	-	2	1000	100, 500, 1000
m	1, 2, . . . , 6	1, 2, . . . 6	5000	1.0	-	1000	100, 500, 1000
I	500, 525, . . . , 5000	500, 100, 1500, 2000	5000	1.0	2	-	100, 500

Table 2.2: Variables and BA parameter values used for grid search experiments. Trees were simulated under the test values, and compared to a grid of trees simulated under the grid values. Kernel scores were used to calculate point estimates and credible intervals for the test values.

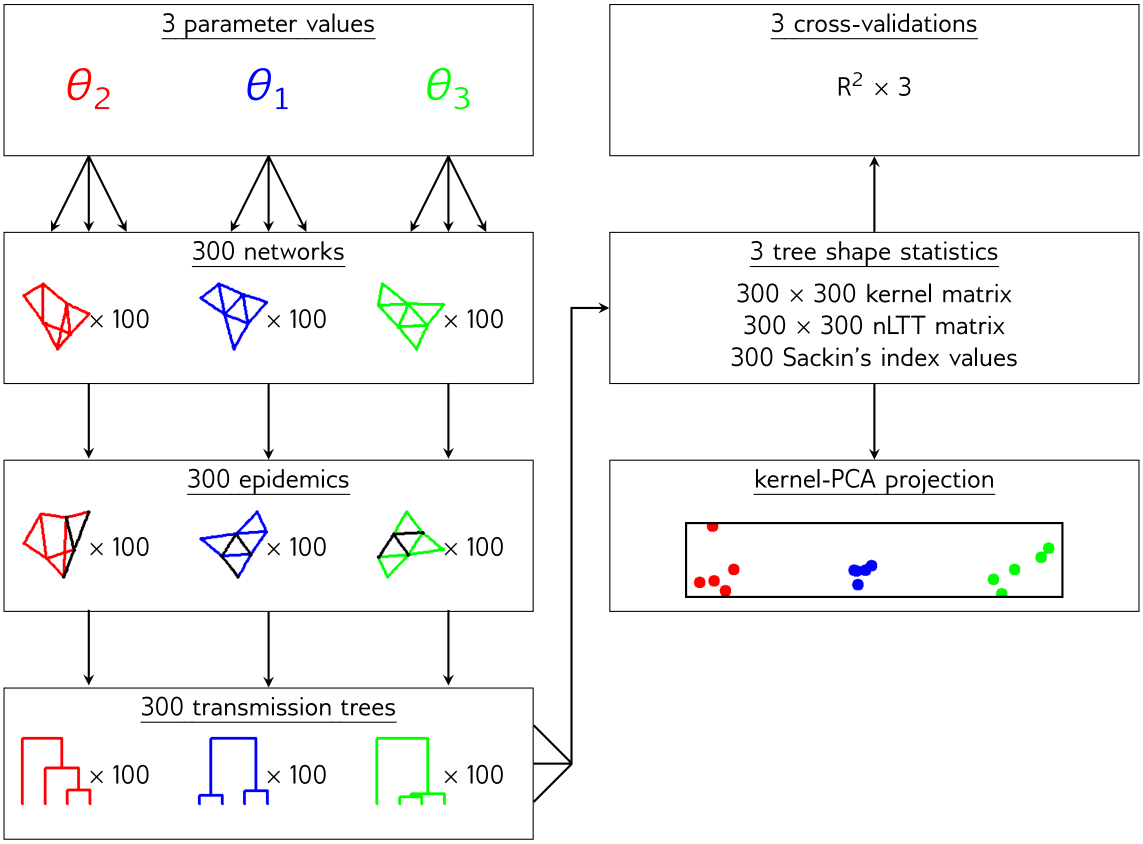

Figure 2.2: Schematic of investigation of BA model parameters using the tree kernel.

Grid search

As in the previous section, we will begin by describing a single experiment, and then list the variables for which similar experiments were performed. We varied α along a narrowly spaced grid of values: 0, 0.01, ..., 2. For each value, fifteen networks were generated with *igraph*, and transmission trees were simulated over each using *nettree*. These trees will be referred to as "grid trees", and their associated values "grid values". Next, one further test tree was simulated with the test value $\alpha = 0$. Both the grid trees and the test tree had 500 tips, and were simulated with the other BA parameters set to $N = 5000$, $m = 2$, and $I = 1000$. The test tree was compared to each of the grid trees using the tree kernel, with the meta-parameters set to $\lambda = 0.3$ and $\sigma = 4$, using the *treekernel* binary. The median kernel score was calculated for each grid value, and the scores were normalized such that the area under the curve was equal to 1. The grid value with the highest median kernel score was taken as the point estimate for the test value, and a 95% credible interval was obtained using the *hpd* function in the *TeachingDemos* package.

Each experiment of the type just described was repeated ten times with the same test value. Similar experiments were performed for each of the four BA parameters, with several test values and trees of varying sizes. The variables are listed in table 2.2. A graphical schematic of the grid search experiments is shown in fig. 2.3.

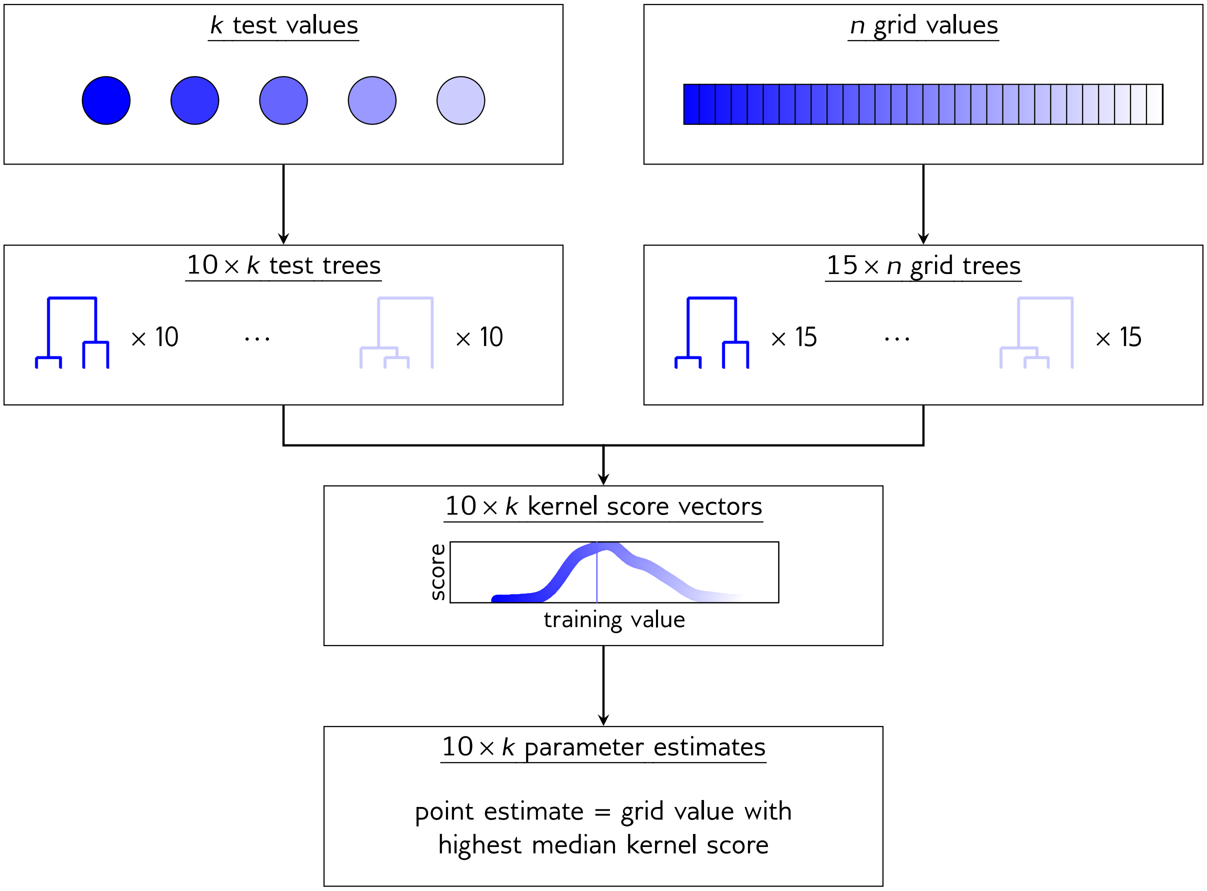
k test values
n grid values
10 × *k* test trees
× 10
· · ·
× 10
15 × *n* grid trees
× 15
· · ·
× 15
10 × *k* kernel score vectors
score
training value
10 × *k* parameter estimates
point estimate = grid value with
highest median kernel score

Figure 2.3: Graphical schematic of grid search experiments used to investigate BA model parameters.

Approximate Bayesian computation

To test the full kernel-ABC algorithm, we simulated four trees each under a variety of parameter values, and ran the *netabc* program to estimate posterior distributions for the parameters. The parameter values and priors used are listed in 2.3. The tree kernel meta-parameters were set to $\lambda = 0.3$ and $\sigma = 4$. The SMC algorithm was run with 1000 particles, five sampled datasets per particle, and the α parameter (not to be confused with the BA preferential attachment parameter, see section 2.1.1) set to 0.95. The algorithm was stopped when the acceptance rate of the Metropolis-Hastings (MH) kernel dropped below 1.5%, the same criterion used by Del Moral, Doucet, and Jasra. Approximate marginal posterior densities for each parameter were calculated using the *density* function in *R* applied to the final weighted population of particles. Credible intervals were obtained for each parameter using the *HPDinterval* function in the *coda* package [111].

parameter or variable	test values	prior
N	5000	Uniform(500, 15000)
α	0, 0.5, 1, 1.5, 2	Uniform(0, 2)
m	2, 3, 4	Uniform(1, 5)
I	1000, 2000	Uniform (1000, 2000)
tips	500	-

Table 2.3: Variables and BA parameter values used for ABC validation experiments. Trees were simulated under the test values, and kernel-ABC was used to re-estimate posterior distributions for the BA parameters without training.

parameter	values
N	500, 600, . . . , 15000
α	0, 0.01, . . . , 2
m	1, 2, . . . , 8

Table 2.4: BA model parameters used as input to GLM predicting power law exponent γ. One network was simulated with each combination of parameters, and γ was calculated for each network. A GLM with Gamma-distributed errors and a log link function was fit to the γ values with all parameters and interaction terms as predictors.

2.1.3 Real data experiments

Because the BA model assumes a single connected contact network, it is most appropriate to apply to groups of individuals who are epidemiologically related. Therefore, we searched for published HIV datasets which originated from existing clusters, either phylogenetically or geographically defined. In addition, we analysed an in-house dataset sampled from HIV-positive individuals in British Columbia, Canada (the "BC data"). The datasets are summarized in table 2.5.

We downloaded all sequences associated with each published study from GenBank. For the Novitsky et al. [112] data, each *env* sequence was aligned pairwise to the HXB2 reference sequence (GenBank accession number K03455) and the hypervariable regions were clipped out with *BioPython* version 1.66+ [113]. Sequences were multiply aligned using *MUSCLE* version 3.8.31 [114], and alignments were manually inspected with *Seaview* version 4.4.2 [115]. Phylogenies were constructed from the nucleotide alignments by approximate maximum likelihood using *FastTree2* version 2.1.7 with the generalized 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 [116]) as described previously [117].

Three of the datasets [112, 118, and the BC data] were initially much larger than the others, containing 1265, 1299, and 7923 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. and BC datasets, we detected clusters of size 280 and 399 respectively using a patristic distance cutoff of 0.02 as described previously [66]. Only sequences within these clusters were carried forward. For the Novitsky et al. [112] data, no large clusters were detected using the same cutoff, so we analysed a subtree of size 180 chosen arbitrarily.

Reference	Sequences (n)	Location	Risk group	Gene
Wang et al. [61]	173	Beijing, China	MSM	*pol*
Cuevas et al. [119]	287	Basque Country, Spain	mixed	*pol*
Novitsky et al. [120] Novitsky et al. [112]	180	Mochudi, Botswana	mixed	*env*
Li et al. [118]	280	Shanghai, China	MSM	*pol*
Niculescu et al. [121]	136	Romaina	IDU	*pol*
unpublished	399	British Columbia, Canada	IDU	*pol*

Table 2.5: Characteristics of published HIV datasets analyzed with kernel-ABC.

2.2 Results

2.2.1 Analysis of Barabási-Albert model

Classifiers for parameters based on tree shape

Trees simulated under different values of α were visibly quite distinct (fig. 2.4). In particular, higher values of α produce networks with a small number of highly connected nodes which, once infected, are likely to transmit to many other nodes. This results in a more unbalanced, ladder-like structure in the phylogeny, compared to networks with lower α values. None of the other three parameters produced trees which were as easily distinguished from each other (figs. S1 to S3). Sackin's index, which measures tree imbalance, was significantly correlated with all four parameters (for α, I, m, and N respectively: Spearman's rho = 0.85, -0.12, -0.13, 0.09; p-values $<10^{-5}$, 0.003, $<10^{-5}$, $<10^{-5}$) The ratio of internal to terminal branch lengths was negatively correlated with α and I, and positively corelated with N and m (Spearman's rho -0.84, -0.69, 0.1, 0.18; all $p < 10^{-5}$).

Figure 2.5 shows kPCA projections of the simulated trees onto the first two principal components of the kernel matrix. The figure shows only the simulations with 500-tip trees and 1000 infected nodes. The three α and I values considered are well separated from each other in feature space. On the other hand, the three N values overlap significantly, and the three m values are virtually indistinguishable. Similar observations can be made for other values of I and the number of tips (figs. S8 to S11). The values of I and N separated more clearly with larger numbers of tips, and in the case of N, larger epidemic sizes.

Accuracy of the kSVR classifiers varied based on the parameter being tested (fig. 2.6, 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. 2.6, centre and right). The results were largely robust to variations in the tree kernel meta-parameters λ and σ, although accuracy varied between different epidemic and sampling scenarios (figs. S4 to S7).

When classifying α, the kSVR classifier had an average R^2 of 0.92, compared to 0.56 for the nLTT-based SVR, and 0.75 for the linear regression against Sackin's index. There was little variation about the mean for different tree and epidemic sizes. No classifier could accurately identify the m parameter in any epidemic scenario, with average R^2 values of 0.12 for kSVR, 0.01 for the nLTT, and 0.06 for

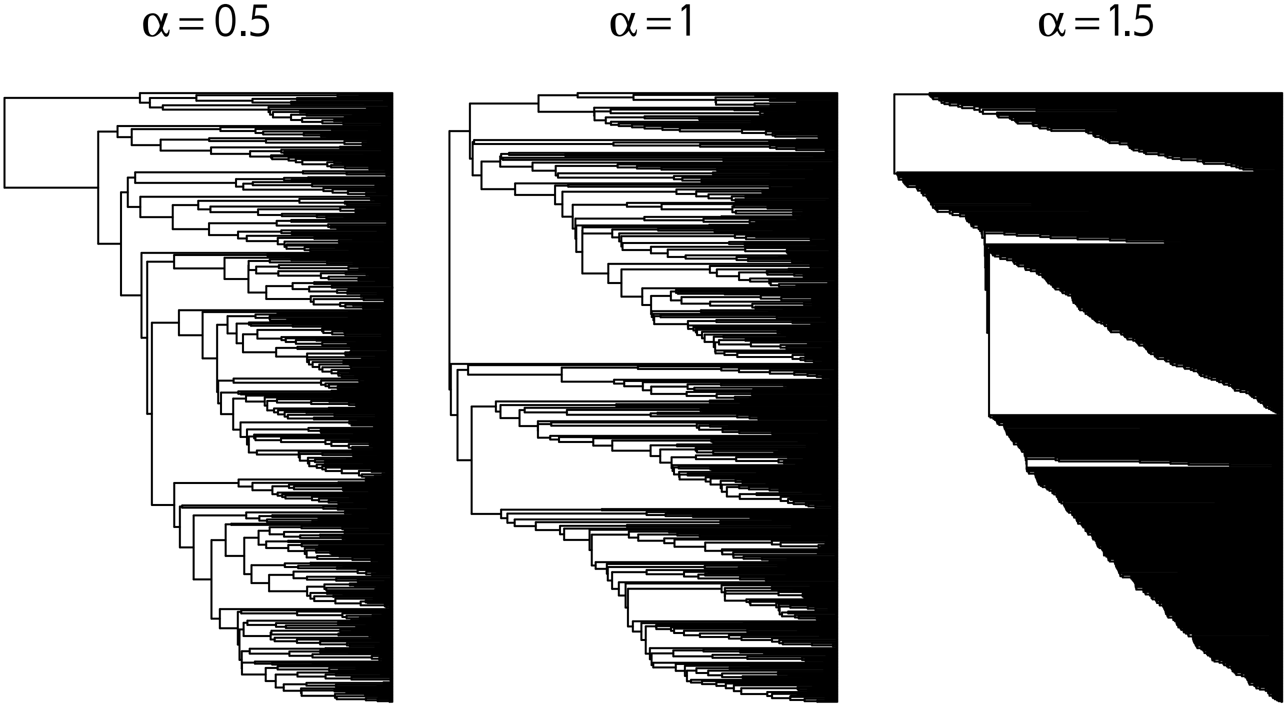

Figure 2.4: Epidemics simulated on BA networks of 5000 nodes, with α equal to 0.5, 1.0, or 1.5, until 1000 individuals were infected. Transmission trees were created by sampling 500 infected nodes. Higher α values produced networks with a small number of highly-connected nodes, resulting in highly unbalanced, ladder-like trees.

Sackin's index. Again, there was little variation in accuracy between epidemic scenarios, although the accuracy of the kSVR was slightly higher on 1000-tip trees (average R^2 0, 0, 0 for 100, 500, and 1000 tips respectively).

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 kSVR classifier ranged from 0.08 for the smallest epidemic and smallest sample size, to 0.82 for the largest. Likewise, R^2 for the nLTT-based SVR ranged from 0.01 to 0.54. Sackin's index did not accurately classify N in any scenario, with an average R^2 of 0.03 and little variation between scenarios.

Marginal parameter estimates with grid search

The accuracy of grid search estimates largely paralleled that of the kSVR classifiers. Figure 2.7 shows point estimates and 95% highest density intervals for each of the BA parameters, for one replicate experiment with 500-tip trees. For all parameters except m, the error of point estimates was negatively correlated with the number of sampled tips in the tree (for α, I, and N respectively: Spearman's ρ = −0.22, −0.51, −0.16; p-values 4×10^{-4}, $<10^{-5}$, 0.01). The highest density intervals obtained for all parameters were extremely wide, occcupying >90% of the grid in all cases (fig. 2.7).

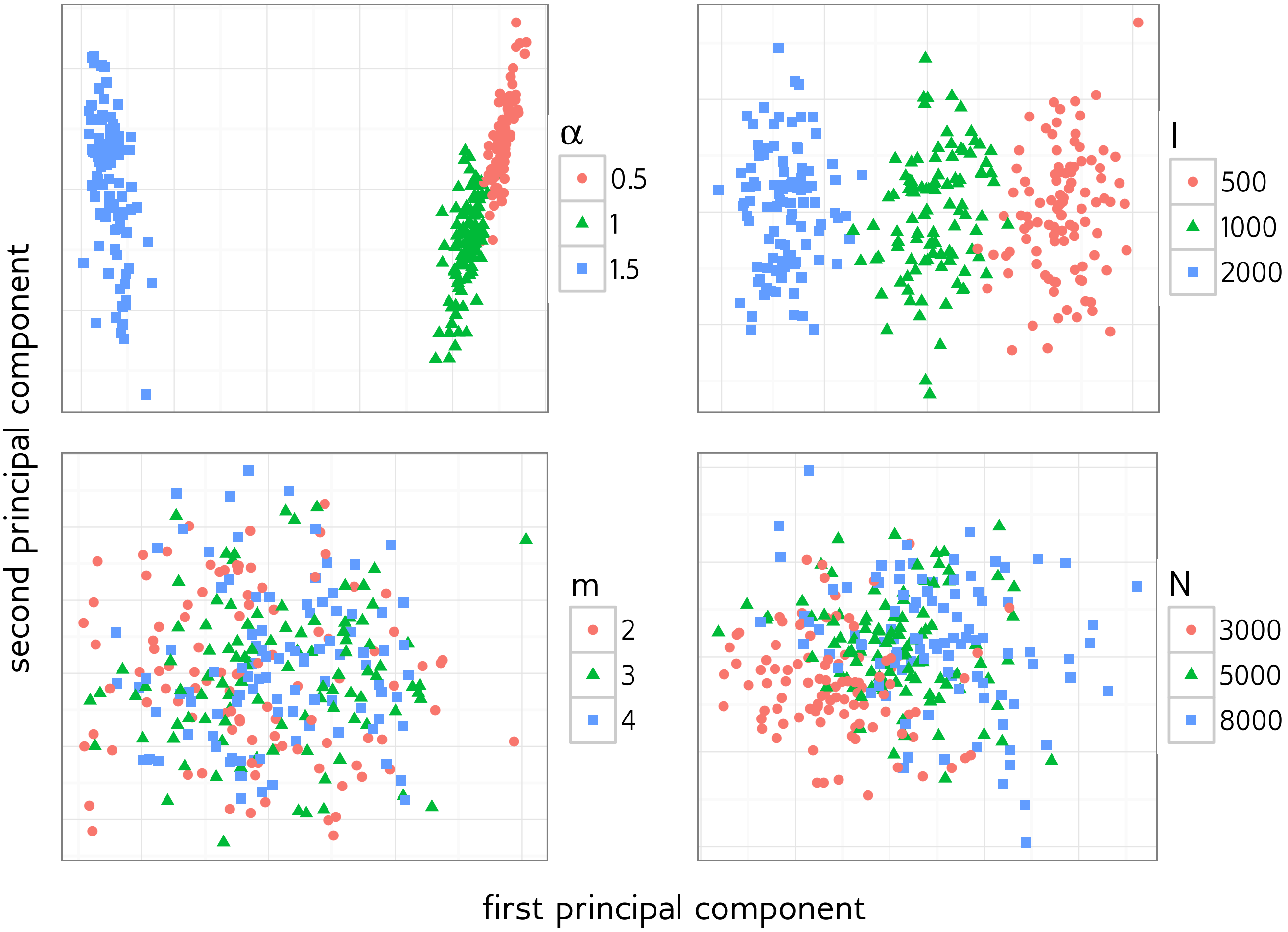

Figure 2.5: Each parameter of the BA model was individually varied to produce 300 simulated trees. Kernel matrices were formed from all pairwise kernel scores among each set of 300 trees. The trees were projected onto the first two principal components of the kernel matrix calculated using kPCA. All trees had 500 tips. The parameters not being varied were set to $\alpha = 1$, $I = 1000$, $m = 2$, and $N = 5000$. The tree kernel meta-parameters were $\lambda = 0.3$ and $\sigma = 4$.

The α parameter was the most accurately estimated, with point estimates having an average deviation of 0.14 from the true value, on a grid from 0 to 2. The error was negatively correlated with the true value of α (Spearman's $\rho = -0.26$, $p = 1\times10^{-5}$), although the relationship was clearly nonlinear (figs. 2.7 and S12). The accuracy was highest for the test value $\alpha = 1.25$ (mean error 0.02) which exhibited markedly different behaviour than the other values in terms of the distribution of kernel scores along the grid (fig. S12). In particular, there was a very pronounced peak in scores around the true value, in contrast to most other values where the scores were flat around the true value. The peak was also observed to a lesser degree for $\alpha = 1$. The average absolute error of the point estimates for I was 310, on a grid of 500 to 5000, and the errors were not significantly correlated with the true value of I. Kernel score distributions for all test values exhibited a similar rounded shape (fig. S13).

The average error for m was 1.31, on a grid from 1 to 6; this error was positively correlated with increasing m (Spearman's $\rho = 0.25$, $p = 6\times10^{-4}$). This positive correlation was apparently due to the

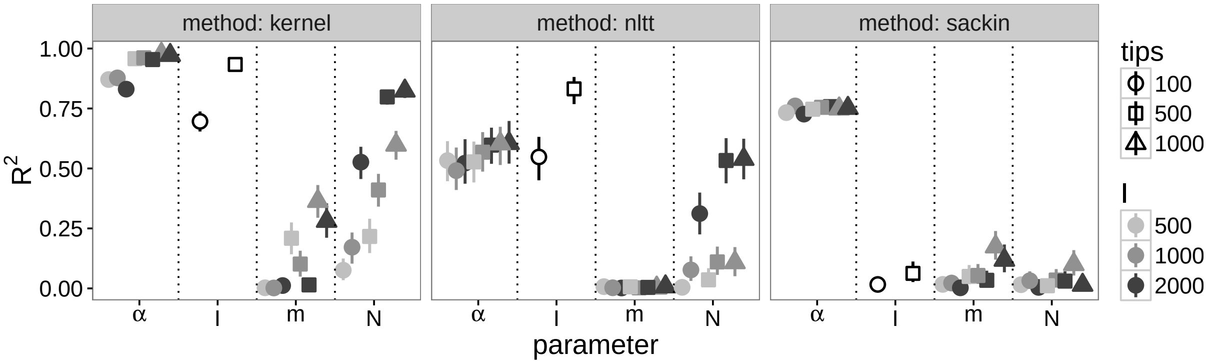

Figure 2.6: Cross-validation accuracy of kernel-SVR classifier (left), SVR classifier using nLTT (centre), and linear regression using Sackin's index (right) for BA model parameters. Kernel meta-parameters were set to λ = 0.3 and σ = 4. Each point was calculated based on 300 simulated transmission trees 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.

much lower error for m = 1 and 2 than for the other m values (mean errors 0.93 for $m \leq 2$ vs. 1.49 for $m > 2$, fig. S14). The value $m = 1$ causes the network to take on a distinct shape relative to higher m values, namely a tree (*i.e.* there are no cycles, see section 1.2.4). The average error for N was 2419, on a grid from 1000 to 15000, and was positively correlated with the true value of N (Spearman's ρ = 0.43, p<10^{-5}).

Joint parameter estimates with kernel-ABC

We used *netabc* to estimate the parameters of the BA model on simulated trees where the true parameter values were known. Point estimates for each parameter are shown in ?? for the simulations with m = 2. The results for the other values of m were similar (figs. S16 and S17). The median [IQR] absolute error of estimates of α across all simulations was 0.11 [0.05-0.18]. 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 = 6×10^{-4}). The error in the estimated value of I was 306 [108-607]. Errors were significantly higher for $\alpha \geq 1$ (Wilcoxon rank-sum test, p = 6×10^{-4}) and for I = 2000 (p = <10^{-5}), but not for any values of m (one-way ANOVA). The m parameter was estimated correctly in only 37 % of simulations. The true values of m and I did not significantly affect the error (one-way ANOVA), but the accuracy was significantly lower for integral than non-integral values of α (Wilcoxon rank-sum test, p = 7×10^{-3}). Finally, the total number of nodes N was consistently over-estimated by about a factor of two (error 6588 [4214 - 8284]). No other parameters influenced the accuracy of the N estimates (one-way ANOVA)

?? shows the ABC approximation to the posterior distribution on the BA parameters for one simulation (equivalent plots for all the simulations can be found in the supplemental materials). Highest posterior density (HPD) intervals around α and I were narrow relative to the region of nonzero prior density, whereas the intervals for m and N were widely dispersed. ?? shows point

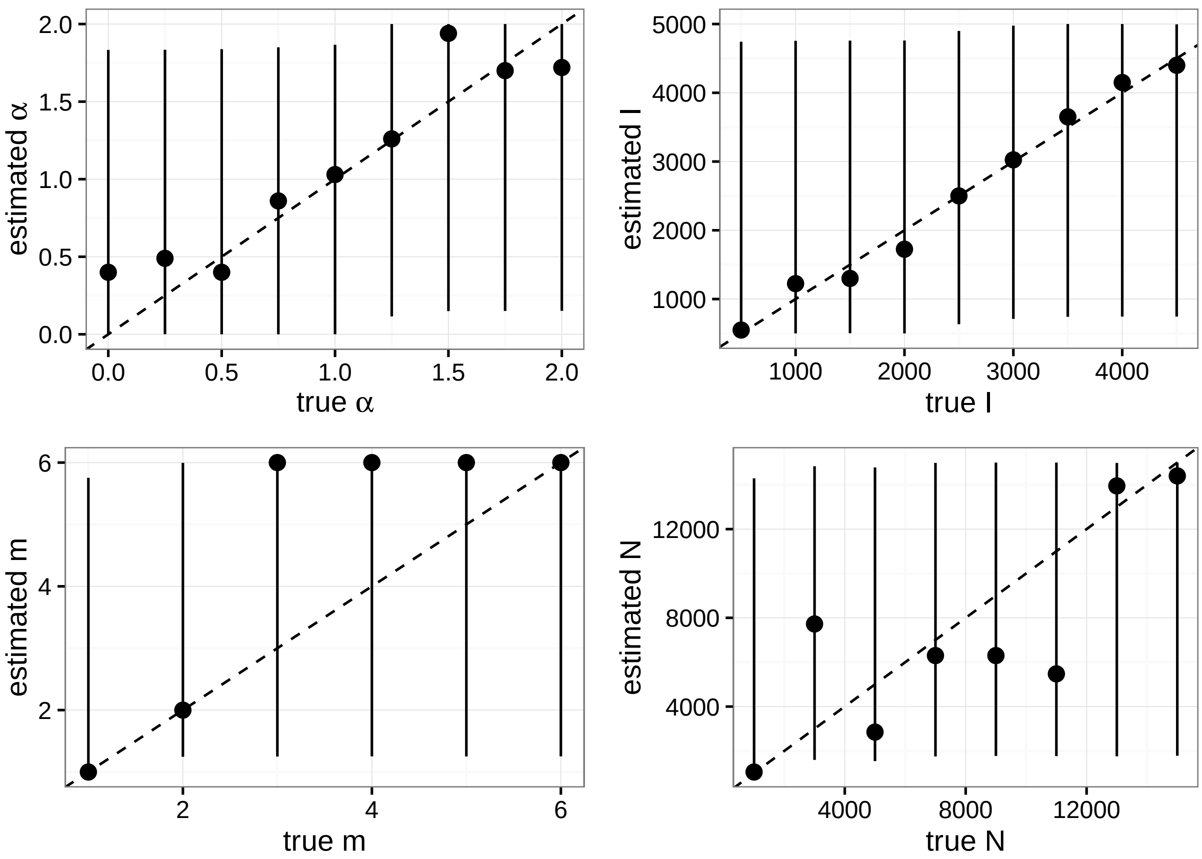

Figure 2.7: Point estimates and 95% highest density intervals for each BA model parameter, obtained using grid search. Networks and transmission trees were simulated over a grid of values for each parameter while holding the others fixed. For a subset of the grid values (x-axis), test networks and trees were created and compared to each tree on the grid using the tree kernel. The kernel scores along the grid were normalized to resemble a probability distribution, from which the mode and highest density interval were calculated. Shown values correspond to one replicate experiment, with trees of size 500.

estimates and 95% HPD intervals averaged over all simulations.

2.2.2 Application to HIV data

We applied kernel-ABC to five published HIV datasets (??), and found substantial heterogeneity among the parameter estimates (fig. 2.9). Two of the datasets [61, 121] had estimated α values near unity (MAP estimates [95% HPD] 1.06 [0.63 - 1.27] and 1 [0.41 - 1.16] respectively). Another two datasets [118, 119] 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. [112] 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 - 1.51×10^3]; Niculescu et al., 175 [138 - 454]) and wider intervals around the other two (Li et al., 1.59×10^3 [284 - 3.81×10^3]; Wang et al., 651 [268 - 4.24×10^3]). The Novitsky et al. data was again the outlier, with

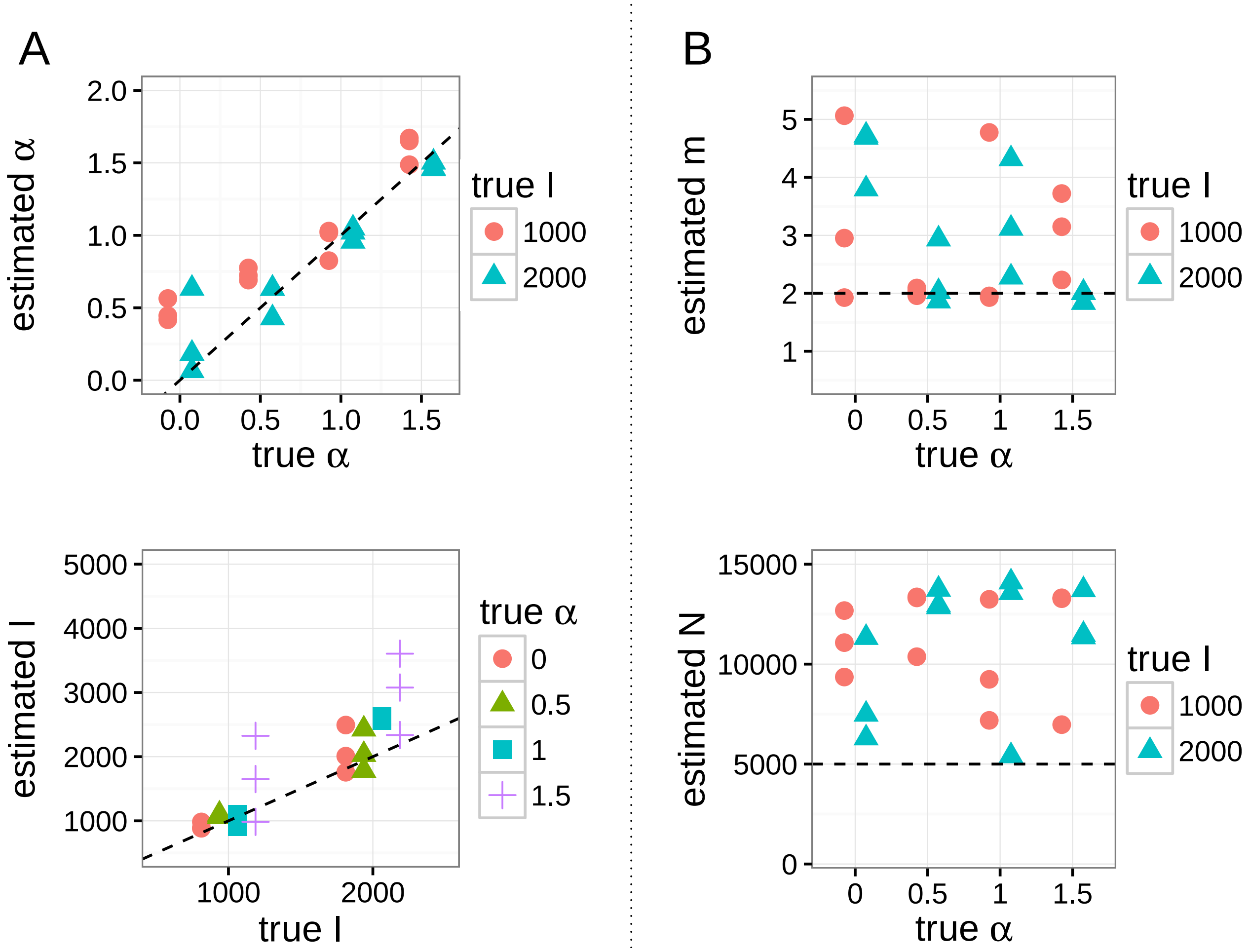

Figure 2.8: Maximum *a posteriori* point estimates for BA model parameters obtained by running *netabc* on simulated data. Values shown are for simulations with $m = 3$. Dashed lines indicate true values. (A) Estimates of α and I which were varied in these simulations against known values. (B) Estimates of m and N which were held fixed in these simulations at the values $m = 3$ and $N = 5000$.

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 2.6: Average widths of 95% confidence intervals for BA model parameters estimated with kernel-ABC.

a very high estimated I, and HPD interval spanning almost the entire prior region (7.55×10^3 [228 - 8.92×10^3]). 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 (5.84×10^3) than the estimates for other datasets (average 8.93×10^3).

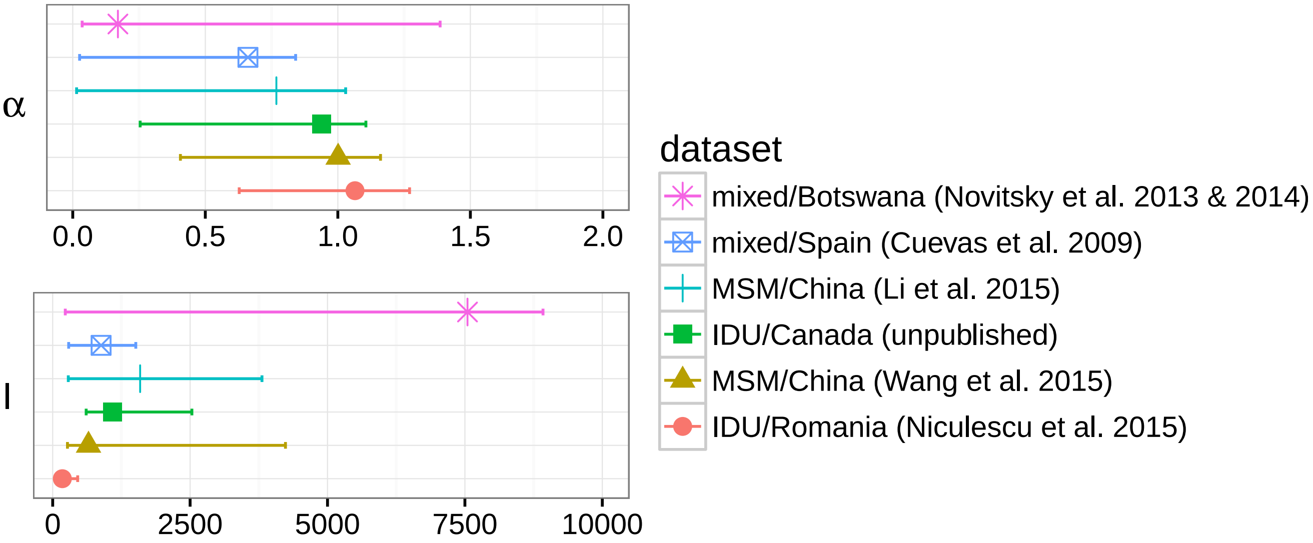

Figure 2.9: Maximum *a posteriori* point estimates and 95% HPD intervals for parameters of the BA network model, fitted to five published HIV datasets with kernel-ABC.

2.3 Discussion

2.3.1 *Netabc*: uses, limitations, and possible extensions

The method behind *netabc* is model-agnostic, meaning that it can be used to infer parameters of any network model, as long as it allows simulated networks can be easily generated. We have included generators for the BA model discussed here, as well as the ER and Watts-Strogatz (WS) network models. Instructions for adding additional models are available in the project's online documentation. We have made *netabc* publicly available at github.com/rmcclosk/netabc under a permissive open source license, to encourage other researchers to apply and extend our method.

In retrospect, there are a few algorithmic improvements which could have been made to *netabc*. Firstly, our implementation uses simple multinomial sampling for the resampling step, as well as for generating the final sample from the target distribution. Several other sampling strategies have been developed [122], and it is possible that the use of a more sophisticated scheme might increase the algorithm's accuracy. Secondly, the model parameters I, N, and m are all integer-valued, but we treat them as continuous-valued when performing SMC. This is not inherently wrong, in that it simply assigns equal probability density to the interval $[i, i + 1)$ for all integers i. However, computational effort could be wasted if particles are moved around within these intervals. To improve efficiency, alternative proposals could be used for discrete-valued particles.

The two-step process of simulating a contact network and subsequently allowing an epidemic to spread over that network carries with it the assumption that the contact network is static over the duration of the epidemic. Clearly this assumption is invalid, as people make and break partnerships on a regular basis. Our work has not addressed this assumption, primarily due to our desire to avoid the additional complexity required to address the dynamic nature of networks. This simplifying assumption is made by most studies using contact network models in an epidemiological context [12, 54]. However, in principle, kernel-ABC could be adapted to dynamic contact networks by using a method such as that developed by Robinson, Cohen, and Colijn [123] to simulate a dynamic contact network, while concurrently simulating the spread of an epidemic.

An alternative approach is the deterministic framework outlined by Morris [50], who proposes to apply the standard compartmental modelling framework to contact networks by assigning each individual their own compartment. Thus, each individual is associated with a single ordinary differential equation (ODE), with the entire ODE system parameterized by the adjacency matrix of the contact network. Morris proposes to use log-linear models to parameterize the matrix. This framework is highly expressive, and allows straightforward incorporation of time-dependent dynamics. However, simulating a transmission tree would require the numerical solution of a very large system of ODEs. Given the large number of simulations required for kernel-ABC, it is not clear if this method would be computationally feasible in this context.

It is important to note that *netabc* takes a transmission tree as input, rather than a viral phylogeny. Thus, we have left the estimation of a transmission tree up to the user. There were two reasons for this choice. First and foremost, we wished again to avoid extra complexity and keep the number of estimated parameters small. In theory, it is possible to incorporate the process by which a viral phylogeny is generated along with a transmission tree into our method, for example by simulating within-host dynamics. Although this may be an avenue for future extension, we felt that it would obscure the primary purpose of this work, which is to study contact network parameters. Second, there are a number of different methods available for inferring transmission trees [37–40, 117], some of which incorporate geographic and/or epidemiological data not accommodated by our method. We therefore felt it would be best to allow researchers to use their own preferred tree building method.

2.3.2 Analysis of Barabási-Albert model

The preferential attachment power α had a very strong influence on tree shape in the range of values we considered. Although the tree kernel was the most effective classifier for α, a tree balance statistic performed nearly as well. This result was intuitive: high α values produce networks with few well-connected superspreader nodes which are involved in a large number of transmissions, resulting in a highly unbalanced ladder-like tree structure. The I parameter, representing the prevalence at the time of sampling, was also generally estimable. The dynamics of the SI model, and the coalescent process, offer a potential explanation for this result. In the initial phase of the epidemic, when I is small, each new transmission results in potentially many new discordant edges, thus decreasing the waiting time until the next transmission. Hence, there is an early exponential growth phase, producing many short branches near the root of the tree. As the epidemic gets closer to saturating the network, the number of discordant edges decays, causing longer waiting times. The distribution of coalescent times in the tree should therefore be informative about I. This information is captured by the tree kernel, and also by the nLTT statistic, which both performed quite will in classifying I (fig. 2.6).

The number of nodes in the network, N, exhibited the most variation in terms of being estimable. There was almost no difference between trees simulated under different N values when the number of infected nodes I was very small. There is an intuitive explanation for this result, namely that adding additional nodes does not change the edge density or overall shape of a BA network. This can be illustrated by imagining that we add a small number of nodes to a network after the epidemic simulation has already been completed. It is possible that none of these new nodes attains a connection to any infected node. Thus, running the simulation again on the new, larger network could produce the exact same transmission tree as before. On the other hand, when I is large, the coalescent dynamics discussed above for I also apply, as evidenced by the relative accuracy of the nLTT. The m parameter, which controls the number of connections added to the network per vertex, did not have a measurable impact on tree shape and was not estimable with kernel-ABC. The exception to this was the value $m = 1$, which produces networks without cycles whose associated trees were more easily distinguished. However, all the analyses presented here did not take the absolute size of the transmission trees into account, as the branch lengths were rescaled by their mean. Because higher m values imply higher edge density, an epidemic should spread more quickly for higher m than lower m with the same per-edge transmission probability. Hence, considering the absolute height of the trees may improve our method's ability to reconstruct m.

In addition to the tree height, many summary statistics have been developed to capture particular details of tree shape. Two of these, Sackin's index and the ratio of internal to terminal branch lengths, were correlated with every BA parameter. Classifiers based on Sackin's index and the nLTT similarity measure performed well in some cases, though poorly in others. ABC is often applied using a vector of summary statistics, rather than a kernel-based similarity score as we have done here, and methods have been developed to select an optimal combination of summary statistics for a given inference task [124]. Hence, an avenue for future improvement of our method may be the inclusion of additional summary statistics to supplement the tree kernel. In addition, all four parameters were more accurately classified when the number of tips in the transmission trees was larger, underscoring the importance of adequate sampling for accurate phylodynamic inference.

For the more estimable parameters, the credible intervals attained from the marginal ABC target distributions were much narrower than those obtained through grid search, while point estimates were of comparable accuracy. This was likely due to the fact that SMC employs importance sampling to approximate the posterior distribution, while grid search simply calculates a distance metric which may not have any resemblance to the posterior. Admittedly, our method of finding credible intervals from kernel scores along the grid, namely by normalizing the scores to resemble a probability distribution, was somewhat ad hoc, which may also have played a role. Regardless, this result indicates that there is benefit to applying the more sophisticated method, even if values for some of the parameters are known *a priori*, and especially if credible intervals are desired on the parameters of interest.

2.3.3 Application to HIV data

The analysis of HIV datasets indicated heterogeneity in the contact network structures underlying several distinct local epidemics. The five datasets analysed fell into three categories (Figure 2.9). First, we estimated a preferential attachment power between 0.5 and 1 for the epidemics studied by Cuevas et al. [119] and Li et al. [118], 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 cutoff [66]. 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 in the data introduced errors into our estimated viral phylogeny.

For the outbreaks studied by Niculescu et al. [121] and Wang et al. [61], 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 users (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. [120] sampled approximately 44% of the HIV-infected individuals in the northern area of Mochudi, Botswana. Additional sampling in a later study [112] 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 the majority of highly connected nodes were 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.

Our use of the BA model makes several simplifying assumptions. First, we assume homogeneity across the network with respect to node behaviour and transmission risk. In reality, the attraction to high-degree nodes seems likely to vary among individuals, as does their risk of transmitting or contracting the virus. We have also assumed that all transmission risks are symmetric, which is clearly false for all known modes of HIV transmission, and that infected individuals never recover but remain infectious indefinitely. These assumptions were made for the purpose of keeping the model as simple as possible, since this is the very first attempt to fit a contact network model in a phylodynamic context. However, the Gillespie simulation algorithm built into *netabc* can handle arbitrary transmission and removal rates which need not be homogeneous across the network. Moreover, it is possible to use kernel-ABC to fit a model which relaxes some or all of these assumptions, which may be a fruitful avenue for future investigation.

Chapter 3

Conclusion

This thesis developed *netabc*, a computer program implementing a statistical inference method for contact network parameters from viral phylogenetic data.

Bibliography

[1] William Heaton Hamer. *The Milroy lectures on epidemic disease in England: the evidence of variability and of persistency of type*. Bedford Press, 1906.

[2] William O Kermack and Anderson G McKendrick. "A contribution to the mathematical theory of epidemics". In: *Proceedings of the Royal Society of London A: mathematical, physical and engineering sciences*. Vol. 115. 772. The Royal Society. 1927, pp. 700–721.

[3] S Rushton and AJ Mautner. "The deterministic model of a simple epidemic for more than one community". In: *Biometrika* 42.1/2 (1955), pp. 126–132.

[4] Trevelyan McKinley, Alex R Cook, and Robert Deardon. "Inference in epidemic models without likelihoods". In: *The International Journal of Biostatistics* 5.1 (2009).

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

[6] Ernst Heinrich Haeckel. *Generelle Morphologie der Organismen*. Vol. 2. Verlag von Georg Reimer, 1866.

[7] EF Harding. "The probabilities of rooted tree-shapes generated by random bifurcation". In: *Advances in Applied Probability* (1971), pp. 44–77.

[8] Luigi L Cavalli-Sforza and Anthony WF Edwards. "Phylogenetic analysis. Models and estimation procedures". In: *American journal of human genetics* 19.3 Pt 1 (1967), p. 233.

[9] Sean Nee, Arne O Mooers, and Paul H Harvey. "Tempo and mode of evolution revealed from molecular phylogenies". In: *Proceedings of the National Academy of Sciences* 89.17 (1992), pp. 8322–8326.

[10] Peter Buneman. "A note on the metric properties of trees". In: *Journal of Combinatorial Theory, Series B* 17.1 (1974), pp. 48–50.

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

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

[13] Matt J Keeling and Ken TD Eames. "Networks and epidemic models". In: *Journal of the Royal Society Interface* 2.4 (2005), pp. 295–307.

[14] Eben Kenah et al. “Algorithms linking phylogenetic and transmission trees for molecular infectious disease epidemiology”. In: *arXiv preprint arXiv:1507.04178* (2015).

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

[16] Eddie C Holmes et al. “Revealing the history of infectious disease epidemics through phylogenetic trees”. In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 349.1327 (1995), pp. 33–40.

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

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

[19] Erik M Volz, Katia Koelle, and Trevor Bedford. “Viral phylodynamics”. In: *PLoS Comput Biol* 9.3 (2013), e1002947.

[20] Oliver G Pybus and Andrew Rambaut. “Evolutionary analysis of the dynamics of viral infectious disease”. In: *Nature Reviews Genetics* 10.8 (2009), pp. 540–550.

[21] David G Kendall. “On the generalized" birth-and-death" process”. In: *The annals of mathematical statistics* (1948), pp. 1–15.

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

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

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

[25] Masatoshi Nei and Sudhir Kumar. *Molecular evolution and phylogenetics*. Oxford University Press, 2000.

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

[27] Eben Kenah et al. “Molecular Infectious Disease Epidemiology: Survival Analysis and Algorithms Linking Phylogenies to Transmission Trees”. In: *PLOS Comput Biol* 12.4 (2016), e1004869.

[28] Jerry A Coyne and H Allen Orr. *Speciation*. Vol. 37. Sinauer Associates Sunderland, MA, 2004.

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

[30] Alexandros Stamatakis. "RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies". In: *Bioinformatics* (2014), btu033.

[31] RAJ Shankarappa et al. "Consistent viral evolutionary changes associated with the progression of human immunodeficiency virus type 1 infection". In: *Journal of virology* 73.12 (1999), pp. 10489–10502.

[32] Bette Korber et al. "Timing the ancestor of the HIV-1 pandemic strains". In: *Science* 288.5472 (2000), pp. 1789–1796.

[33] Alexei Drummond, G Oliver, Andrew Rambaut, et al. "Inference of viral evolutionary rates from molecular sequences". In: *Advances in parasitology* 54 (2003), pp. 331–358.

[34] Thu-Hien To et al. "Fast dating using least-squares criteria and algorithms". In: *Systematic biology* (2015), syv068.

[35] Wen-Hsiung Li, Masako Tanimura, and Paul M Sharp. "Rates and dates of divergence between AIDS virus nucleotide sequences." In: *Molecular Biology and Evolution* 5.4 (1988), pp. 313–330.

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

[37] Eleanor M Cottam et al. "Integrating genetic and epidemiological data to determine transmission pathways of foot-and-mouth disease virus". In: *Proceedings of the Royal Society of London B: Biological Sciences* 275.1637 (2008), pp. 887–895.

[38] T Jombart et al. "Reconstructing disease outbreaks from genetic data: a graph approach". In: *Heredity* 106.2 (2011), pp. 383–390.

[39] RJF Ypma et al. "Unravelling transmission trees of infectious diseases by combining genetic and epidemiological data". In: *Proceedings of the Royal Society of London B: Biological Sciences* 279.1728 (2012), pp. 444–450.

[40] Xavier Didelot, Jennifer Gardy, and Caroline Colijn. "Bayesian inference of infectious disease transmission from whole-genome sequence data". In: *Molecular Biology and Evolution* 31.7 (2014), pp. 1869–1879.

[41] Arne O Mooers and Stephen B Heard. "Inferring evolutionary process from phylogenetic tree shape". In: *Quarterly Review of Biology* (1997), pp. 31–54.

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

[43] Mark Kirkpatrick and Montgomery Slatkin. "Searching for evolutionary patterns in the shape of a phylogenetic tree". In: *Evolution* (1993), pp. 1171–1181.

[44] G Udny Yule. "A mathematical theory of evolution, based on the conclusions of Dr. JC Willis, FRS". In: *Philosophical Transactions of the Royal Society of London. Series B, Containing Papers of a Biological Character* 213 (1925), pp. 21–87.

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

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

[47] Christopher JC Burges. “A tutorial on support vector machines for pattern recognition”. In: *Data mining and knowledge discovery* 2.2 (1998), pp. 121–167.

[48] Michael Collins and Nigel Duffy. “New ranking algorithms for parsing and tagging: Kernels over discrete structures, and the voted perceptron”. In: *Proceedings of the 40th annual meeting on association for computational linguistics*. Association for Computational Linguistics. 2002, pp. 263–270.

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

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

[51] Jacob L Moreno. “Who shall survive”. In: *New York* (1953).

[52] John Arundel Barnes. *Class and committees in a Norwegian island parish*. Plenum New York, 1954.

[53] Stanley Wasserman and Katherine Faust. *Social network analysis: Methods and applications*. Vol. 8. Cambridge university press, 1994.

[54] Shweta Bansal, Bryan T Grenfell, and Lauren Ancel Meyers. “When individual behaviour matters: homogeneous and network models in epidemiology”. In: *Journal of the Royal Society Interface* 4.16 (2007), pp. 879–891.

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

[56] Roy M Anderson, Robert M May, and B Anderson. *Infectious diseases of humans: dynamics and control*. Vol. 28. Wiley Online Library, 1992.

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

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

[59] Duncan J Watts and Steven H Strogatz. “Collective dynamics of ‘small-world’ networks”. In: *Nature* 393.6684 (1998), pp. 440–442.

[60] Romualdo Pastor-Satorras and Alessandro Vespignani. “Epidemic spreading in scale-free networks”. In: *Physical review letters* 86.14 (2001), p. 3200.

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

[62] Susan J Little et al. “Using HIV Networks to Inform Real Time Prevention Interventions”. In: *PLoS ONE* 9.6 (2014).

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

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

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

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

[67] David L Yirrell et al. “Molecular epidemiological analysis of HIV in sexual networks in Uganda”. In: *AIDs* 12.3 (1998), pp. 285–290.

[68] Sonia Resik et al. “Limitations to contact tracing and phylogenetic analysis in establishing HIV type 1 transmission networks in Cuba”. In: *AIDS research and human retroviruses* 23.3 (2007), pp. 347–356.

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

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

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

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

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

[74] Hawoong Jeong et al. “The large-scale organization of metabolic networks”. In: *Nature* 407.6804 (2000), pp. 651–654.

[75] John T Kemper. “On the identification of superspreaders for infectious disease”. In: *Mathematical Biosciences* 48.1 (1980), pp. 111–127.

[76] Zhuang Shen et al. “Superspreading SARS events, Beijing, 2003”. In: *Emerging infectious diseases* 10.2 (2004), pp. 256–260.

[77] Herbert A Simon. “On a class of skew distribution functions”. In: *Biometrika* 42.3/4 (1955), pp. 425–440.

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

[79] Edwin J Bernard et al. “HIV forensics: pitfalls and acceptable standards in the use of phylogenetic analysis as evidence in criminal investigations of HIV transmission*”. In: *HIV medicine* 8.6 (2007), pp. 382–387.

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

[81] Vladimir N Minin, Erik W Bloomquist, and Marc A Suchard. “Smooth skyride through a rough skyline: Bayesian coalescent-based inference of population dynamics”. In: *Molecular Biology and Evolution* 25.7 (2008), pp. 1459–1471.

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

[83] Gabriel E Leventhal et al. “Inferring Epidemic Contact Structure from Phylogenetic Trees”. In: *PLoS Computational Biology* 8.3 (2012).

[84] David Welch. “Is network clustering detectable in transmission trees?” In: *Viruses* 3.6 (2011), pp. 659–676.

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

[86] Arnaud Doucet, Nando De Freitas, and Neil Gordon. “An introduction to sequential Monte Carlo methods”. In: *Sequential Monte Carlo methods in practice*. Springer, 2001, pp. 3–14.

[87] Jun S Liu. *Monte Carlo strategies in scientific computing*. Springer Science & Business Media, 2008.

[88] Jun S Liu, Rong Chen, and Tanya Logvinenko. “A theoretical framework for sequential importance sampling with resampling”. In: *Sequential Monte Carlo methods in practice*. Springer, 2001, pp. 225–246.

[89] Adrian Smith et al. *Sequential Monte Carlo methods in practice*. Springer Science & Business Media, 2013.

[90] Olivier Cappé, Simon J Godsill, and Eric Moulines. “An overview of existing methods and recent advances in sequential Monte Carlo”. In: *Proceedings of the IEEE* 95.5 (2007), pp. 899–924.

[91] Pierre Del Moral, Arnaud Doucet, and Ajay Jasra. “Sequential monte carlo samplers”. In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 68.3 (2006), pp. 411–436.

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

[93] Mark A Beaumont. “Approximate Bayesian computation in evolution and ecology”. In: *Annual review of ecology, evolution, and systematics* 41 (2010), pp. 379–406.

[94] Jean-Michel Marin et al. “Approximate Bayesian computational methods”. In: *Statistics and Computing* 22.6 (2012), pp. 1167–1180.

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

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

[97] Donald B Rubin et al. “Bayesianly justifiable and relevant frequency calculations for the applied statistician”. In: *The Annals of Statistics* 12.4 (1984), pp. 1151–1172.

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

[99] Paul Marjoram et al. “Markov chain Monte Carlo without likelihoods”. In: *Proceedings of the National Academy of Sciences* 100.26 (2003), pp. 15324–15328.

[100] Oliver Ratmann et al. “Using likelihood-free inference to compare evolutionary dynamics of the protein networks of H. pylori and P. falciparum”. In: *PLoS Comput Biol* 3.11 (2007), e230.

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

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

[103] Doug Baskins. *Judy arrays*. 2004.

[104] Brian Gough. *GNU scientific library reference manual*. Network Theory Ltd., 2009.

[105] Blaise Barney. “POSIX threads programming”. In: *National Laboratory. Available at:< https://computing. llnl. gov/tutorials/pthreads/> Accessed* 5 (2016).

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

[107] Alessandro Moschitti. “Making Tree Kernels Practical for Natural Language Learning.” In: *EACL*. Vol. 113. 120. 2006, p. 24.

[108] Emmanuel Paradis, Julien Claude, and Korbinian Strimmer. "APE: analyses of phylogenetics and evolution in R language". In: *Bioinformatics* 20.2 (2004), pp. 289–290.

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

[110] David Meyer et al. *e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien*. R package version 1.6-7. 2015. URL: https://CRAN.R-project.org/package=e1071.

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

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

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

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

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

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

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

[118] Xiaoyan Li et al. "HIV-1 Genetic Diversity and Its Impact on Baseline CD4+ T Cells and Viral Loads among Recently Infected Men Who Have Sex with Men in Shanghai, China". In: *PLoS ONE* 10.6 (2015).

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

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

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

[122] Randal Douc and Olivier Cappé. "Comparison of resampling schemes for particle filtering". In: *Image and Signal Processing and Analysis, 2005. ISPA 2005. Proceedings of the 4th International Symposium on*. IEEE. 2005, pp. 64–69.

[123] Katy Robinson, Ted Cohen, and Caroline Colijn. “The dynamics of sexual contact networks: effects on disease spread and control”. In: *Theoretical Population Biology* 81.2 (2012), pp. 89–96.

[124] Paul Fearnhead and Dennis Prangle. “Constructing summary statistics for approximate Bayesian computation: semi-automatic approximate Bayesian computation”. In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 74.3 (2012), pp. 419–474.

Appendix: Supplemental Figures

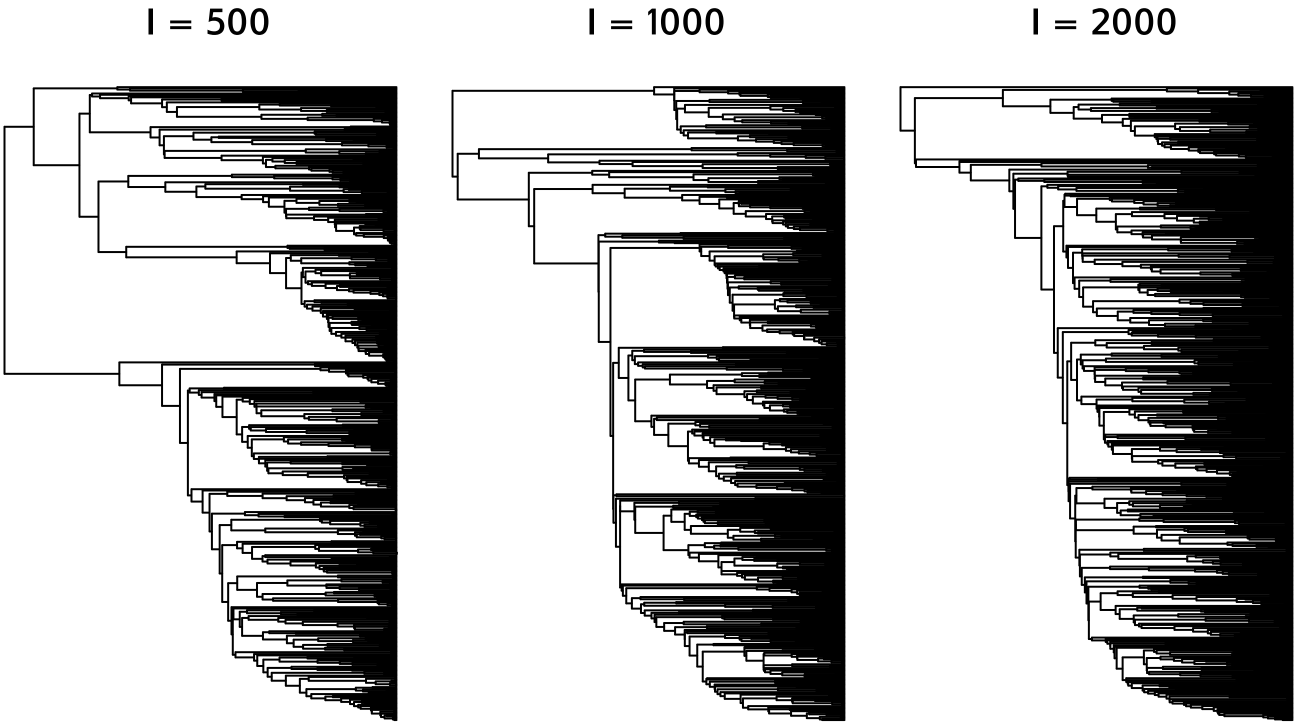

Figure S1: Transmission trees simulated over BA networks with varying values of I, the number of infected nodes when the epidemic simulation was stopped.

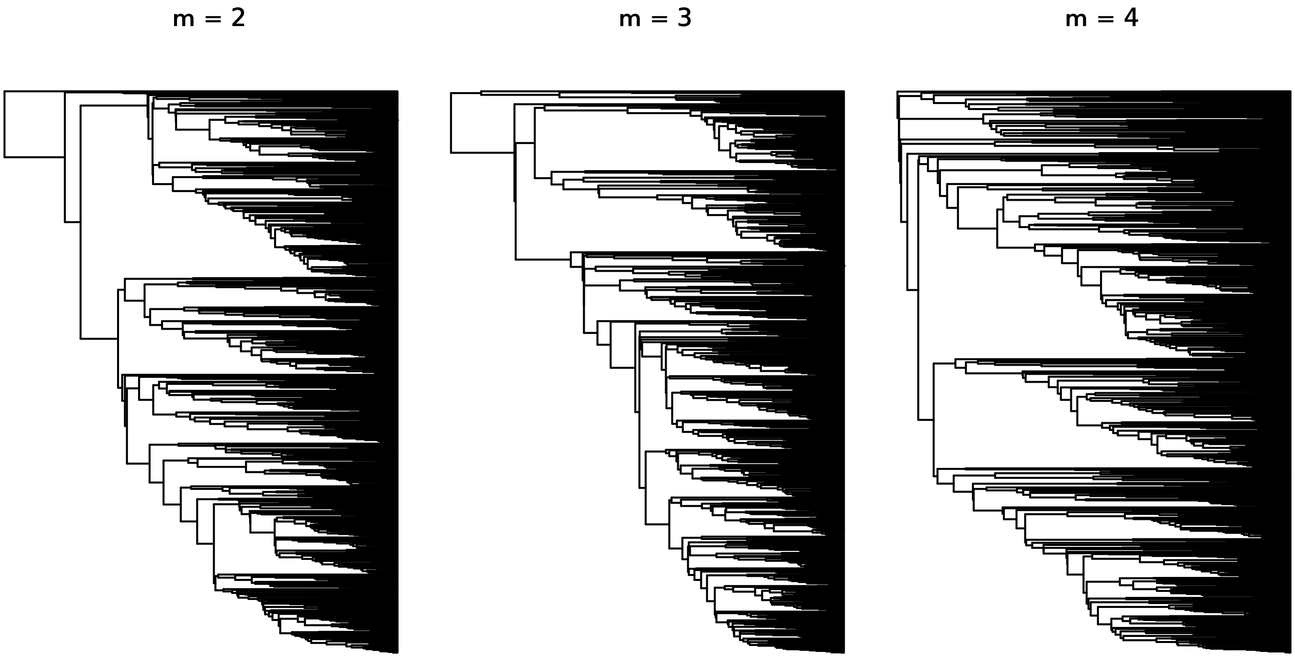

Figure S2: Transmission trees simulated over BA networks with varying values of m, the number of edges added per vertex.

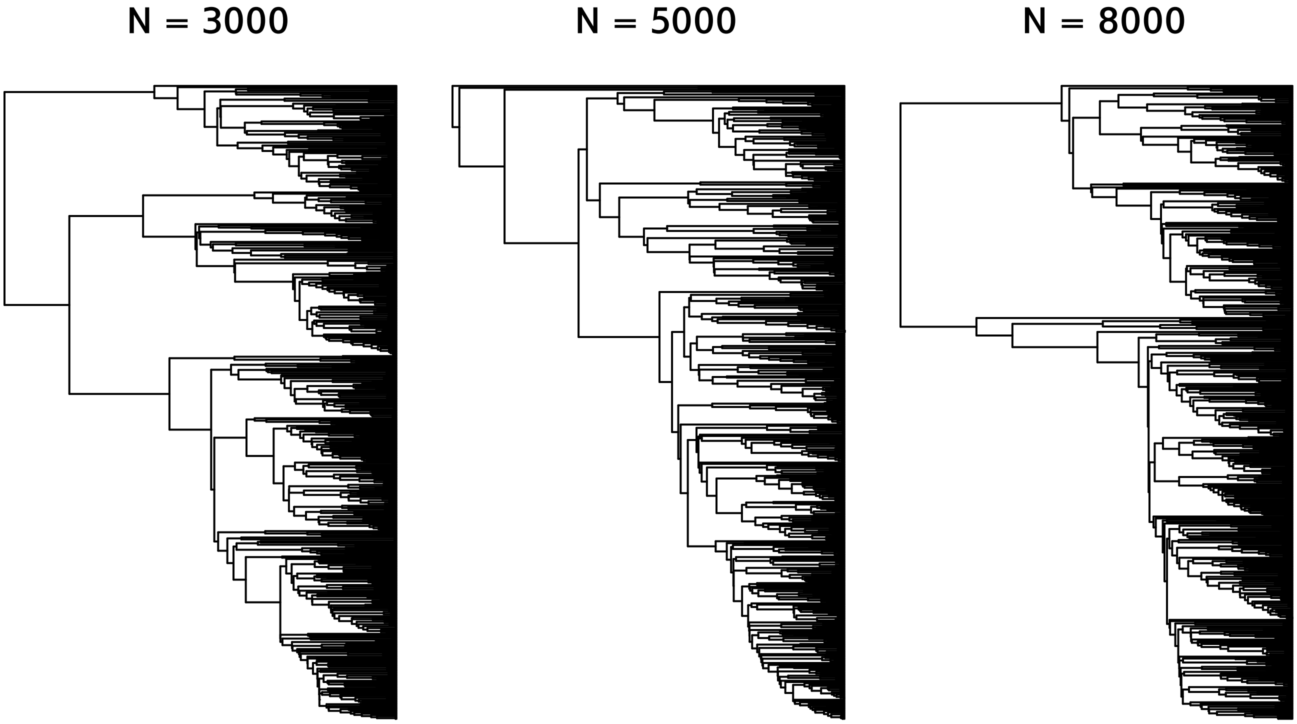

Figure S3: Transmission trees simulated over BA networks with varying values of N, the number of nodes in the network.

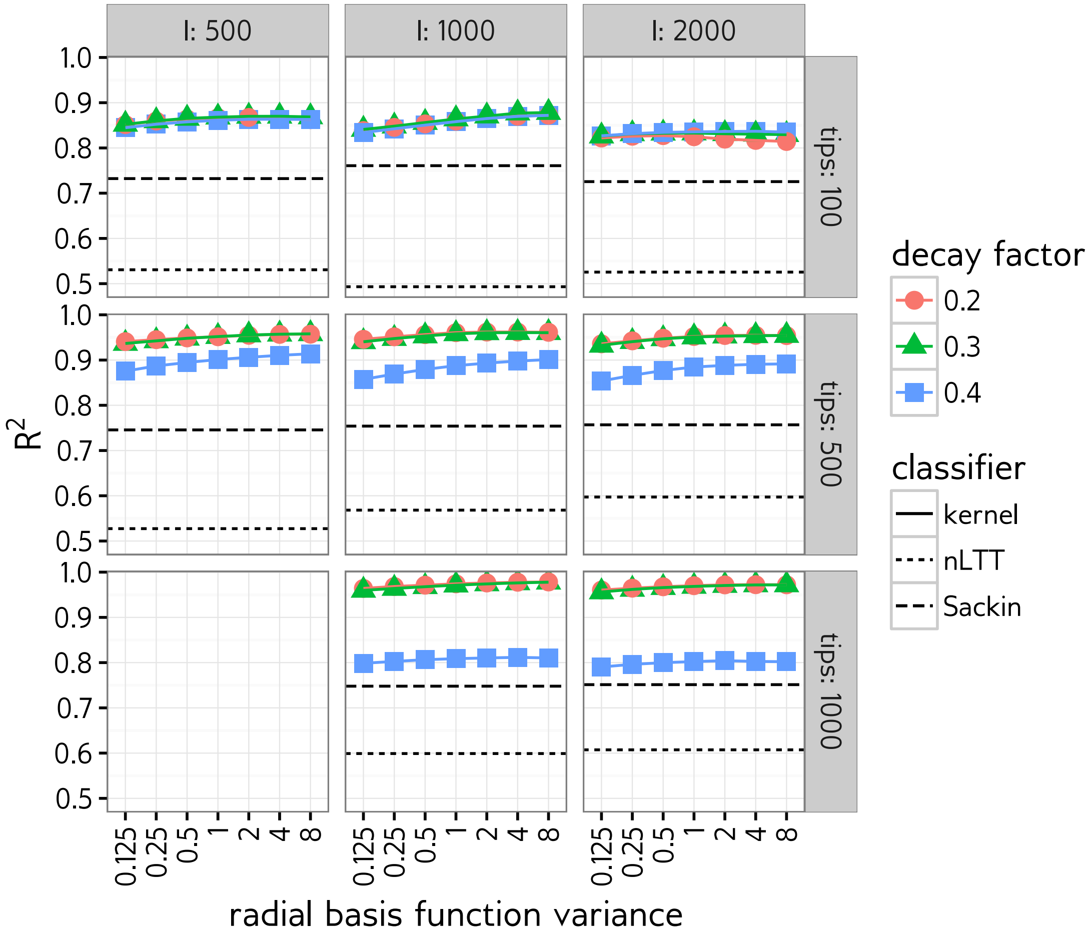

Figure S4: Cross validation accuracy of classifiers for BA model parameter α for eight epidemic scenarios. Solid lines and points are R^2 of tree kernel kSVR under various kernel meta-parameters. Dashed and dotted lines are R^2 of linear regression against Sackin's index, and SVR using nLTT.

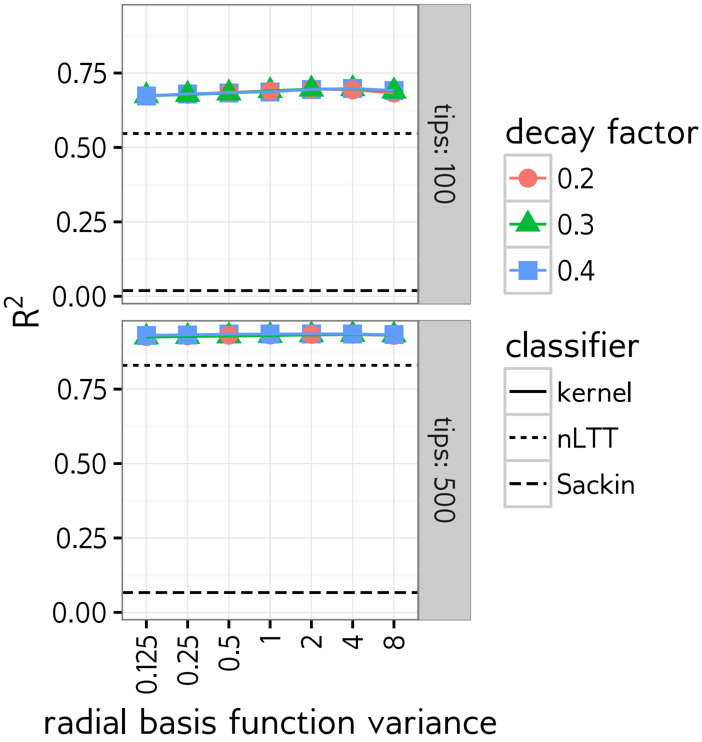

Figure S5: Cross validation accuracy of classifiers for BA model parameter I for eight epidemic scenarios. Solid lines and points are R^2 of tree kernel kSVR under various kernel meta-parameters. Dashed and dotted lines are R^2 of linear regression against Sackin's index, and SVR using nLTT.

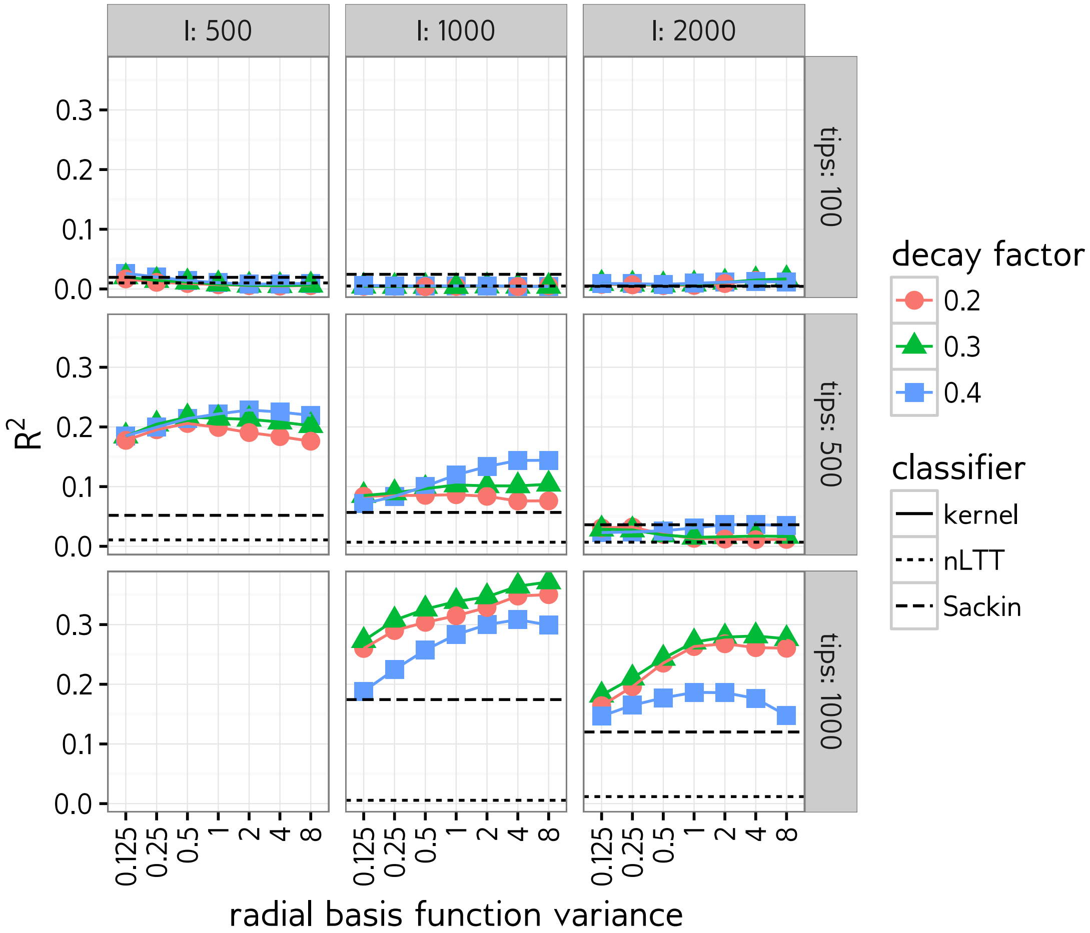

Figure S6: Cross validation accuracy of classifiers for BA model parameter m for eight epidemic scenarios. Solid lines and points are R^2 of tree kernel kSVR under various kernel meta-parameters. Dashed and dotted lines are R^2 of linear regression against Sackin's index, and SVR using nLTT.

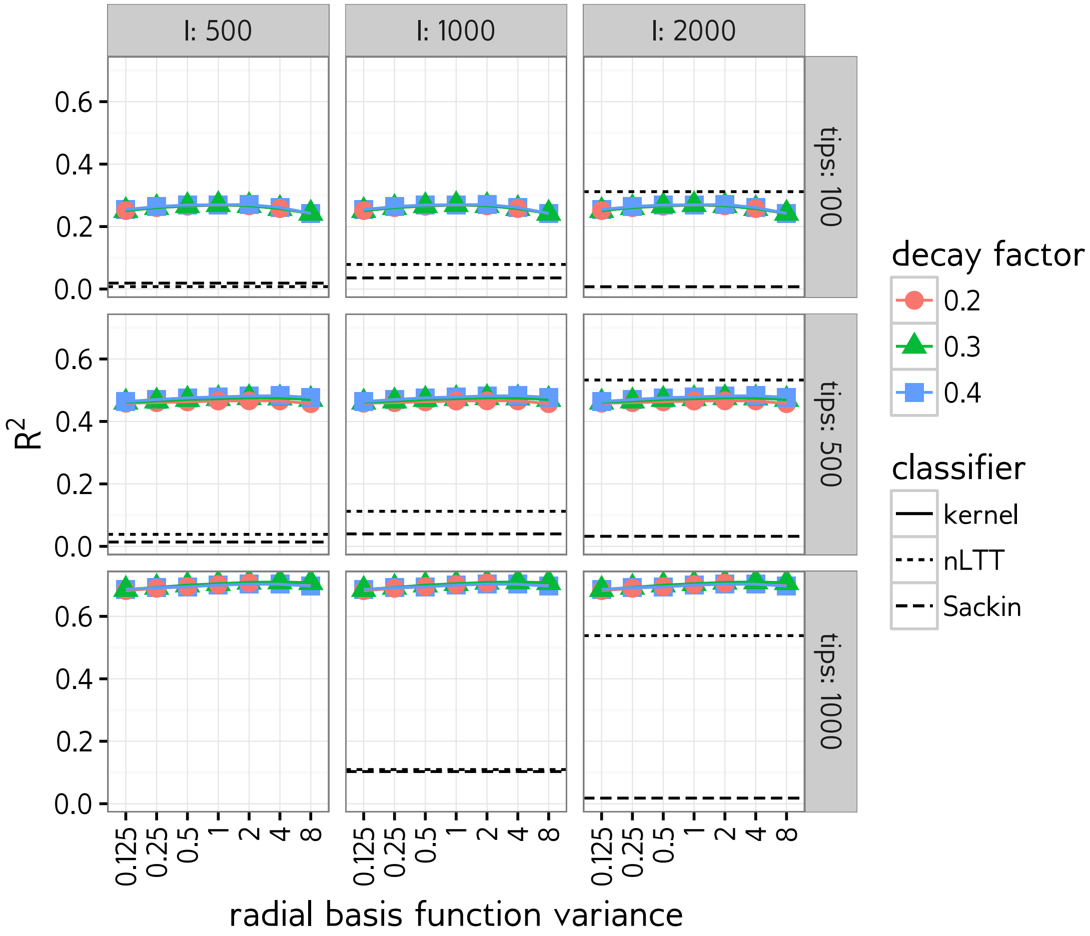

Figure S7: Cross validation accuracy of classifiers for BA model parameter N for eight epidemic scenarios. Solid lines and points are R^2 of tree kernel kSVR under various kernel meta-parameters. Dashed and dotted lines are R^2 of linear regression against Sackin's index, and SVR using nLTT.

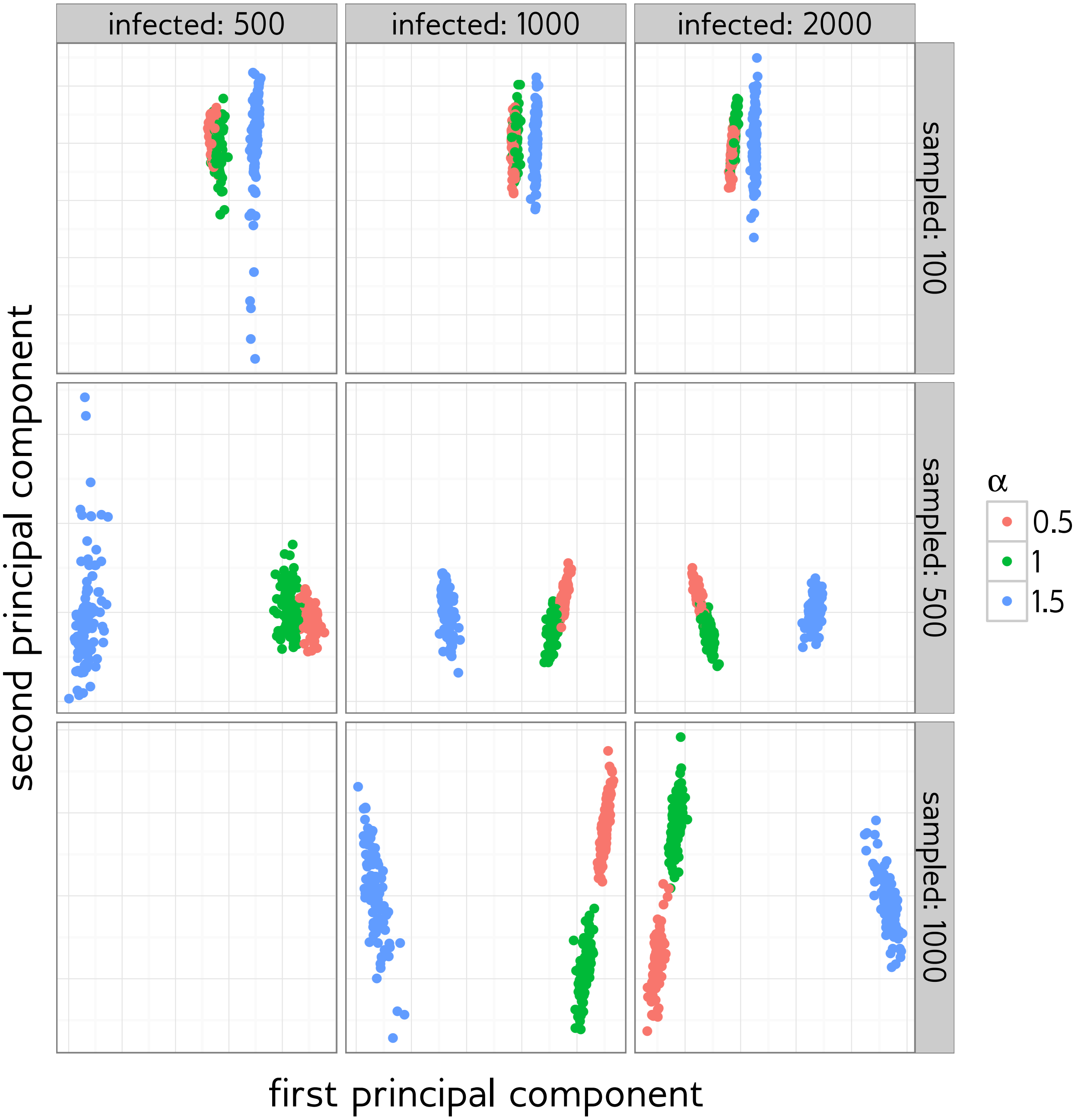

Figure S8: Kernel principal components projection of trees simulated under three different values of BA parameter α, for eight epidemic scenarios.

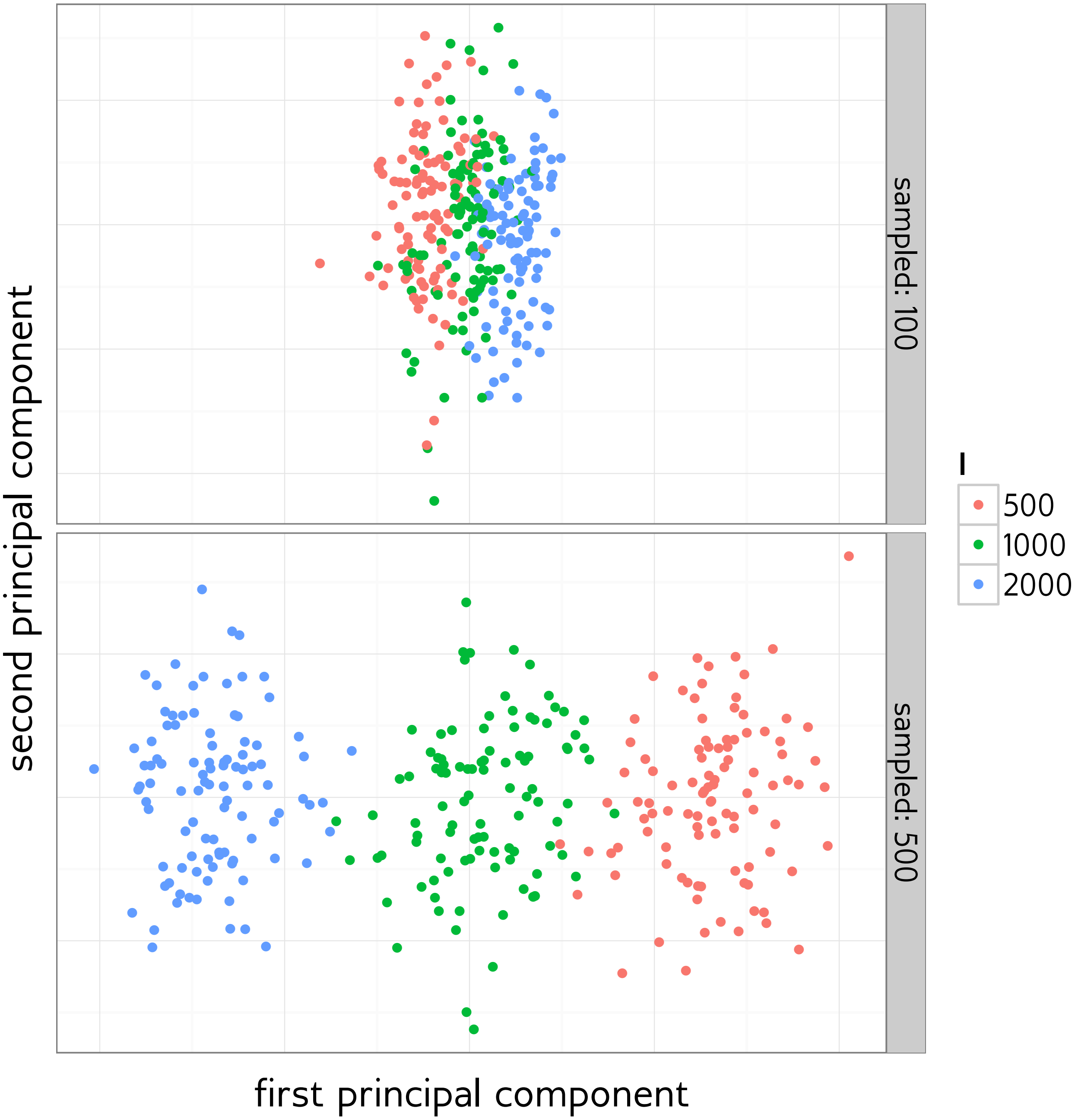

Figure S9: Kernel principal components projection of trees simulated under three different values of BA parameter I, for eight epidemic scenarios.

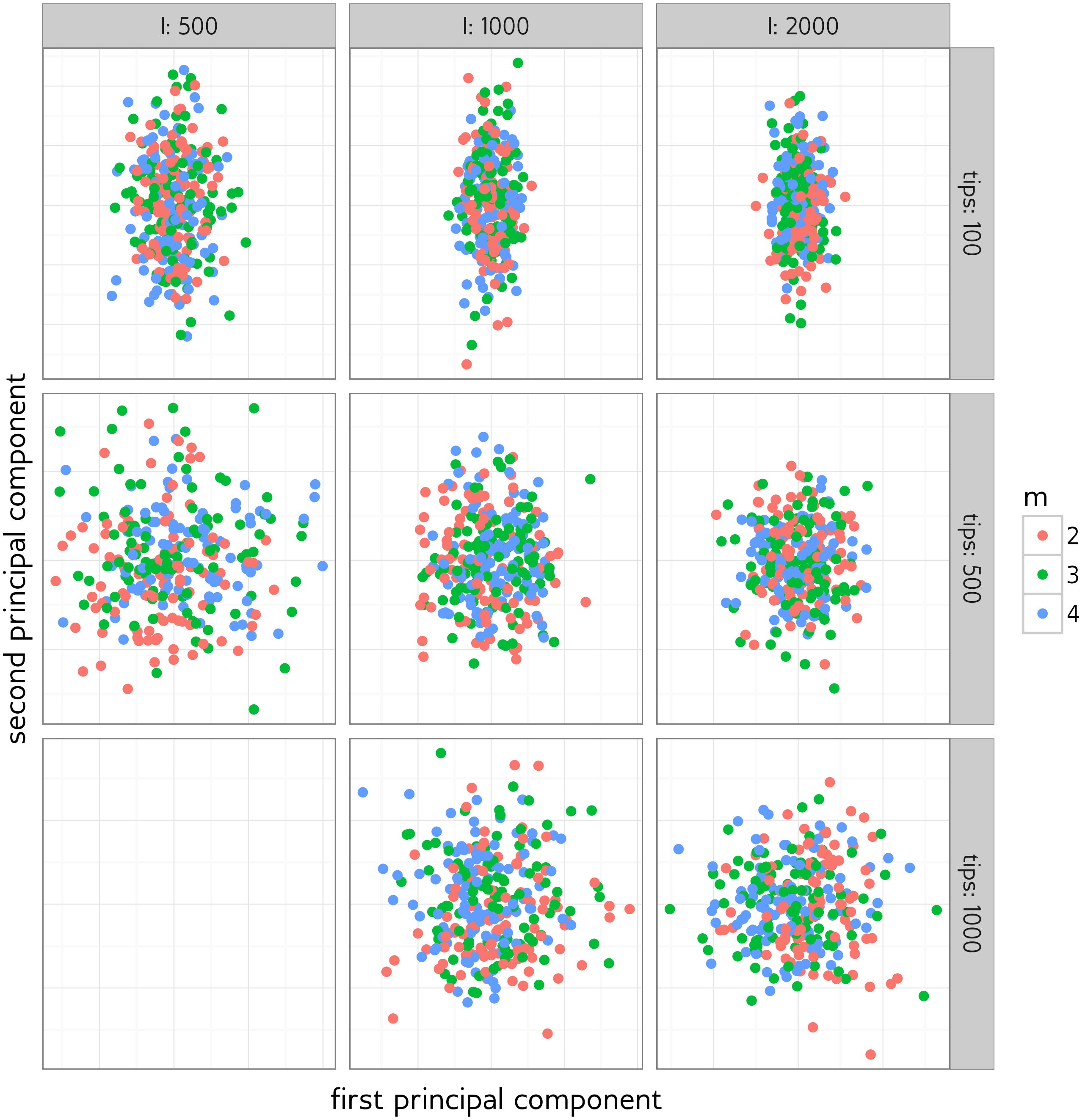

Figure S10: Kernel principal components projection of trees simulated under three different values of BA parameter m, for eight epidemic scenarios.

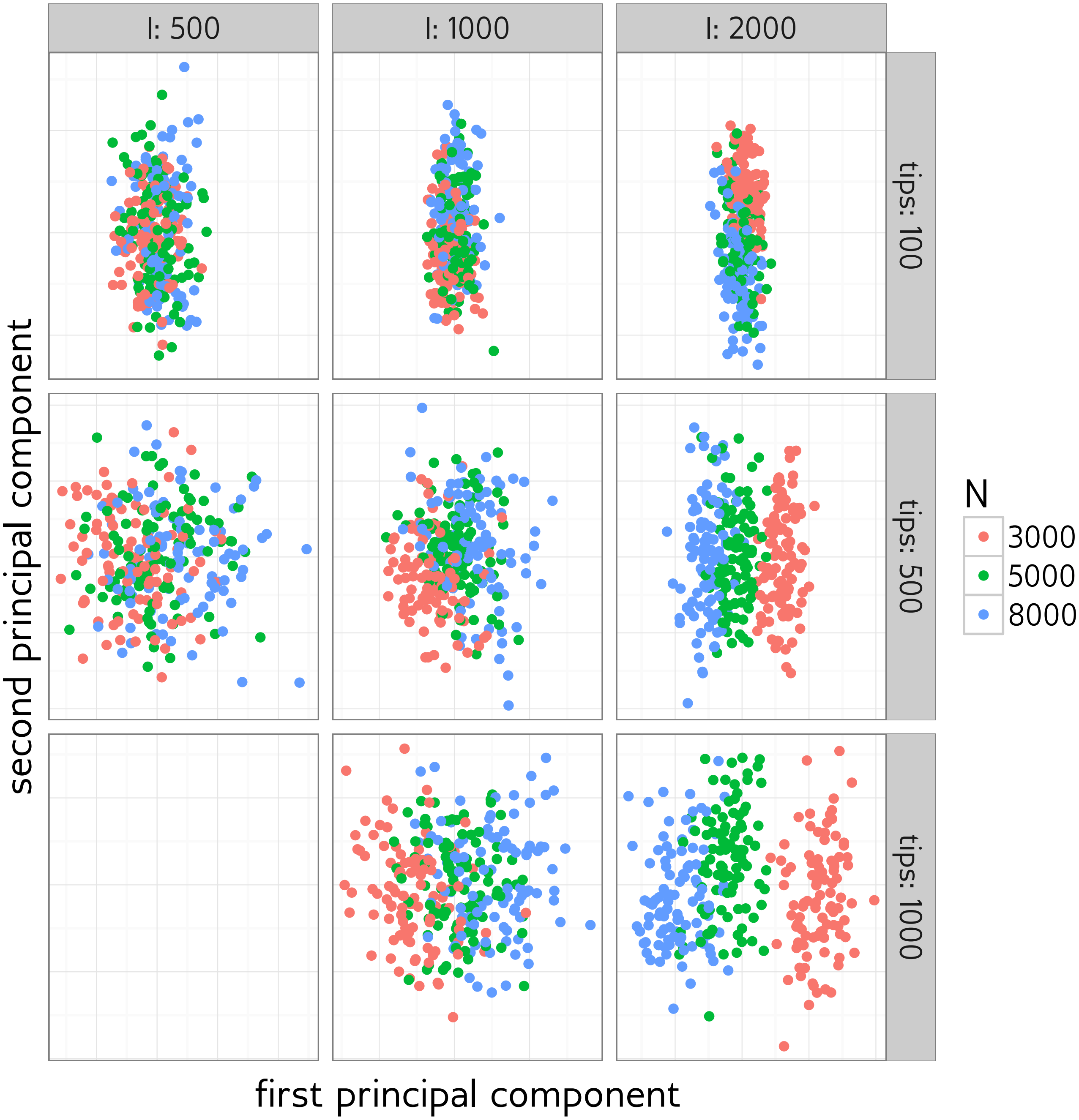

Figure S11: Kernel principal components projection of trees simulated under three different values of BA parameter N, for eight epidemic scenarios.

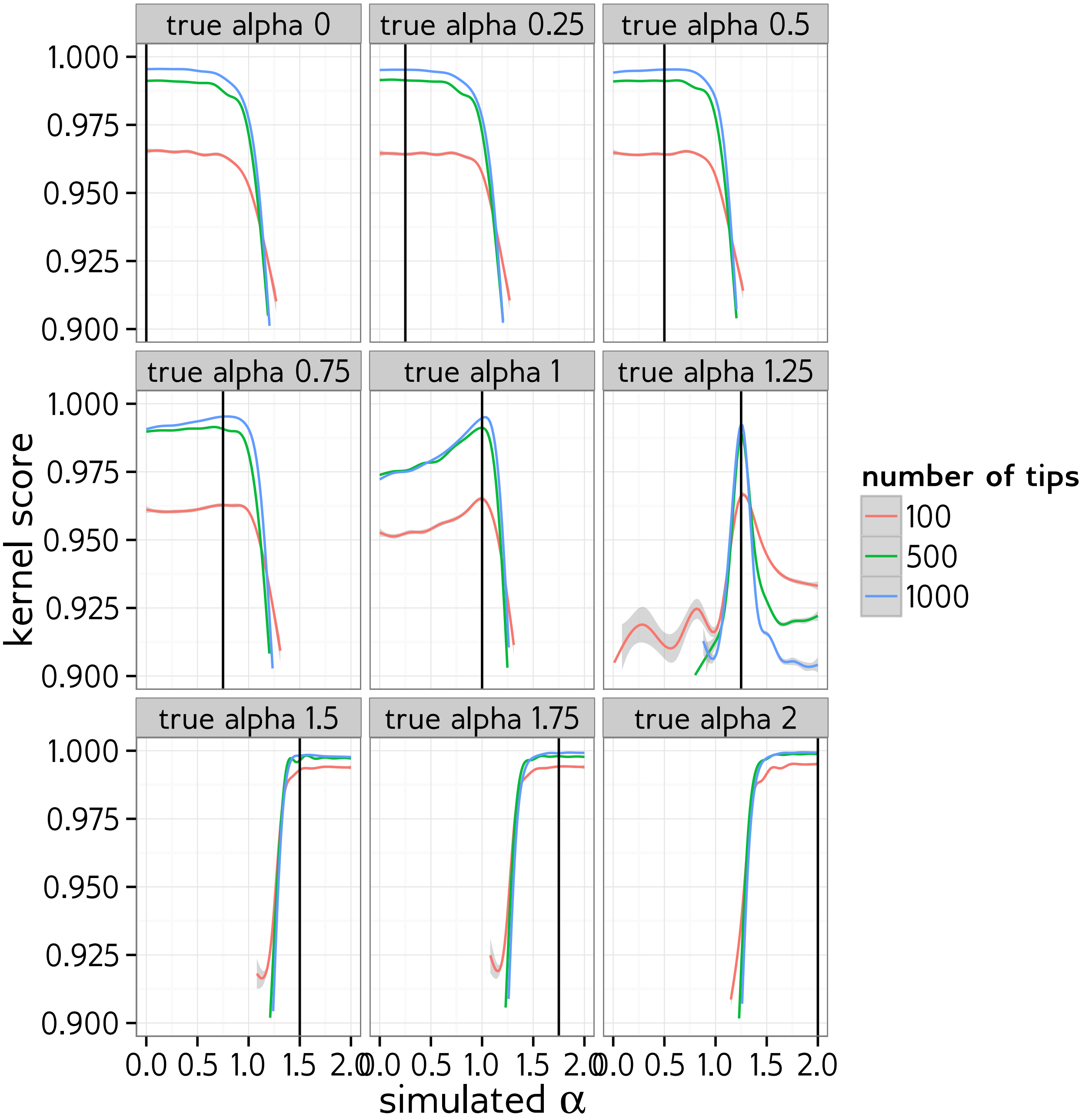

Figure S12: Grid search kernel scores for testing trees simulated under various α values. The other BA parameters were fixed at $I = 1000$, $N = 5000$, and $m = 2$.

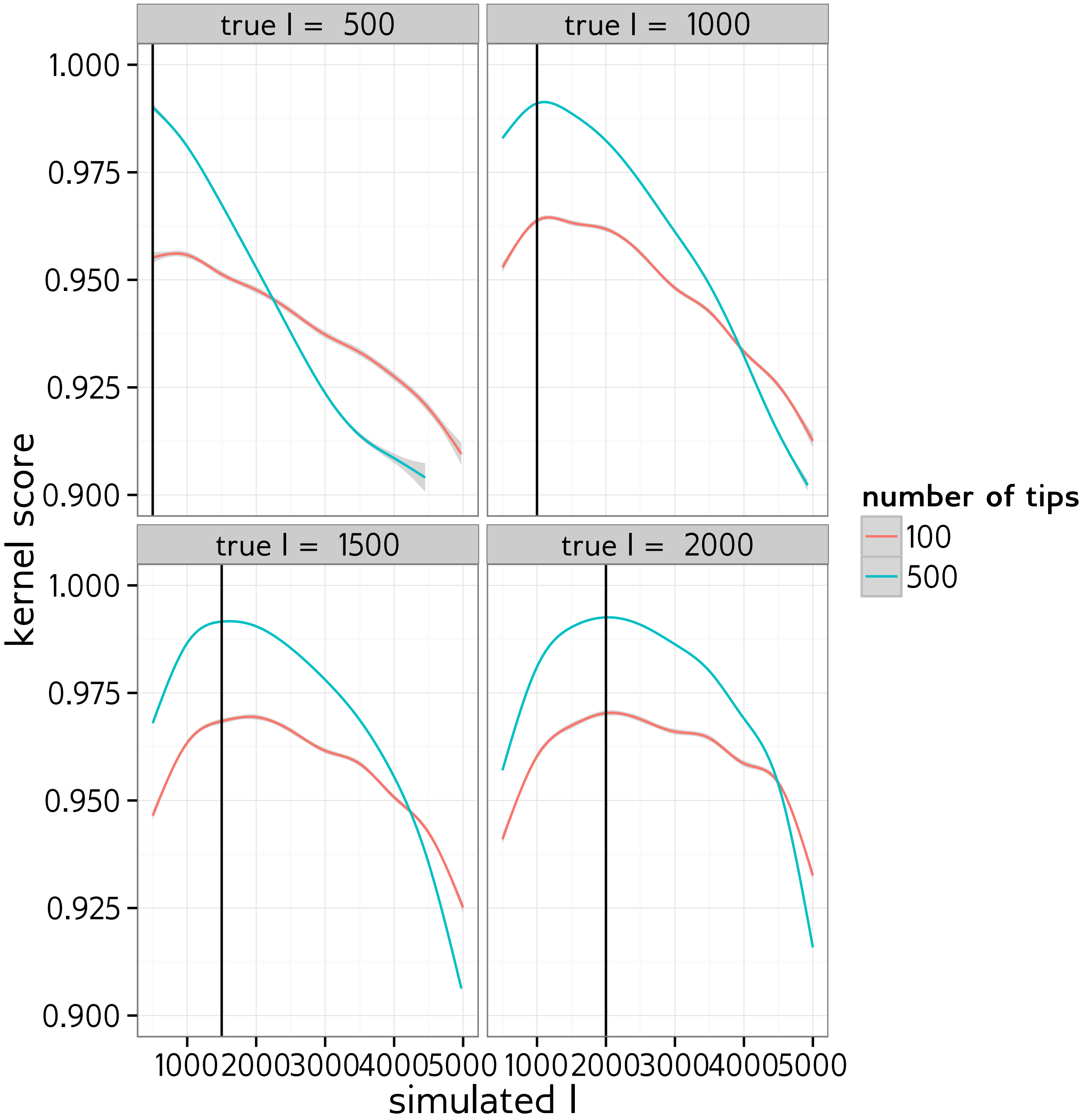

Figure S13: Grid search kernel scores for testing trees simulated under various I values. The other BA parameters were fixed at $\alpha = 1.0$, $N = 5000$, and $m = 2$.

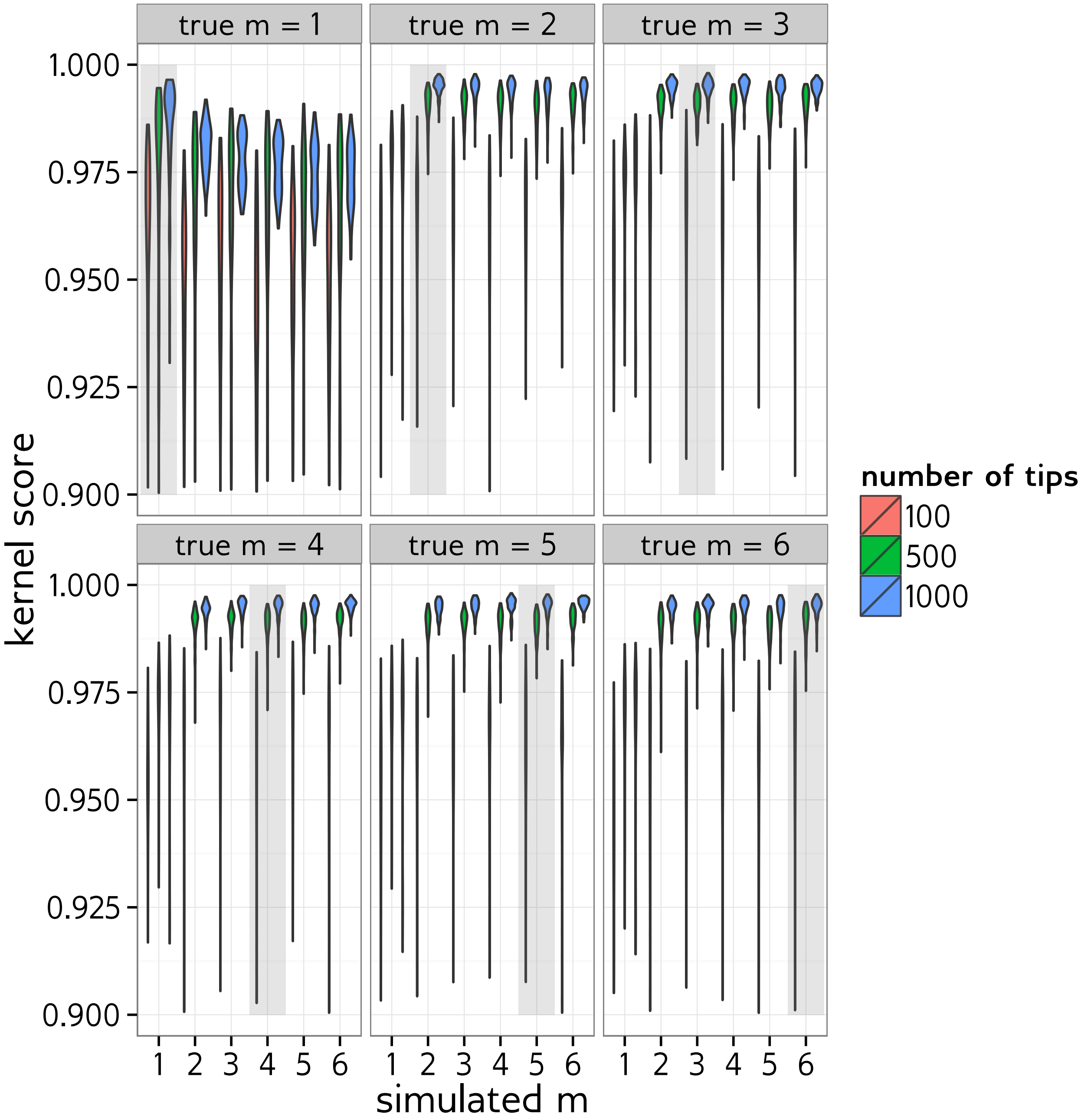

Figure S14: Grid search kernel scores for testing trees simulated under various m values. The other BA parameters were fixed at $\alpha = 1.0$, $I = 1000$, and $N = 5000$.

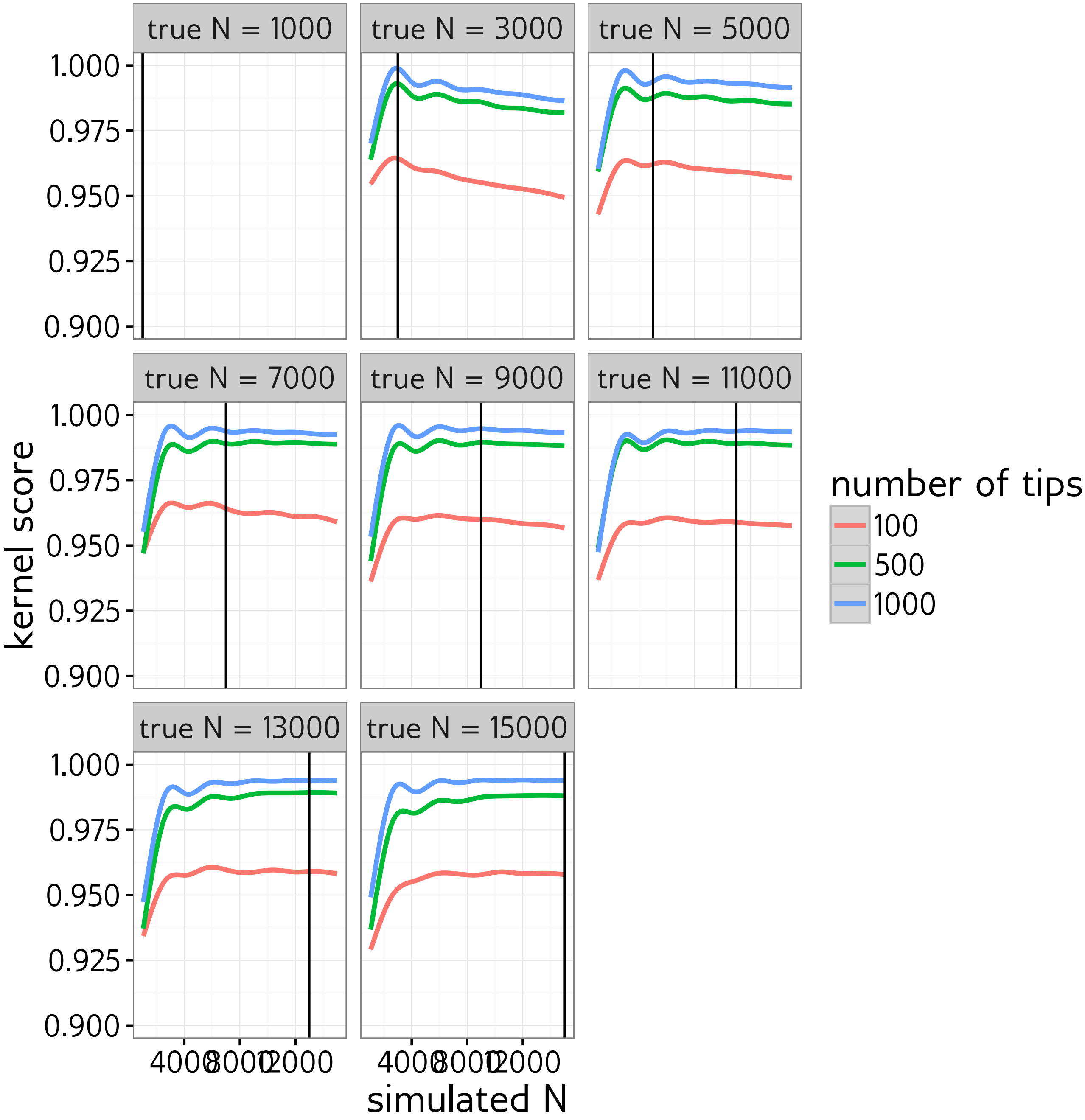

Figure S15: Grid search kernel scores for testing trees simulated under various N values. The other BA parameters were fixed at $\alpha = 1.0$, $I = 1000$, and $m = 2$.

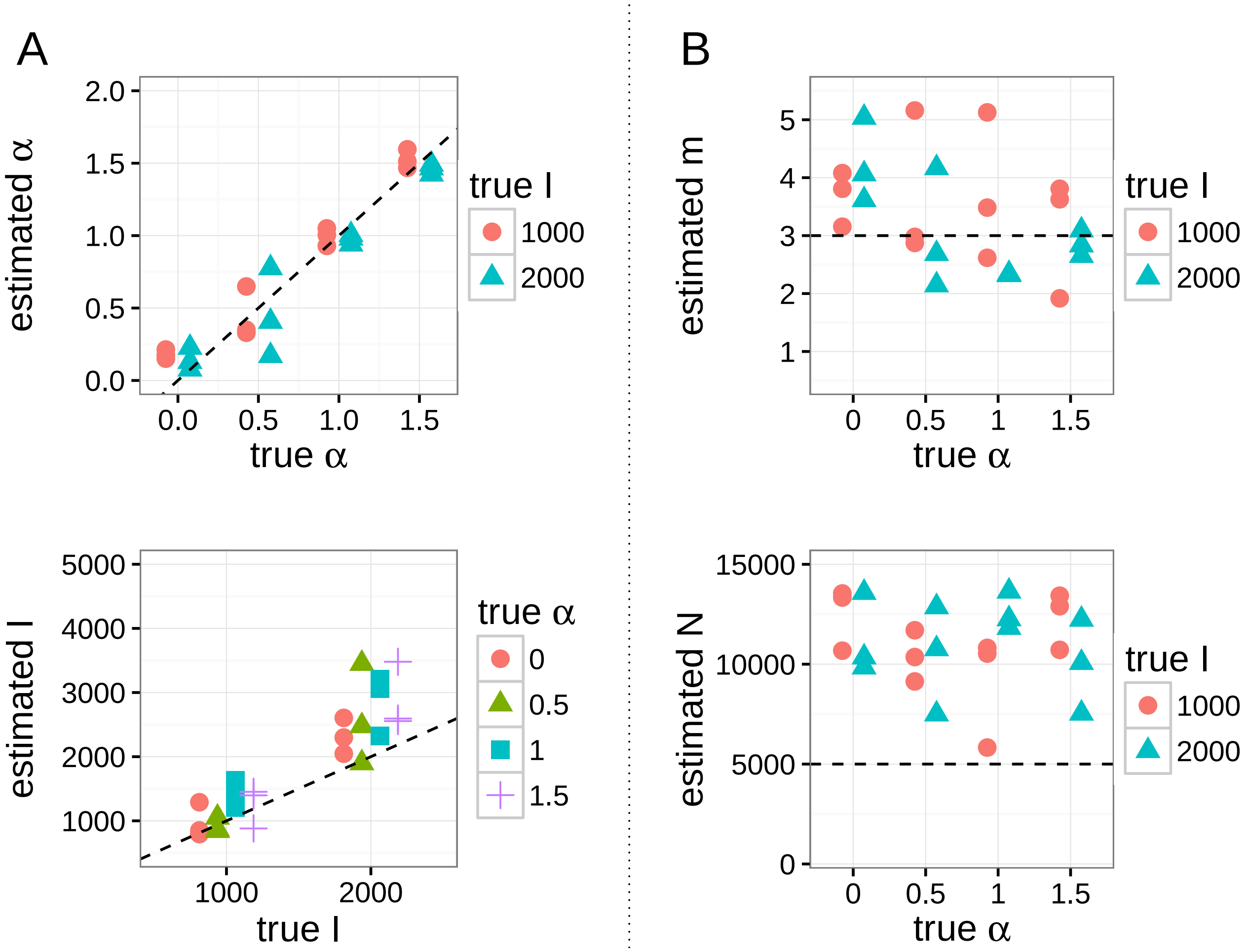

Figure S16: Maximum *a posteriori* point estimates for BA model parameters obtained by running *netabc* on simulated data. Values shown are for simulations with $m = 3$. Dashed lines indicate true values. (A) Estimates of α and I which were varied in these simulations against known values. (B) Estimates of m and N which were held fixed in these simulations at the values $m = 3$ and $N = 5000$.

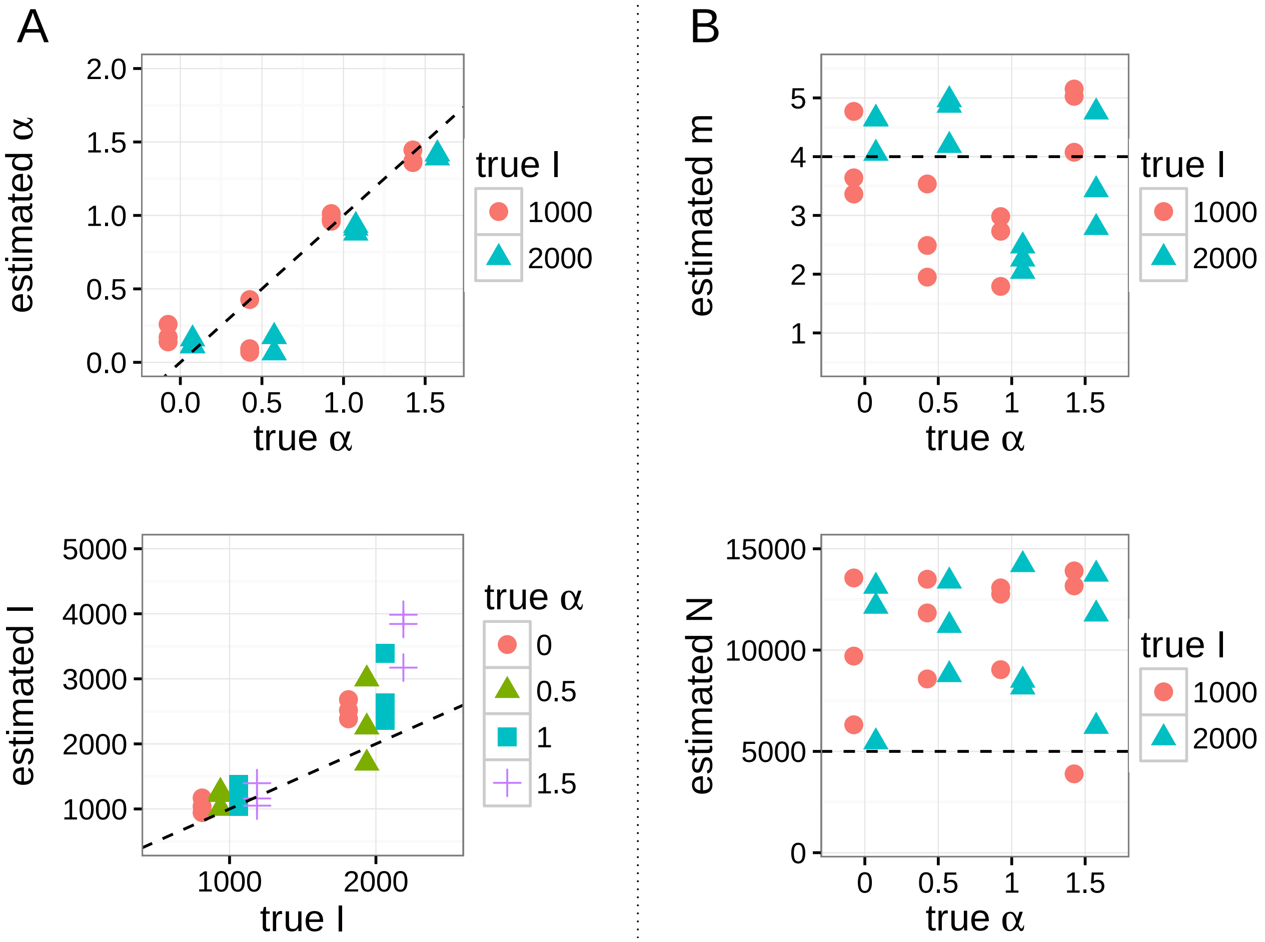

Figure S17: Maximum *a posteriori* point estimates for BA model parameters obtained by running *netabc* on simulated data. Values shown are for simulations with $m = 4$. Dashed lines indicate true values. (A) Estimates of α and I which were varied in these simulations against known values. (B) Estimates of m and N which were held fixed in these simulations at the values $m = 4$ and $N = 5000$.