# PHYLOGENETIC ESTIMATION OF CONTACT NETWORK PARAMETERS WITH 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)

July 2016

# Abstract

Models of the spread of disease in a population often make the simplifying assumption that the population is homogeneously mixed, or is divided into homogeneously mixed compartments. However, human populations have complex structures formed by social contacts, which can have a significant influence on the rate and pattern of epidemic spread. Contact networks capture this structure by explicitly representing each contact that could possibly lead to a transmission. Contact network models parameterize the structure of these networks, but estimating their parameters from contact data requires extensive, often prohibitive, epidemiological investigation.

We developed a method based on approximate Bayesian computation (ABC) for estimating structural parameters of the contact network underlying an observed viral phylogeny. The method combines adaptive sequential Monte Carlo for ABC, Gillespie simulation for propagating epidemics though networks, and a previously developed kernel-based tree similarity score. Our method offers the potential to quantitatively investigate contact network structure from phylogenies derived from viral sequence data, complementing traditional epidemiological methods.

We applied our method to the Barabási-Albert network model. This model incorporates the preferential attachment mechanism observed in real world social and sexual networks, whereby individuals with more connections attract new contacts at an elevated rate ("the rich get richer"). Using simulated data, we found that the strength of preferential attachment and the number of infected nodes could often be accurately estimated. However, the mean degree of the network and the total number of nodes appeared to be weakly- or non-identifiable with ABC.

Finally, the Barabási-Albert model was fit to eleven real world HIV datasets, and substantial heterogeneity in the parameter estimates was observed. Posterior means for the preferential attachment power were all sub-linear, consistent with literature results. We found that the strength of preferential attachment was higher in injection drug user populations, potentially indicating that high-degree "superspreader" nodes may play a role in epidemics among this risk group. Our results underscore the importance of considering contact structures when investigating viral outbreaks.

# Preface

The initial idea to use approximate Bayesian computation (ABC) to infer contact network model parameters was Dr. Poon's, based on his previous work using ABC to infer parameters of population genetic models. 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 was mine, but Dr. Alexandre Bouchard-Côté made me aware of the adaptive version used in this work. Dr. Sarah Otto suggested the experiments involving a network with a heterogeneous $\alpha$ parameter and peer-driven sampling. Dr. Richard Liang provided guidance in the development of the Gillespie simulation algorithm and statistical advice. The *netabc* program, and all supplementary analysis programs, were written by me.

A version of chapter 2 has been submitted for publication with the title "Reconstructing network parameters from viral phylogenies." An oral presentation entitled "Phylodynamic inference of contact network parameters with kernel-ABC" was given based on chapter 2 to the 23rd HIV Dynamics and Evolution meeting on April 25, 2016, in Woods Hole, Massachusetts, USA (the presentation was delivered remotely). A poster based on chapter 2 entitled "Likelihood-free estimation of contact network parameters from viral phylogenies" was presented at the Intelligent Systems for Molecular Biology meeting on July 8, 2016, in Orlando, Florida, USA.

Use of the BC data is in accordance with an ethics application that was reviewed and approved by the UBC/Providence Health Care Research Ethics Board (H07-02559).

Source code for the *netabc* program is freely available at https://github.com/rmcclosk/netabc under the GPL-3 license. Scripts to run all computational experiments, as well as the source code for this thesis, are available at https://github.com/rmcclosk/thesis.

# Table of Contents

# List of Tables

# List of Figures

# List of Symbols

$I$ number of infected nodes in a contact network at the time of transmission tree sampling.

$M$ number of simulated datasets per particle in ABC-SMC.

$N$ total number of nodes in a contact network.

$R_0$ basic reproductive number; average number of infections ultimately caused by one infected individual.

$T$ a transmission tree.

$\Theta$ parameter space for a given mathematical model.

$\alpha_{\text{ESS}}$ per-iteration rate of expected sample size decay in adaptive ABC-SMC.

$\alpha$ preferential attachment power parameter in Barabási-Albert networks.

$\beta$ transmission rate in susceptible-infected and susceptible-infected-removed epidemiological models.

$\gamma$ exponent of power-law degree distribution in scale-free networks.

$\lambda$ decay factor meta-parameter for tree kernel.

$\mathcal{D}$ set of discordant edges in Gillespie simulation.

$\mathcal{I}$ set of infected nodes in Gillespie simulation.

$\mathcal{R}$ set of recovered nodes in Gillespie simulation.

$\mathcal{S}$ set of recovered nodes in Gillespie simulation.

$\nu$ recovery rate in the susceptible-infected-removed epidemiological model.

$\rho$ distance function for approximate Bayesian computation.

$\sigma$ radial basis function variance meta-parameter for tree kernel.

$\varepsilon$ distance function for approximate Bayesian computation.

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

$n$ number of particles used for sequential Monte Carlo.

$q$ proposal function for Metropolis-Hastings kernel.

$t_{\max}$ user-defined cutoff time at which to stop Gillespie simulation.

$t$ time since initial infection of index case in an epidemic.

# List of Abbreviations

**ABC** approximate Bayesian computation.
**AIC** Akaike information criterion.

**BA** Barabási-Albert.

**ER** Erdős-Rényi.
**ERGM** exponential random graph model.
**ESS** expected sample size.

**GLM** generalized linear model.
**GSL** GNU scientific library.
**GTR** generalized time-reversible.

**HDI** highest density interval.
**HIV** human immunodeficiency virus.
**HMM** hidden Markov model.
**HPD** highest posterior density.

**IDU** injection drug users.
**IQR** interquartile range.
**IS** importance sampling.

**kPCA** kernel principal components analysis.
**kSVM** kernel support vector machine.
**kSVR** kernel support vector regression.

**LTT** lineages-through-time.

**MAP** maximum *a posteriori*.
**MCMC** Markov chain Monte Carlo.
**MH** Metropolis-Hastings.
**ML** maximum likelihood.
**MPI** message passing interface.
**MSM** men who have sex with men.

**nLTT** normalized lineages-through-time.

**PA** preferential attachment.
**PCA** principal components analysis.
**pdf** probability density function.
**POSIX** Portable Operating System Interface.

**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

First and foremost I am deeply indebted to my supervisor, Dr. Art Poon, for mentoring me through the majority of my academic career, and somehow always managing to make time for me. My time in his lab has been challenging and rewarding, and I feel well prepared for opportunities to come. I also thank my labmates, especially Jeff, Richard, and Thuy. I'll miss our Thursday lunches.

I am grateful for the guidance of my committee members Dr. Alexandre Bouchard-Côté, Dr. Richard Harrigan, and Dr. Sarah Otto. Thanks also to my rotation supervisors Dr. Sara Mostafavi and Dr. Ryan Morin, who accepted me into their labs and always had time for my many questions.

I owe thanks to my fellow graduate students of the bioinformatics training program, especially Ogan, who helped keep me on an even keel when times got tough.

Finally, I dedicate this thesis to my parents, who have supported me morally and financially, endured my frustrations, and celebrated my successes.

# 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 *mass action* [4], or *panmixia*, implies that any two individuals in the same compartment are equally likely to come into contact making transmission possible at some predefined rate. Although this provides a reasonable approximation in many cases [5], the error introduced by assuming a panmictic population can be substantial when significant contact heterogeneity exists in the underlying population [6–8]. 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 [9, 10]. Phylodynamic methods, which link viruses' evolutionary and epidemiological dynamics, have been used to fit many different types of models to phylogenetic data [11, 12]. However, these models generally assume a panmictic population. The primary objective of this work is *to develop a method to fit contact network models, and thereby relax the assumption of homogeneous mixing, in a phylodynamic framework.*

In this work, we take a Bayesian approach: our goal is to estimate the posterior distribution on model parameters given our data,

$$\pi(\theta \mid T) = \frac{f(T \mid \theta)\pi(\theta)}{\int_\Theta f(T \mid \theta)\pi(\theta)\mathrm{d}\,\theta}, \tag{1.1}$$

where $f(T \mid \theta)$ is the likelihood of the parameters given $T$, $\pi(\theta)$ is the prior on $\theta$, and $\Theta$ is the space of possible model parameters. The denominator on the right-hand side is the marginal probability of $T$ which acts as a normalizing constant on the posterior (see appendix A for a review of mathematical modelling and Bayesian inference, including definitions of these concepts). As we shall show (section 2.1.4), estimating this distribution presents computational challenges beyond those usually encountered in Bayesian inference. Both the likelihood $f(T \mid \theta)$ and the normalizing constant seem to be

intractable, which rules out the use of most common maximum likelihood and Bayesian methods.

An alternative to likelihood-based methods, which could be applied to any network model, is provided by approximate Bayesian computation (ABC) [13–16]. 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. Those shapes can be compared using a highly informative similarity measure called the "tree kernel" [17]; similar kernel functions have been demonstrated to work well as distance functions in ABC [18]. ABC can be implemented with several algorithms, but sequential Monte Carlo (SMC) has advantages over others, including improved accuracy in low-density regions and parallelizability [19]. A recently-developed adaptive algorithm requiring minimal tuning on the part of the user makes SMC an even more attractive approach [20]. 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. Our method will expand on the framework developed by [21], who combined ABC with the tree kernel to infer parameters of population genetic models from viral phylogenies using Markov chain Monte Carlo (MCMC).

Empirical studies of sexual contact networks have found that these networks tend to be scale-free [22–25], meaning that their degree distributions follow a power law (although there has been some disagreement, see [6, 26]). Preferential attachment has been postulated as a mechanism by which scale-free networks could be generated [27]. The Barabási-Albert (BA) model [27] is one of the simplest preferential attachment models, which makes it a natural choice to explore with our method. The second aim of this work is *to use simulations to investigate the parameters of the Barabási-Albert model, including whether they have a detectable impact on tree shape, and whether they can be accurately recovered using ABC.*

Due to its high global prevalence and fast mutation rate, 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 settings. At the time of this writing, there were $635,400$ HIV sequences publicly available in GenBank, annotated with 172 distinct countries of origin. Since HIV is almost always spread through either sexual contact or sharing of injection drug supplies, the contact networks underlying HIV epidemics are driven by social dynamics and are therefore likely to be highly structured [25]. 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 [28, 29]. 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 ABC to fit the Barabási-Albert model to existing HIV outbreaks*.

To summarize, this work has three objectives. First, we will develop a method which uses 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 ABC. Finally, we will apply the method to fit the BA model to several real-world HIV datasets.

The remainder of this background chapter is organized in four sections. The first section introduces phylogenies and transmission trees, which are the input data from which our method aims to make statistical inferences. This section also introduces phylodynamics, a family of methods that, like ours, aim to infer epidemiological parameters from evolutionary data. The second section focuses on contact networks and network models, whose parameters we are attempting to infer. The relationship between contact networks and transmission trees is also discussed. The third and fourth sections introduce SMC and ABC respectively, which are the two algorithmic components of the method we will implement. In particular, ABC refers to the general approach of using simulations to replace likelihood calculations in a Bayesian setting, while SMC is a particular algorithm which can be used to implement ABC.

## 1.2 Phylogenetics and phylodynamics

### 1.2.1 Phylogenetic trees

In evolutionary biology, a *phylogeny*, or *phylogenetic tree*, is a graphical representation of the evolutionary relationships among a group of organisms or species (generally, *taxa*) [30]. The *tips* of a phylogeny, that is, the nodes without any descendants, correspond to *extant*, or observed, taxa. The *internal nodes* correspond to their common ancestors, usually extinct (although occasionally the internal nodes may be observed as well, *e.g.*[31]). The edges or *branches* of the phylogeny connect ancestors to their descendants. Phylogenies may have a *root*, which is a node with no parent distinguished as the most recent common ancestor of all the extant taxa [32]. 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* [33].

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 and 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, reference points [34]. 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 [35]. These concepts are illustrated in fig. 1.1.

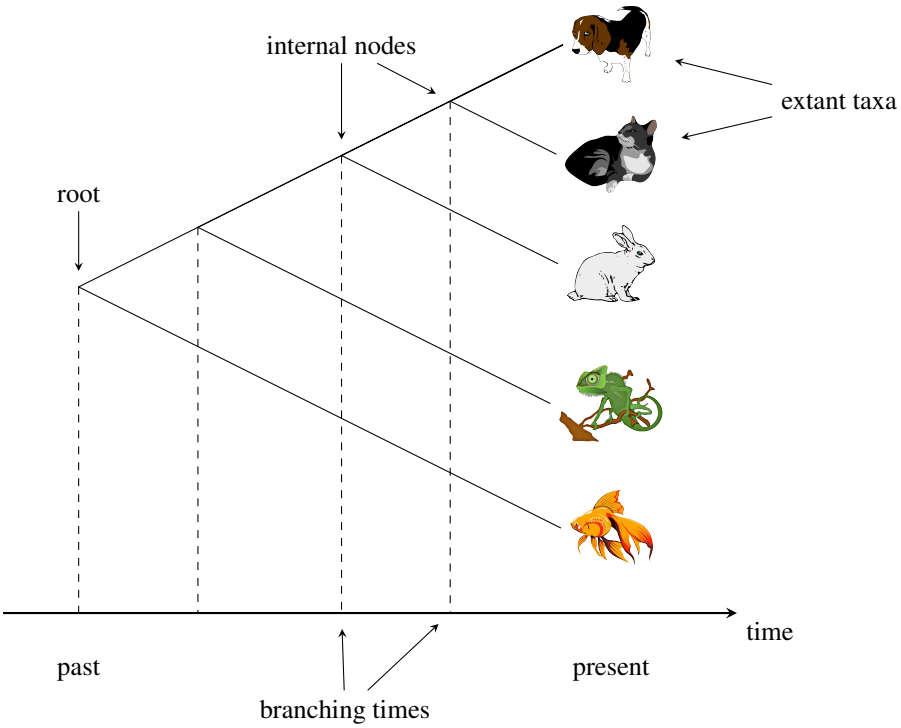


Figure 1.1: Illustration of a rooted, ultrametric, time-scaled phylogeny. The tips of the tree, which represent extant taxa, are placed at the present day on the time axis. Internal nodes, representing extinct common ancestors to the extant taxa, fall in the past. The topology of the tree indicates that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

### 1.2.2 Transmission trees

In epidemiology, a *transmission tree* is a graphical representation of an epidemic's progress through a population [36]. Like phylogenies, transmission trees have tips, nodes, edges, and branch lengths. However, rather than recording an evolutionary process (speciation), they record an epidemiological process (transmission). The tips of a transmission tree represent the removal by sampling 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 [37]. 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 [38], in contrast to the *isochronous* taxa found in most phylogenies of macro-organisms. A transmission tree is illustrated in fig. 1.2 (right). The object on the right of the figure is called a *contact network*, which depicts the entire susceptible population along with all possible routes of disease transmission. Contact networks, and their relationships to transmission trees, will be

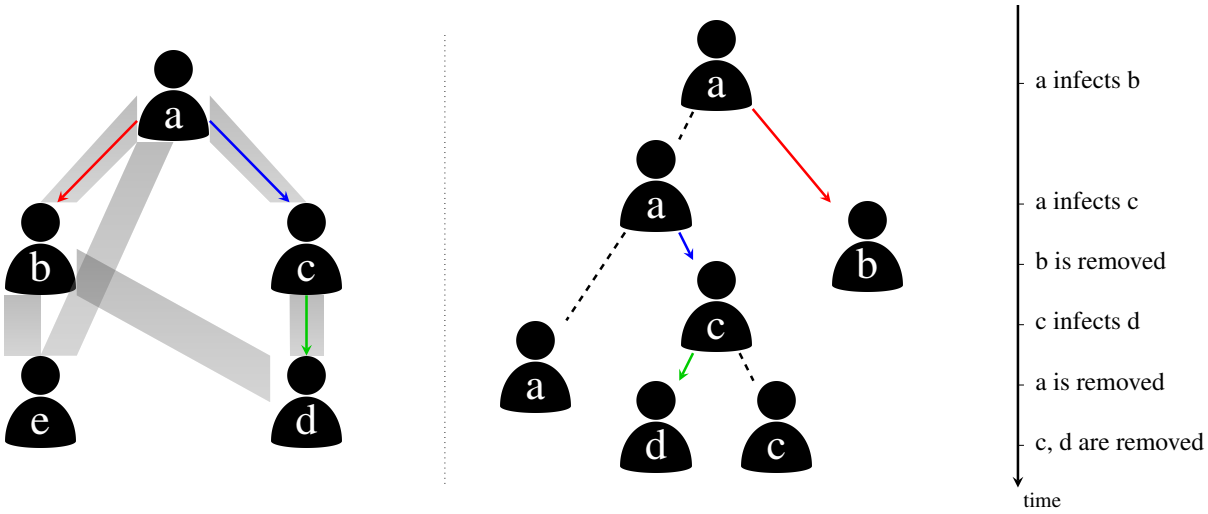


Figure 1.2: Illustration of epidemic spread over a contact network, and the corresponding transmission tree. (Left) A contact network with five hosts, labelled *a* through *e*. Thick shaded edges indicate symmetric contacts among the hosts. The transmission network is indicated by coloured arrows. The 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.

discussed further in section 1.3.

Each infected individual in an epidemic may appear at nodes of 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 [8, 39]. 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 trees in the graph theoretical sense, and hence are sometimes also referred to as transmission trees [*e.g.* 40]. In this work, we reserve the term "transmission tree" for the objects depicted on the right side of fig. 1.2, following *e.g.* [37]. The term "transmission network" is taken to mean the subgraph of the contact network along which transmissions occurred, following *e.g.* [8, 39].

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 [34], 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 [41]. However, in all but the most well-studied of epidemics, transmission trees are not possible to assemble through traditional epidemiological methods [39]. The time and effort to conduct detailed interviews and contact tracing of a sufficient number of infected individuals is usually prohibitive, and may additionally be confounded by misreporting and other challenges [42]. 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* [43] addresses the challenge of estimating epidemiological parameters from viral sequence data [12].

### 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 [38]. 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 [11, 12], 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 [11, 12]. 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 [44, 45] or coalescent [46, 47] models. These methods either assume the viral phylogeny is known, as we do in this work, or (more commonly) integrate over phylogenetic uncertainty in a Bayesian framework. Phylogenetic inference is a complex topic which we shall not discuss here; see *e.g.* [48] for a full review.

Due to the relationship between the aforementioned processes, there is a degree of correspondence between viral phylogenies and transmission trees [36, 40, 49, 50]. In particular, the transmission process is quite similar to *allopatric speciation* [51], 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 [52]. 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 [53]. Modern likelihood-based methods of phylogenetic reconstruction [*e.g.* 54, 55] 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 [11]. 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 [56–58], which we apply in this work, and least-square dating [59]. Alternatively, the tree may be rooted separately with an outgroup [60] 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 far from exact [36, 61]. 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 in one order, but the transmitted lineages originated within the donor in a different order. 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. [62] and Paraskevis et al. [63] were able to accurately recover known transmission trees using viral phylogenies. The problem of accurately estimating transmission trees is an ongoing area of research [31, 53, 64–67]. For example, Hall, Woolhouse, and Rambaut [53] developed a Bayesian method to jointly estimate a transmission tree and viral phylogeny by combining models of agent-based transmission, within-host

population dynamics, and sequence evolution.

### 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 [68]. 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 [69], which measures the imbalance or asymmetry in a rooted tree. For the $i$th 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 [70] derive the expected value of Sackin' index under the Yule model [71], under which each lineage has an equal probability of splitting (creating a branching in the tree) at all times. 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 normalized lineages-through-time (nLTT) [72], which compares the LTT [34] 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 [41], so large values may indicate that the trees being compared were produced by different epidemic trajectories [72].

Poon et al. [17] developed an alternative pairwise measure which applies the concept of a *kernel function* to phylogenies. Kernel functions, whose best known use is in support vector machines (SVMs) [73], compare objects in a space $\mathcal{X}$ by mapping them into a feature space $\mathcal{F}$ of high or infinite dimension via a function $\varphi$. 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 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 [17] will henceforth be referred to as the *tree kernel*. This kernel maps trees into 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 [74] and did not incorporate branch length information. The version developed by Poon et al. [17] 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 $T_2$. Following Poon et al. [17], let $N(T)$ denote the set of all nodes in $T$, $\mathrm{nc}(n)$ be the number of children of node $n$, $c_n^j$ be the $j$th child of node $n$, and $l_n$ be the ordered vector of branch lengths connecting node $n$ to its $\mathrm{nc}(n)$ children. Furthermore, let $\mathrm{nl}(n)$ be the number of children of $n$ which are leaves (we always have $\mathrm{nl}(n) \leq \mathrm{nc}(n)$). The *production rule* of $n$ is the pair $(\mathrm{nc}(n), \mathrm{nl}(n))$. 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 Euclidean norm and $\sigma$ is a variance parameter. The tree kernel is defined as

$$K(T_1,T_2) = \sum_{n_1 \in N(T_1)} \sum_{n_2 \in N(T_2)} \Delta(n_1,n_2), \tag{1.2}$$

where

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

The parameter $\lambda$ in the above expression, is called the *decay factor [75], and takes a value between 0 and 1. Without this parameter, terms in the sum 1.2 corresponding to large subset trees with the same topology would be similarly large and tend to dominate the kernel score.* $\lambda$ *penalizes* $\Delta$ *more strongly as the number of recursive calls increases, which downweights the largest matching substructures and allows smaller matches to contribute more to the kernel score.* In this work, we refer to the parameters $\lambda$ and $\sigma$ as *meta-parameters*, to avoid confusing them with model parameters we are trying to estimate. When evaluating the tree kernel, it is helpful to reorder the children of each internal node such that the larger of the two subtrees is on the right-hand side. If the two subtrees have equal sizes, then the child with the longer branch length can be put on the right-hand side. This operation is referred to as *ladderizing*. Since the ordering of children is arbitrary in phylogenies, this operation ensures that a maximal number of matching subset trees are counted by the tree kernel without making meaningful changes to the trees.

The tree kernel was later shown to be highly effective in differentiating trees simulated under a compartmental model with two risk groups of varying contact rates [21]. In that paper, Poon used the tree kernel as the distance function in approximate Bayesian computation (ABC) (see section 1.5), to fit epidemiological models to observed trees.

## 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 human immunodeficiency virus (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 [8, 76, 77]. 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* [78, 79], 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 80].

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 would use directed edges, since the recipient has no chance of transmitting to the donor. Directed edges are also useful when the transmission risk is not equal between the hosts, such as with HIV transmission among men who have sex with men (MSM), where the receptive partner carries a higher risk of infection than the insertive partner [81]. 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 [80]. 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 *in-degree* and *out-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 [5], 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 [82]. However, these models still assume that every possible contact between a member of group $i$ and a member of group $j$ is equally likely. Real human communuties are not necessarily homogeneously mixed; therefore, this assumption can lead to errors in predicted epidemic trajectories when there is substantial heterogeneity present [6, 83, 84]. 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 [5]. Using this assumption, researchers have been able to derive estimates of the transmission rate and the basic repro-

ductive number of various outbreaks, which have agreed with values obtained by on-the-ground data collection [85]. 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 [80].

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.* 7, 86]. 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 [87]. 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 [5, 88]. Hence, the contact network can provide an idea of what to expect as an epidemic unfolds. In terms of actionable information, the efficacy of different vaccination strategies may depend on the topology of the network [8–10, 89]. On a local level, contact networks can be informative about the groups or individuals who are at highest risk of acquiring or transmitting infection who would therefore benefit most from public health interventions [28, 29].

Contact networks are a challenging type of data to collect, requiring extensive epidemiological investigation in the form of contact tracing [8, 39, 42, 77]. 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 [80]. One of the most well-studied of these is the degree distribution mentioned above, which can theoretically be estimated by simply asking each person how many contacts they had in some interval of time. However, the degree distributions often observed in real-world sexual networks are heavy-tailed [22–24], so dense or respondent-driven sampling [90] would be needed to capture the high-degree nodes characterizing the tail of the distribution.

An alternative approach has been the analysis of other types of network, which can be directly 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 [91]. However, this network may show little to no agreement with contact data obtained through epidemiological methods [92–94] and therefore may be a poor proxy for the contact network. Other studies [95] have investigated the *transmission network*, which is the subgraph of the contact network consisting of infected nodes and the edges that led to their infections [39] (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 [95]. These studies again mostly focus

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 [96] 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 [97] 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 [86] 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 [83] 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 $\gamma$ [27]. 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 [22–25]. Interestingly, many other types of network, including computer networks [88], biological metabolic networks [98], and academic co-author networks [99], 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* [100], which have been postulated to contribute in varying degree to the spread of diseases such as HIV [37] and severe acute respiratory syndrome (SARS) [101]. Scale-free networks have no epidemic threshold [88], meaning that diseases with arbitrarily low transmissibility have a chance, however small, of persisting 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 [5].

One mechanism which has been shown to lead to scale-free networks is *preferential attachment* [27, 102]. The simplest preferential attachment model is known as the Barabási-Albert (BA) model after its inventors [27]. 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 plus one.

There has been some contention over the idea that contact networks are scale-free. Handcock and

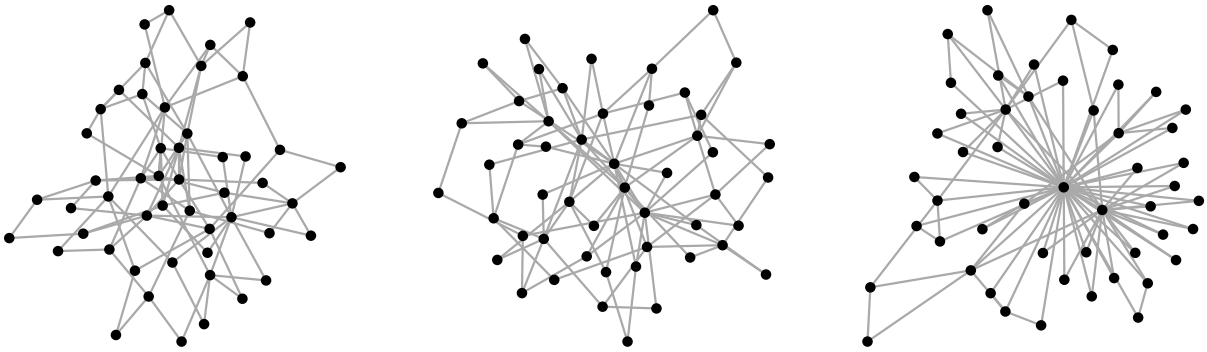

Figure 1.3: Examples of Barabási-Albert networks with preferential attachment power $\alpha$ = 0 (left), 1 (centre), and 2 (right). All networks have $N = 50$ nodes and were constructed with $m = 2$ edges per vertex. When $\alpha = 0$, attachments are formed at random and most nodes have low degree. When $\alpha = 1$, preferential attachment is linear and several higher-degree nodes are observable. When $\alpha = 2$, preferential attachment is quadratic and nearly every vertex is attached to a small number of hub nodes.

Jones [26] 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 [6] found that an exponential distribution, rather than a power law, was the best fit to degree distributions of six social and sexual networks. Dombrowski et al. [103] contend that sexual networks are shaped more by homophily ("like attract like") than by preferential attachment, but find that injection drug users (IDU) network do demonstrate a scale-free structure.

In the paper describing the BA model, Barabási and Albert suggest an extension where the probability of choosing a partner of degree $d$ is proportional to $d^{\alpha} + 1$ for some constant $\alpha$. When $\alpha \neq 1$, the degree distribution no longer follows a power law [104]. For $\alpha < 1$, the distribution is a stretched exponential, meaning that the number of nodes of degree $k$ is proportional to $\exp(-k^{\beta})$ for some constant $\beta$. For $\alpha > 1$, the distribution takes on a characteristic called *gelation*, where a one or a few high-degree hub nodes are connected to nearly every other node in the graph. We do not believe these departures from the power law affect the applicability of the model to real world networks. In fact, de Blasio, Svensson, and Liljeros [105] were able to estimate the preferential attachment power from partner count data collected from the same individuals for consecutive time intervals, and found a value less than one in all cases. It is also worth noting that, in addition to the BA model, other investigations of the interaction between contact networks and transmission trees have studied the Erdős-Rényi and Watts-Strogatz models [106], whose degree distributions do not generally follow a power law under any parameter settings.

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 other $m$ values. Examples of BA networks with three different values of the preferential attachment power $\alpha$ are shown in fig. 1.3.

### 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 is impossible to distinguish who transmitted to whom from a transmission tree alone [107].

A number of studies have made progress in quantifying the relationship between contact networks and transmission trees. O'Dea and Wilke [108] simulated epidemics over networks with four types of degree distribution. They then estimated the Bayesian skyride [109] population size trajectory in two ways: from the phylogeny, using MCMC; and from the incidence and prevalence trajectories, using the method developed by Volz et al. [52]. 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. [106] 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. The authors also investigated a real-world HIV phylogeny and found a level of imbalance inconsistent with a randomly mixing population. Welch [110] 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. [111] investigated the correspondence between contact network clusters and transmission tree clusters and did find a moderate correspondence between the two in some cases. Goodreau [112] combined a dynamic contact network model with a model of within-host viral evolution to simulate viral phylogenies over eight types of contact network. Estimates of prevalence and effective population size were calculated for each simulated phylogeny under three models of epidemic growth. The author found that estimates for networks with a small high-risk subgroup and networks involving commercial sex workers were substantially different than estimates for random networks or networks with segregated equal-risk groups.

## 1.4 Sequential Monte Carlo

### 1.4.1 Overview and notation

Recall that the primary objective of our work is to develop a statistical inference method for estimating contact network parameters from transmission trees. For a network model with parameters $\theta$ and an input transmission tree $T$, our goal is to estimate statistics, such as means and credible intervals, of the

posterior distribution (1.1). As alluded to in section 1.1, both the likelihood $f(T \mid \theta)$ and the normalizing constant are likely computationally intractable (this will be discussed further in section 2.1.4). Hence, rather than computing the posterior distribution analytically, we will approximate it using a *Monte Carlo* approach. The fundamental idea behind Monte Carlo methods is succinctly expressed by Liu, Chen, and Logvinenko [113]:

> Monte Carlo's view of the world is that any probability distribution $\pi$, regardless of its complexity, can always be *represented* by a discrete sample from it. By "represented", we mean that any computation of expectations using $\pi$ can be replaced to an acceptable degree of accuracy by using the empirical distribution resulting from the discrete sample.

In other words, if we are able to sample enough points from a distribution of interest, we will be able to make reasonably accurate statements about the distribution itself. For example, the expected value of the distribution can be estimated by the sample's population mean. The reason Monte Carlo methods will be useful in this work is that algorithms exist for obtaining samples from distributions that are analytically intractable and from which direct sampling is not possible (for a review see [114]). Sequential Monte Carlo (SMC) [115–117] is one such algorithm.

SMC considers a population of points or "particles", here denoted $\left\{x^{(k)}\right\}$ and indexed by an integer $k$. The particles are associated with weights, $\left\{w(x^{(k)})\right\}$ representing their probability density. After the algorithm has been run, the weighted particles are a Monte Carlo representation for the target distribution. For example, the expected value is approximated by

$$\frac{1}{n}\sum_{k=1}^{n} x^{(k)} w(x^{(k)}),$$

where $n$ is the number of particles. The weights of the whole population always sum to 1 so that their weighted histogram defines a true probability density function integrating to 1. Initially, the particles do not represent the target distribution but rather a more tractable distribution from which direct sampling is straightforward. The word "sequential" is used to describe the iterative process of perturbation, resampling, and reweighting applied to the particles in such a way that they converge, collectively, to a representation of the target. SMC is also known as the *particle filter*.

In this work, the distribution of interest is the posterior distribution **??**. The particles are particular values of the parameters $\theta$ of the contact network model being studied. If we were taking a typical Bayesian Monte Carlo approach to this problem, the particles would end up weighted by their posterior probability and distributed in such a way that the weighted population was a reasonable representation of $\pi(\theta \mid T)$. In our case, due to the intractable likelihood, we will need to consider an approximation to the posterior (see section 1.5). However, for now, nothing is lost by assuming that our target distribution is the posterior itself.

In this section, we review an algorithm called the SMC sampler, developed by Del Moral, Doucet, and Jasra [118], which forms the basis of the adaptive ABC-SMC algorithm we apply toward the main objective of this work. We begin by describing sequential importance sampling (SIS), which is a precursor to SMC that samples from a sequence of distributions defined on spaces of increasing dimension.

We then describe SMC itself, which extends SIS with a resampling step to fight particle degeneracy. Finally, we outline the SMC sampler, which allows SMC to be applied to sequences of distributions all defined on the same space. This terminology will become clear as the methods are described.

To fully describe SMC, we will introduce some notation and terminology. For a sequence $x_1, \ldots, x_d$, we will write $\mathbf{x_i}$ to mean the partial sequence $x_1, \ldots, x_i$. The subscript $^{(k)}$ will be used to indicate the $k$th 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

$$\int_Y K(x,y)\mathrm{d}y = 1 \tag{1.3}$$

for all $x \in X$. This is an "operational" definition of Markov kernel which will be suitable for our purposes. A more rigorous definition can be found in *e.g.* [119]. Note that 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). Also, the variable $\pi$ refers to a generic distribution in the following descriptions of SMC and related algorithms, and does not refer to the prior or posterior distributions we will eventually want to work with.

### 1.4.2 Sequential importance sampling

Sequential importance sampling (SIS) [120] is a particle-based method whose aim is to sample from a distribution $\pi$ 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 $\pi$ 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 $\pi$ 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 $\eta$, called the *importance distribution*, from which we are able to sample. Define the *importance weight* as the ratio $w(x) = \gamma(x)/\eta(x)$. We can write the expectation of interest as

$$\int f(x)\pi(x)\mathrm{d}x = \frac{1}{Z}\int w(x)\eta(x)f(x)\mathrm{d}x. \tag{1.4}$$

Since $\eta$ can be sampled from exactly, and $\gamma$ and $f$ can both be evaluated pointwise, the integral $\int w(x)\eta(x)f(x)\mathrm{d}x$ can be approximated by a Monte Carlo estimate. We simply take a sample from $\eta$, multiply each point in the sample by $f$ and $\gamma$ evaluated at that point, and sum the results. Moreover, the normalizing constant $Z$ can be expressed in terms of the importance weight and distribution, $Z = \int w(x)\eta(x)\mathrm{d}x$. Therefore, we have all the ingredients we need to obtain an estimate of $\pi(f)$ using eq. (1.4). Although this is a simple and elegant approach, the drawback is that the variance of the esti-

mate is proportional to the variance of the importance weights [117], which may be quite large if $\eta$ and $\gamma$ are very different. Therefore, the practical use of IS on its own is limited, since it depends on finding an importance distribution similar to $\pi$, which we usually know very little about *a priori*.

The objective of SIS is to build up an importance distribution $\eta$ for $\pi$ 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 $\eta$ for $\pi$ 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 $\eta$ can be written recursively as

$$\begin{aligned} w(\mathbf{x_i}) &= \frac{\pi(\mathbf{x_i})}{\eta(\mathbf{x_i})} && \text{definition of importance weight} \\ &= \frac{\pi(x_i \mid \mathbf{x_{i-1}})\pi(\mathbf{x_{i-1}})}{\eta(x_i \mid \mathbf{x_{i-1}})\eta(\mathbf{x_{i-1}})} && \text{definition of conditional probability} \\ &= \frac{\pi(x_i \mid \mathbf{x_{i-1}})}{\eta(x_i \mid \mathbf{x_{i-1}})} \cdot w(\mathbf{x_{i-1}}) && \text{definition of importance weight.} \end{aligned} \tag{1.5}$$

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.* [117, 121] 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 $\eta(x_i \mid \mathbf{x_{i-1}})$ which minimizes the variance of the weights is called the *optimal kernel* [122], the name deriving from the fact that $k(x_i, \mathbf{x_{i-1}}) = \eta(x_i \mid \mathbf{x_{i-1}})$ is a Markov kernel. In some applications, it is possible to approximate the optimal kernel or even compute it explicitly.

The recursive definition 1.5 suggests an algorithm for obtaining a sample from $\eta$ and using it to obtain an approximate sample from $\pi$ by IS (algorithm 1). We begin with $n$ particles, which have been sampled from the importance distribution $\eta(x_1)$ for $\pi(x_1)$. Each particle is extended by drawing $x_2$ from $\eta(x_2 \mid x_1)$, and the importance weights are updated by eq. (1.5). This procedure is repeated $d$ times until the particles have dimension equal to that of $\pi$.

### 1.4.3 Sequential Monte Carlo

The importance distribution $\eta$ constructed with SIS is merely an approximation to $\pi$, and may be a fairly poor one in practice depending on the application. Try as we might to keep the variances of the

**Algorithm 1** Sequential importance sampling.

**for** $k = 1$ to $n$ **do**
  Sample $x_1^{(k)}$ from $\eta(x_1)$ ▷ Initialize the $k$th particle
  $w^{(k)} \leftarrow \pi(x_1^{(k)}) / \eta(x_1^{(k)})$ ▷ Initialize importance weight
**end for**
**for** $i = 2$ to $d$ **do**
  **for** $k = 1$ to $n$ **do**
    Sample $x_i^{(k)}$ from $\eta(x_i \mid \mathbf{x_{i-1}^{(k)}})$ ▷ Extend the $k$th particle
    $w^{(k)} \leftarrow [\pi(x_i^{(k)} \mid \mathbf{x_{i-1}^{(k)}}) \,/\, \eta(x_i^{(k)} \mid \mathbf{x_{i-1}^{(k)}})] \cdot w^{(k)}$ ▷ Update importance weight
  **end for**
  Normalize the weights so that $\sum w = 1$
**end for**

weights low, the cumulative errors at each sequential step tend to push many of the weights to very low values [115]. This results in a poor approximation to $\pi$, since only a few particles retain high importance weights after all $d$ sequential steps, a problem known as *particle degeneracy*. To mitigate this problem, Gordon, Salmond, and Smith [120] introduced technique they called the *bootstrap filter*, which involves a resampling of the population of particles after each sequential step in accordance with their importance weights. A similar idea, termed *particle rejuvenation*, was proposed by Liu and Chen [123]. These approaches cause particles with high importance weights to be replicated in the population, while particles with low weights may be removed. After each resampling step, the importance weights for all particles are set equal.

The resampling step was formally integrated with SIS by Doucet, Godsill, and Andrieu [115] to form the first SMC algorithm (algorithm 2). Rather than resample at every step as the bootstrap filter proposed, the authors use a criterion based on the expected sample size (ESS) the particle population to determine when resampling is necessary. 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 in the population. Resampling is triggered when the ESS drops below the threshold (conventionally $n/2$ [117]). Various resampling strategies beyond basic sampling with replacement have been proposed [124], but we will not discuss those here.

### 1.4.4 The sequential Monte Carlo sampler

The SIS and SMC algorithms described above aim to sample from a high-dimensional distribution $\pi(\mathbf{x})$, by sequentially sampling from $d$ distributions of lower but increasing dimension. Del Moral, Doucet, and Jasra [118] developed the *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 $\pi_i$ 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 each $\gamma_i$ is known pointwise and the normalizing constants $Z_i$ are unknown.

**Algorithm 2** Sequential Monte Carlo [115].

**for** $k = 1$ to $n$ **do**
  Sample $x_1^{(k)}$ from $\eta(x_1)$ ▷ Initialize the $k$th particle
  $w^{(k)} \leftarrow \pi(x_1^{(k)}) / \eta(x_1^{(k)})$ ▷ Initialize importance weight
**end for**
**for** $i = 2$ to $d$ **do**
  **for** $k = 1$ to $n$ **do**
    Sample $x_i^{(k)}$ from $\eta(x_i \mid \mathbf{x_{i-1}^{(k)}})$ ▷ Extend the $k$th particle
    $w^{(k)} \leftarrow [\pi(x_i^{(k)} \mid \mathbf{x_{i-1}^{(k)}}) \,/\, \eta(x_i^{(k)} \mid \mathbf{x_{i-1}^{(k)}})] \cdot w^{(k)}$ ▷ Update importance weight
  **end for**
  **if** $\mathrm{ESS}(w) < T$ **then** ▷ $T$ is a user-defined threshold
    Resample the particles according to $w$
    **for** $k = 1$ to $n$ **do**
      $w^{(k)} \leftarrow 1/n$
    **end for**
  **end if**
**end for**

Both algorithms involve progression through a sequence of related distributions. For SIS and SMC, 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 $\pi_i$ can be easily transformed into one for $\pi_{i+1}$. In particular, the sequence of importance distributions $\eta_i$ is constructed as

$$\eta_i(x') = \int \eta_{i-1}(x) K_i(x, x') \mathrm{d}x, \tag{1.6}$$

where $K_i$ is a Markov kernel and the integral is over the space on which the $\pi_i$ are defined. The choice of $K_i$ should be based on the perceived relationship between $\pi_{i-1}$ and $\pi_i$. Del Moral, Doucet, and Jasra [118] propose the use of a MCMC kernel with equilibrium distribution $\pi_i$. That is,

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

where $q(x, x')$ is a proposal function such as a Gaussian distribution centred at $x$ from which $x'$ is drawn (see appendix A).

Although this method of building up $\eta$ appears straightforward, the drawback is that the importance distribution itself becomes intractable. In particular, evaluating $\eta_i(x)$ involves an $i$-dimensional integral of the type in eq. (1.6). 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 $\pi_i$. Del Moral, Doucet, and Jasra [118] 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; they will later be precisely defined on a problem-specific basis. Second, the authors 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. That is, $\tilde{\pi}_i$ has dimension equal to the original dimension of $\pi_i$ raised to the power of $i$. This brings us back to the SIS setting described above (section 1.4.2), namely of building up an importance distribution sequentially through lower-dimensional distributions. The dimension of the importance distributions $\eta$ is similarly augmented with the forward kernels,

$$\eta_i(\mathbf{x_i}) = \eta_1(x_1) \prod_{k=1}^{i-1} K_k(x_k, x_{k+1}) \tag{1.7}$$

Thanks to the backwards kernels, we can write $\tilde{\pi}_i$ in terms of $\tilde{\pi}_{i-1}$ as follows.

$$\begin{aligned}
\tilde{\pi}_i(\mathbf{x}_i) &= \pi_i(x_i) \prod_{k=1}^{i-1} L_k(x_{k+1}, x_k) && \text{definition of } \tilde{\pi}_i \\
&= \pi_i(x_i) L_{i-1}(x_{i-1}, x_i) \prod_{k=1}^{i-2} L_k(x_{k+1}, x_k) && \text{pull out } k = i-1 \text{ term from product} \\
&= \frac{\pi_i(x_i) L_{i-1}(x_{i-1}, x_i)}{\pi_{i-1}(x_{i-1})} \cdot \pi_{i-1}(x_{i-1}) \prod_{k=1}^{i-2} L_k(x_{k+1}, x_k) && \text{multiply and divide by } \pi_{i-1}(x_{i-1}) \\
&= \frac{\pi_i(x_i) L_{i-1}(x_{i-1}, x_i)}{\pi_{i-1}(x_{i-1})} \cdot \tilde{\pi}_{i-1}(\mathbf{x_{i-1}}) && \text{definition of } \tilde{\pi}_{i-1}.
\end{aligned}$$

The importance distributions can also be expressed recursively using the forward kernels, which follows directly from eq. (1.7),

$$\eta_i(\mathbf{x_i}) = \eta_{i-1}(\mathbf{x_{i-1}}) K_i(x_{i-1}, x_i).$$

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

$$\begin{aligned}
w(\mathbf{x_i}) &= \frac{\tilde{\pi}_i(\mathbf{x_i})}{\eta_i(\mathbf{x_i})} && \text{definition of importance weight} \\
&= \frac{\tilde{\pi}_i(\mathbf{x_i})}{\eta_{i-1}(\mathbf{x_{i-1}}) K_i(x_{i-1}, x_i)} && \text{recursive definition of } \eta_i \\
&= \frac{\tilde{\pi}_{i-1}(\mathbf{x_{i-1}}) \pi_i(x_i) L_{i-1}(x_i, x_{i-1})}{\eta_{i-1}(\mathbf{x_{i-1}}) \pi_{i-1}(x_{i-1}) K_i(x_{i-1}, x_i)} && \text{recursive definition of } \pi_i \\
&= w(\mathbf{x_{i-1}}) \cdot \frac{\pi_i(x_i) L_{i-1}(x_i, x_{i-1})}{\pi_{i-1}(x_{i-1}) K_i(x_{i-1}, x_i)} && \text{definition of importance weight} \\
&\propto w(\mathbf{x_{i-1}}) \cdot \frac{\gamma_i(x_i) L_{i-1}(x_i, x_{i-1})}{\gamma_{i-1}(x_{i-1}) K_i(x_{i-1}, x_i)} && \text{remove normalizing constant } Z_i/Z_{i-1}
\end{aligned} \tag{1.8}$$

The final key piece of information is to notice that, because the $L_i$ are Markov kernels, $\pi_i$ 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 $\pi_i$, by considering only the $i$th component of $\mathbf{x_i}$. In fact, since the weight update eq. (1.8) depends only on the $i$th and $i-1$st components of each particle, we do not even need to keep track of the complete particles if we are only interested in the final distribution. The sequences of kernels $L$ and $K$ should be chosen based on the problem at hand to minimize the variance in the importance weights as well as possible. For a fixed choice of $K$, the backward kernels which minimize this variance are called the *optimal* backward kernels. The full SMC sampler algorithm is presented as algorithm 3. A resampling step is applied whenever the ESS of the population drops too low, as discussed in the previous section.

**Algorithm 3** Sequential Monte Carlo sampler of Del Moral, Doucet, and Jasra [118].

**for** $k = 1$ to $n$ **do**
  Sample $x_1^{(k)}$ from $\eta_1(x_1)$ ▷ Initialize the $k$th particle
  $w^{(k)} \leftarrow \gamma_1(x_1^{(k)})/\eta_1(x_1^{(k)})$ ▷ Initialize the importance weights
  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
    $w^{(k)} \leftarrow w^{(k)} \cdot \dfrac{\gamma_i(x_i)L_{i-1}(x_i, x_{i-1})}{\gamma_{i-1}(x_{i-1})K_i(x_{i-1}, x_i)}$ ▷ Update the importance weights
  **end for**
  Normalize the weights so that $\sum w = 1$
  **if** $\mathrm{ESS}(w) < T$ **then** ▷ $T$ is a user-defined threshold
    Resample the particles according to $w$
    **for** $k = 1$ to $n$ **do**
      $w^{(k)} \leftarrow 1/n$
    **end for**
  **end if**
**end for**

## 1.5 Approximate Bayesian computation

### 1.5.1 Overview and motivation

Sequential Monte Carlo, and the SMC sampler, were developed for sampling from distributions which can be evaluated up to a normalizing constant. We claim, and shall argue more thoroughly below (section 2.1.4) that the posterior distribution

$$\pi(\theta \mid T) \propto f(T \mid \theta)\pi(\theta)$$

for a contact network model with parameters $\theta$ and an input transmission tree $T$ does not fall in this category. Therefore, SMC, and other Bayesian and maximum likelihood (ML) techniques for fitting

mathematical models (see appendix A), cannot be directly applied to our problem. In particular, MCMC and the SMC sampler are designed for distributions $\pi$ which can be evaluated up to a normalizing constant $Z$, that is, $\pi(x) = \gamma(x)/Z$. Both algorithms calculate a ratio of the form $\pi(x)/\pi(x') \propto \gamma(x)/\gamma(x')$ for a current value $x$ and proposed updated value $x'$ - for MCMC, this is part of the Metropolis-Hastings ratio, while for the SMC sampler, a similar ratio is required to calculate the importance weights. In the context of Bayesian inference, this ratio contains a likelihood ratio, which must be calculated by computing the individual likelihoods and dividing them. If the likelihood is intractable, this is clearly not a viable approach.

Approximate Bayesian computation (ABC) [13–15] was developed to estimate posterior distributions with intractable likelihoods, which have arisen frequently in the domain of population genetics [16, 125]. ABC navigates around the intractable likelihood by replacing the posterior as the target of inference by an *approximate* posterior. This distribution is constructed in such a way that the ratios required for MCMC and the SMC sampler can be computed, conveniently allowing us to apply those algorithms with minimal changes. In the next section, we shall demonstrate how this is done, but first we give the definition of the approximate posterior.

By targeting the posterior distribution, Bayesian inference makes the assertion that model parameters with higher posterior density are "better" in the sense that they offer a more credible explanation for the observed data. The approximate posterior targeted by ABC uses an alternative metric for parameter credibility: 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, let $y$ be the observed data to which we are trying to fit a model with parameters $\theta$. In the case of this work, the data is a transmission tree $T$, but we shall stick with the generic variable $y$ for now. Suppose we have a distance measure $\rho$ 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 $\varepsilon$ of the observed data $y$,

$$\pi_\varepsilon(\theta, z \mid y) = \frac{\pi(\theta) f(z \mid \theta) \mathbb{I}_{A_{\varepsilon,y}}(z)}{\int_{A_{\varepsilon,y} \times \Theta} \pi(\theta) f(z \mid \theta) \mathrm{d}\,\theta}. \tag{1.9}$$

Here, $A_{\varepsilon,y}$ is an $\varepsilon$-ball around $y$ with respect to $\rho$, $\Theta$ is the space of all possible model parameters, and $\mathbb{I}$ is the indicator function [126] (that is, $\mathbb{I}_{A_{\varepsilon,y}}(x) = 1$ if $x \in A_{\varepsilon,y}$ or 0 otherwise). The distribution $\pi_\varepsilon(\theta, z \mid y)$ will be referred to as the *ABC target distribution*. The term $f(z \mid \theta)$ appears to be the bothersome likelihood again, but this will turn out not to be a problem because we are simulating $z$ ourselves.

To return to the context of this thesis, the observed data $y$ is an estimated transmission tree for a viral epidemic under investigation. The model in question is a contact network model with parameters $\theta$. The simulated dataset $z$ is a transmission tree, obtained by first generating a contact network under the model, and then simulating the spread of an epidemic over that network. A transmission tree can be constructed by keeping track of who infected whom during the simulated epidemic (further details will be given in section 2.1.1). The distance function $\rho$ must compare two transmission trees - the observed tree, which

takes the place of $y$, and the simulated tree $z$. We will define this distance function using the tree kernel discussed in section 1.2.4. To write eq. (1.9) in words, the approximate posterior we consider here is a distribution which assigns a joint probability density to model parameters and simulated transmission trees under those parameters. The probability density is proportional to the product of the prior on the parameters, and the likelihood of the simulated transmission tree under those parameters, but only if the simulated transmission tree is sufficiently close to the observed tree. Otherwise, the probability density is zero.

In fact, it is not the ABC target distribution itself, but rather its marginal in $z$, which approximates the posterior distribution. In other words, we claim that

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

The intuition for why this approximation might be reasonable comes from considering the case when $\varepsilon = 0$ and $\rho$ has the property that $\rho(z,y)$ if and only if $x = y$. In that case, the $\varepsilon$-ball around $y$ should contain only $y$ itself, hence the integral on the left is exactly equal to the posterior. Thus, by taking $\varepsilon$ small, we should attain something close to the posterior if $\rho$ captures the similarity between datasets reasonably well. However, the accuracy of the ABC approximation depends heavily on the choice of distance function [127, 128].

**Distance functions and summary statistics in ABC**

In many applications (eg. [15, 129]), $\rho$ is defined as $\rho(S(\cdot), S(\cdot))$ where $S$ is a function which maps data points into a vector of summary statistics. In the context of ABC, a summary statistic $S$ is called *sufficient* if

$$\pi(\theta \mid y) = \pi(\theta \mid S(y)).$$

That is, sufficiency implies that the data can be replaced with the summary statistic without losing any information about the posterior distribution [130]. For most problems, it is not possible to find sufficient summary statistics [130]. A number of sophisticated methods have been developed for selecting and weighting summary statistics based on various optimality criteria [127, 128, and references therein]. We do not apply these methods in this work, instead focusing on a distance function which is not based on summary statistics.

Although summary statistics are often presented as a fundamental part of ABC, they are not strictly necessary. For instance, if the data are numeric and of low dimension, the distance function may simply be the Euclidean distance [131]. Park et al. [18] proposed the use of a kernel function (as defined in section 1.2.4) in place of a distance function. The authors referred to their approach as "double-kernel ABC" due to the use of a second (unrelated) kernel function to compute the weights of the particles. The work by Poon [21], upon which ours is based, employed a similar approach, replacing the likelihood ratio in Bayesian MCMC with a ratio of kernel scores.

### 1.5.2 Algorithms for ABC

Algorithms for performing ABC can be grouped into three categories: rejection, MCMC, and SMC [126]. To simplify the notation, 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 $(\theta, z)$, so that the ABC target distribution can be written $\pi_\varepsilon(x \mid y)$. This makes our notation consistent with section 1.4.

Rejection ABC is the simplest method, and also the one that was first proposed [13, 14]. The algorithm, outlined in algorithm 4, repeats the following steps until a desired number of samples from the target distribution are obtained. Parameter values $\theta$ 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 $\varepsilon$, that is, with probability $\mathbb{I}_{A_{\varepsilon,y}}(z)$. Putting this all together, the parameters $\theta$ 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, $\theta$ represents an unbiased sample from the approximate posterior.

**Algorithm 4** Rejection ABC.

```
loop
    Draw θ according to π(θ)
    Simulate a dataset z from the model with parameters θ
    if ρ(y,z) < ε then
        Sample θ
    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 that is close to $z$. The inefficiency is compounded by the curse of dimensionality - the measure of the $\varepsilon$-ball around $y$ decreases exponentially with the number of dimensions. ABC-MCMC (algorithm 5) was designed to overcome these hurdles [132]. The approach is similar to ordinary Bayesian MCMC (appendix A), except that a distance cutoff replaces the likelihood ratio. That is, the transition probability between states $x$ and $x'$ is defined as

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

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

**Algorithm 5** ABC-MCMC.

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

data is low. Various strategies have been developed to mitigate this, including reducing the tolerance level $\varepsilon$ as the chain progresses [133].

The most recently developed class of algorithm for ABC is ABC-SMC [131, 134]. As with ABC-MCMC, the algorithm is a straightforward modification of an existing Bayesian inference method, in this case the SMC sampler (section 1.4.4). The sequence of target distributions is defined as $\pi_i(x) = \pi_{\varepsilon_i}(x \mid y)$ for a decreasing sequence of tolerances $\varepsilon_i$. The intention is for the algorithm to progress smoothly through a sequence of target distributions which ends at the ABC approximation to the posterior. The initial value $\varepsilon_1$ is set to $\infty$, which makes the first distribution in the sequence

$$\pi_1(\theta, z) = \frac{\pi(\theta) f(z \mid \theta)}{\int_{\mathbb{R}\times\Theta} \pi(\theta) f(z \mid \theta) \mathrm{d}\,\theta \mathrm{d}z}.$$

This initial distribution does not depend on the observed data $y$. In the terminology of the SMC sampler (algorithm 2), the numerator is the first of the $\gamma$'s, that is, $\gamma_1 = \pi(\theta) f(z \mid \theta)$. Sampling in proportion to $\gamma_1$ is straightforward and was already demonstrated for rejection ABC above. Because the sampling is exact, the initial importance weights are all set equal to 1 and normalized to $1/n$ where $n$ is the number of particles.

As discussed in section 1.4.4, the choices of the kernels $K$ and $L$ are problem-specific, and so appropriate kernels must be chosen for ABC. Several options have been proposed [20, 131, 134]. With an appropriately chosen kernel, the weight update (eq. (1.8)) will simplify into a computable expression. Sisson, Fan, and Tanaka [131] and Beaumont et al. [134] both suggest random walk kernels for $K$, where each particle is perturbed according to a Gaussian distribution. The backwards kernels $L$ are chosen to approximately minimize the variance in importance weights. In this thesis, we use an MCMC kernel, as proposed by Del Moral, Doucet, and Jasra [20]. The associated backwards kernels and weight updates are discussed in the next chapter (section 2.1.3). With generic kernels $K$ and $L$, the ABC-SMC algorithm is almost identical to the SMC sampler (algorithm 3), with $\gamma_i$ replaced with $\pi_i$. Therefore we will not repeat the full algorithm here.

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 $\varepsilon$-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

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

## 1.6 Summary

In section 1.1, we outlined three research objectives that will be addressed in this thesis. First, we aim to develop a method for fitting contact network models to estimated transmission trees. Although transmission trees can be estimated for any epidemic by thorough contact tracing, viral diseases are the most useful context for this method due to the possibility of inferring the trees from sequence data. Transmission and sequence evolution occur on similar time scales for RNA viruses, resulting in viral phylogenies whose shapes are heavily constrained by the transmission process. The study of this interaction between evolution and epidemiology is called *phylodynamics*; phylodynamic methods make it possible to estimate transmission trees from viral sequence data. These estimated trees form the input data for our method.

The desired output of our method is a posterior distribution of the parameters of a contact network model. Rather than assuming a homogeneously mixed population, as most epidemiological models do, network models take the more realistic view that human populations are structured. For example, contacts may tend to occur more often between "like" individuals (with respect to geography, socioeconomic status, or other factors) than at random. A network model provides a formal representation of this structure. In particular, our second research objective is to characterize our ability to fit the Barabási-Albert (BA) model, which incorporates preferential attachment to generate networks with realistic degree distributions.

To fit these models, our method will apply approximate Bayesian computation (ABC), a simulation-based approach. Simulating a transmission tree according to a network model is straightforward: a network can be generated according to the model, and the spread of an epidemic can be simulated over the network and recorded in a transmission tree. ABC uses the concordance between these simulated transmission trees and the "true" estimated tree as an indicator of parameter credibility. The closer the simulated transmission trees appear to the true tree, the more weight is assigned to the associated network parameters.

ABC can be implemented by at least three classes of algorithm, but the one we choose to apply in this work is sequential Monte Carlo (SMC). SMC uses a population of parameter "particles" to approximate a distribution of interest, in this case the approximate posterior distribution targeted by ABC. After running the ABC-SMC algorithm, statistics on the model parameters can be approximated by the weighted population of particles. For example, a weighted average would give an approximate expected value for each parameter.

Thus, our method integrates four research topics: phylogenetics, contact networks, sequential Monte Carlo, and approximate Bayesian computation. The first two topics together form the problem domain. Phylogenetic data are the input to our method, while estimates of the parameters of contact network models are the desired output. The latter two topics define the algorithm and statistical framework that our inference method will use.

# Chapter 2

# Reconstructing contact network parameters from viral phylogenies

In this chapter, we will address the three research aims of this thesis introduced in section 1.1. First, in section 2.1, we describe *netabc*, a computer program that implements an approximate Bayesian computation-based algorithm to fit contact network models to phylogenetic data. We also provide a justification for the use of ABC for this problem by arguing that the likelihood functions required to fit these models by more conventional means are likely to be computationally intractable. Second, in section 2.2, we perform a simulation study to investigate the Barabási-Albert network model, which uses a preferential attachment mechanism to generate networks with the power law degree distributions observed in real world social and sexual networks. We progress through two exploratory analyses testing the identifiability of the model's parameters, and conclude by testing *netabc*'s ability to recover the parameters from simulated transmission trees. Third, in section 2.3, we apply *netabc* to fit the BA model to six real world HIV datasets, with the understanding of the model parameters' identifiability gained through the simulation experiments. We conclude the chapter with a unified discussion of the three research aims, including interpretation of the results of both the simulated and real data experiments, as well as an examination of the limitations of our approach and opportunities for future investigation.

## 2.1 *Netabc*: a computer program for estimation of contact network parameters with ABC

*Netabc* is a computer program to perform statistical inference of contact network parameters from an estimated transmission tree using ABC. As discussed in section 1.1, the principal statistical algorithm used by *netabc* is adaptive ABC-SMC [20]. In addition, there are two supplementary components which are specific to the domain of phylogenetics and contact networks: Gillespie simulation [135], to simulate transmission trees on contact networks; and the tree kernel [17], which is used as the distance function in ABC to compare transmission trees [21] (see section 1.5). We give a high-level overview of the program here, before describing these components in detail. *Netabc* takes as input an estimated transmission tree,

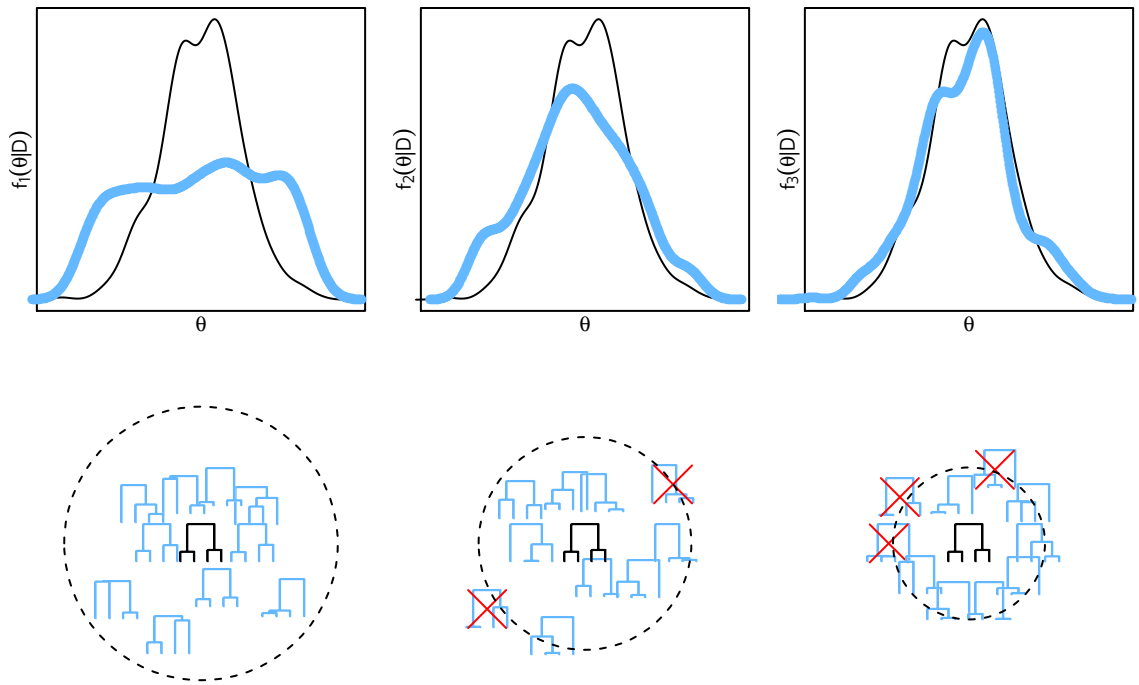


Figure 2.1: Graphical schematic of the ABC-SMC algorithm implemented in *netabc*. Particles are initially drawn from their prior distributions, making the initial population a Monte Carlo approximation to the prior. At each iteration, particles are perturbed, and a distance threshold around the input 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.

which can be derived from a viral phylogeny by rooting and time-scaling as described in section 1.2.3 or estimated by other methods [31, 53, 64–67]. We variously refer to this estimated transmission tree as the observed tree, estimated tree, or input tree.

As described in section 1.4, *netabc* keeps track of a population of particles $x^{(k)}$ indexed by an integer $k$, each of which contains particular parameter values $\theta^{(k)}$ for the contact network model we are trying to fit to the input tree. A small number $M$ of contact networks $z^{(k,i)}$, $1 \leq i \leq M$, are generated under the model 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 input tree using the tree kernel. Particles are weighted according to the similarity of their associated simulated trees with the input 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 input 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 shown in fig. 2.1.

*Netabc* is written in the *C* programming language. The *igraph* library [136] is used to generate and

store contact networks and phylogenies. Judy arrays [137] are used for hash tables and dynamic programming matrices. The GNU scientific library (GSL) [138] is used to generate random draws from probability distributions, and to solve for the next $\varepsilon$ by bisection in the adaptive ABC-SMC algorithm. Parallelization is implemented with Portable Operating System Interface (POSIX) threads [139]. In addition to the *netabc* binary to perform ABC, we provide three additional stand-alone utilities: *treekernel*, to calculate the tree 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.

To check that our implementation of Gillespie simulation was correct, we reproduced Figure 1A of Leventhal et al. [106] (our fig. B.1), which plots the imbalance of transmission trees simulated over four network models at various levels of pathogen transmissibility. Our implementation of adaptive ABC-SMC was tested by applying it to the same mixture of Gaussians used by Del Moral, Doucet, and Jasra [20] to demonstrate their method (originally used by Sisson, Fan, and Tanaka [131]). We were able to obtain a close approximation to the function (see fig. B.2), and attained the stopping condition used by the authors in a comparable number of steps. To check that the algorithm would converge to a bimodal distribution, we also applied it to a mixture of two Gaussians with means $\pm 4$ and variances 1. The algorithm was able to recover both peaks (fig. B.3).

### 2.1.1 Simulation of transmission trees over contact networks

The simulation of epidemics, and the corresponding transmission trees, over contact networks is performed in *netabc* using the Gillespie simulation algorithm [135]. This method has been independently implemented and applied by several authors [*e.g.* 94, 97, 106, 108, 111]. Groendyke, Welch, and Hunter [97] 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* as part of *netabc*.

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 $\beta_e$, which indicates that, once $u$ becomes infected, the waiting time until $u$ infects $v$ is distributed as Exponential$(\beta_e)$. Note that $v$ may become infected before this time has elapsed, if $v$ has other incoming edges. $v$ also has a removal rate $\nu_v$, so that the waiting time until removal of $v$ from the population is Exponential$(\nu_v)$. 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. In the epidemiology literature, the symbol $\gamma$ is usually used in place of $\nu$; we use $\nu$ here to distinguish the recovery rate from the power law exponent of scale free networks (see section 1.3.2).

To describe the algorithm, we introduce some notation and variables. Let $\text{inc}(v)$ be the set of incoming edges to $v$, and $\text{out}(v)$ be the set of outgoing edges from $v$. Let $\mathcal{I}$ be the set of infected nodes in the network, $\mathcal{R}$ be the set of removed nodes, $\mathcal{S}$ be the set of susceptible nodes, and $\mathcal{D}$ be the set of discordant edges in the network. Let $\beta$ be the total transmission rate over all discordant edges, and $\nu$ be

the total removal rate of all infected nodes,

$$\beta = \sum_{e \in \mathcal{D}} \beta_e, \quad \nu = \sum_{v \in \mathcal{I}} \nu_v.$$

The variables $\mathcal{S}$, $\mathcal{I}$, $\mathcal{R}$, $\mathcal{D}$, $\beta$, and $\nu$ are all updated as the simulation progresses. When a node $v$ becomes infected, it is deleted from $\mathcal{S}$ and added to $\mathcal{I}$. Any formerly discordant edges in $\text{inc}(v)$ are deleted from $\mathcal{D}$, and edges in $\text{out}(v)$ to nodes in $\mathcal{S}$ are added to $\mathcal{D}$. If $v$ is later removed, it is deleted from $\mathcal{I}$ and added to $\mathcal{R}$, and any discordant edges in $\text{out}(v)$ are deleted from $\mathcal{D}$. At the time of either infection or removal, the variables $\beta$ and $\nu$ are updated to reflect the changes in the network.

The Gillespie simulation algorithm is given as section 2.1.1. The transmission tree $T$ is simulated along with the epidemic. We keep a map called "*tip*", which maps infected nodes in $\mathcal{I}$ 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 ($I$). We use the notation $\text{Uniform}(0,1)$ to indicate a number drawn from a uniform distribution on $(0,1)$, and $\text{Exponential}(\lambda)$ to indicate a number drawn from an exponential distribution with rate $\lambda$. The combined number of internal nodes and tips in $T$ is denoted $|T|$. The updates to $\mathcal{S}$, $\mathcal{I}$, $\mathcal{R}$, $\mathcal{D}$, $\beta$, and $\nu$ described in the previous paragraph are not written explicitly in section 2.1.1, as they are quite straightforward and would only obfuscate the pseudocode.

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

```
infect a node v at random, updating S, I, D, β, and ν
T ← a single node with label 1
tip[v] ← 1
t ← 0
while D ≠ ∅ and |I| + |R| < I and t < t_max do
    s ← min(t_max − t, Exponential(β + ν))
    for v ∈ tip do
        extend the branch length of tip[v] by s
    end for
    t ← t + s
    if t < t_max then
        if Uniform(0, β + ν) < β then
            choose an edge e = (u, v) from D with probability β_e/β and infect v
            tip[v] ← |T| + 1                          ▷ add new tips to tree and tip array
            tip[u] ← |T| + 2                          ▷ corresponding to u and 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
        else
            choose a node v from I with probability ν_v/ν and remove v
            delete v from tip
        end if
        update S, I, R, D, β, and ν
    end if
end while
```

### 2.1.2 Phylogenetic kernel

The tree kernel developed by Poon et al. [17] provides a comprehensive similarity score between two phylogenetic trees, via the dot-product of the two trees' feature vectors in the space of all possible subset trees with branch lengths (see section 1.2.4). Because the branch lengths are continuous, there are infinitely many possible subset trees; hence, the feature space is infinite-dimensional. The kernel was implemented using the fast algorithm developed by Moschitti [75]. 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 $O(|T_1||T_2|)$ for the two trees $T_1$ and $T_2$ being compared.

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 that is minimized when $T_1 = T_2$. This is similar to the normalization used by Poon et al. [17] and Collins and Duffy [74].

### 2.1.3 Adaptive sequential Monte Carlo for Approximate Bayesian computation

We implemented the adaptive SMC algorithm for ABC developed by Del Moral, Doucet, and Jasra [20]. This algorithm is similar to the reference ABC-SMC algorithm described in section 1.5.2, except that the sequence of tolerances $\varepsilon_i$ 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. A sudden precipitous drop in ESS would indicate that only a small number of particles had non-zero importance weights, which would result in a very poor Monte Carlo approximation to the target distribution. This situation is referred to as the collapse of the approximation or particle degeneracy (see section 1.4.3) and is mitigated by the adaptive approach. A single parameter controls the decay rate. In the original paper of Del Moral, Doucet, and Jasra [20], the parameter is called $\alpha$, but to avoid confusion with the BA parameter of the same name we will refer to it here as $\alpha_{\mathrm{ESS}}$. The tolerance $\varepsilon_i$ is chosen to satisfy

$$\mathrm{ESS}(w_i) = \alpha_{\mathrm{ESS}}\,\mathrm{ESS}(w_{i-1}),$$

where, $w_i$ is the vector of weights at the $i$th step. Note that, since $w_i$ depends on $\varepsilon_i$, this equation solves for the updated weights and the updated tolerance simultaneously. As pointed out by Del Moral, Doucet, and Jasra [20], 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 $\pi_{\varepsilon_i}$ and proposal distributions

$$q_i\left(\theta^{(k)}, \theta^{(k)\prime}\right) \prod_{j=1}^{M} f\left(z^{(j,k)\prime} \middle| \theta^{(k)\prime}\right),$$

where $\theta^{(k)}$ is the vector of model parameters associated with particle $x^{(k)}$ and $z^{(j,k)\prime}$, $1 \leq j \leq M$, are $M$ datasets simulated according to $\theta^{(k)\prime}$ [20]. In our implementation, the $q_i$ are constructed componentwise for $\theta$ out of Gaussian proposals for continuous parameters [20] and Poisson proposals for discrete parameters. For the Poisson proposals, the number of discrete 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. The variance of each proposal distribution is set equal to twice the empirical variance of the particles, following [20, 134]. The backwards kernels are

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

Here we have written $x'$ for $x_i$ and $x$ for $x_{i-1}$ to emphasize that $x_{i-1}$ is the current value of the particle and $x_i$ is the proposed value. When substituted into eq. (1.8), the forward kernels $K_i(x, x')$ and densities $\pi_i(x') = \pi_{\varepsilon_i}(x')$ cancel out, and we are left with the following weight update.

$$\begin{aligned}
w_i(x) &\propto w_{i-1}(x) \frac{\pi_i(x') L_{i-1}(x', x)}{\pi_{i-1}(x) K_i(x, x')} && \text{importance weight update from SMC} \\
&= w_{i-1}(x) \frac{\pi_i(x') \pi_i(x) K_i(x, x')}{\pi_{i-1}(x) K_i(x, x') \pi_i(x')} && \text{substitute } L_{i-1} \\
&= w_{i-1}(x) \frac{\pi_i(x)}{\pi_{i-1}(x)} && \text{cancel } K_i(x, x') \text{ and } \pi_i(x') \\
&= w_{i-1}(x) \frac{\pi(\theta) \prod_{j=1}^{M} f(z^{(j)\prime} \mid \theta) \sum_{j=i}^{M} \mathbb{I}_{A_{\varepsilon_i, y}}(z^{(j)})}{\pi(\theta) \prod_{j=1}^{M} f(z^{(j)\prime} \mid \theta) \sum_{j=i}^{M} \mathbb{I}_{A_{\varepsilon_{i-1}, y}}(z^{(j)})} && \text{definition of } \pi_i(x) \\
&= w_{i-1}(x) \frac{\sum_{j=i}^{M} \mathbb{I}_{A_{\varepsilon_i, y}}(z^{(j)})}{\sum_{j=i}^{M} \mathbb{I}_{A_{\varepsilon_{i-1}, y}}(z^{(j)})} && \text{cancel prior and likelihood.}
\end{aligned}$$

In other words, when the distance threshold $\varepsilon_{i-1}$ is contracted to $\varepsilon_i$, the particles' weights are multiplied by the proportion of simulated datasets that are still inside the new threshold. The user may specify a final tolerance $\varepsilon$, or a final acceptance rate of the MCMC kernel, and the algorithm will be stopped when either of these termination conditions is reached. The latter condition stops the algorithm when the particles are not moving around very much, implying little change in the estimated target.

### 2.1.4 Justification for approach

We present here a non-rigorous justification for the use of ABC for the problem at hand, as opposed to more frequently-used approaches for fitting mathematical models (see appendix A). Consider a contact network model with parameters $\theta$, and an estimated transmission tree $T$. Taking a Bayesian approach,

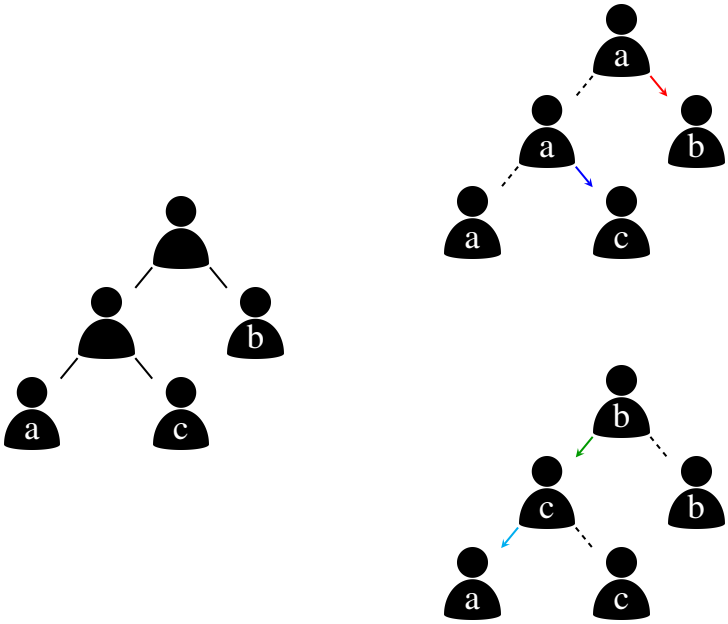


Figure 2.2: Illustration of an estimated transmission tree without labels (left) and two possible underlying complete transmission trees with labels (right). In the top right scenario, the epidemic began with node $a$ who transmitted first to $b$ and then to $c$. In the bottom right scenario, $b$ was the index case; $b$ infected $c$, who went on to infect $a$. A transmission tree estimated from a viral phylogeny would have the same topology and tip labels in both cases.

our aim is to obtain a sample from the posterior distribution on the model's parameters given our data,

$$\pi(\theta \mid T) = \frac{f(T \mid \theta)\pi(\theta)}{\int_\Theta f(T \mid \theta)\pi(\theta)\mathrm{d}\,\theta}.$$

For all but the simplest models, the normalizing constant in the denominator is an intractable integral. What we shall argue here is that, in contrast to most commonly studied mathematical models, the likelihood $f(T \mid \theta)$ is also likely to be intractable in our case.

As discussed in section 1.2.2, the internal nodes of transmission trees represent transmission events, and are labelled with the donor in the associated transmission pair. However, when we estimate a transmission tree from viral sequence data, we generally only know the labels of the tips of the tree, not the labels of the internal nodes. In viral phylogenies, the transmissions are at least partially preserved through the evolutionary relationships among the viruses, but the directionality of those transmissions is unknown. Thus, a single estimated transmission tree can correspond to many possible pathways of the epidemic through the network. Figure 2.2 illustrates this concept for a simple transmission tree with three tips. When calculating a likelihood given a transmission tree, we must sum over all possible labellings of the internal nodes. Let $\mathcal{L}$ be the set of such labellings. Then

$$f(T \mid \theta) = \sum_{l \in \mathcal{L}} f(T, l \mid \theta). \tag{2.1}$$

A contact network model assigns a probability density to each possible contact network. Transmission trees are realized over particular contact networks, not over the model itself. Therefore, we must also sum over all contact networks which could be generated by the model. Let $\mathcal{G}$ be the set of all possible contact networks. Summing eq. (2.1) over $\mathcal{G}$ gives

$$f(T \mid \theta) = \sum_{G \in \mathcal{G}} \sum_{l \in \mathcal{L}} f(T, l \mid G, \theta) f(G \mid \theta). \tag{2.2}$$

This can be simplified somewhat by noticing that, given a specific contact network, the labelled transmission tree depends only on that network and not on the model that generated it. That is, $f(T, \nu \mid G, \theta) = f(T, l \mid G)$, and

$$f(T \mid \theta) = \sum_{G \in \mathcal{G}} f(G \mid \theta) \sum_{l \in \mathcal{L}} f(T, \nu \mid G) \tag{2.3}$$

Under the assumption that both transmission and removal are Poisson processes, calculating $f(T, l \mid G)$ can be accomplished by a straightforward modification of the Gillespie simulation algorithm (section 2.1.1). Rather than choosing transmission or removal events according to their probabilities, the events would be deterministically chosen to mimic the events recorded in the transmission tree. Assuming efficient data structures for storing lists of nodes and edges, the complexity of this calculation would be $O(|T|)$. The number of possible labellings of internal nodes of $T$ is easily seen to be $2^{(|T|-1)/2}$ by noticing that each of the $(|T|-1)/2$ internal nodes must be labelled with the same label as either its right child or its left child. Although exponential calculations of this nature can often be simplified on trees using dynamic programming (*e.g.*[140]), it cannot be straightforwardly applied in this case because the subtrees' probabilities depend on the existing epidemic progress (their parents and siblings). Hence, calculating the inner sum over labels may take time $O(2^{(|T|-1)/2})$.

The outer sum, over all contact networks, is also difficult to evaluate in general. There are $2^{N(N-1)}$ directed graphs on $N$ nodes [141]. There must be at least as many nodes in the contact network as the number of tips in the tree, which is $(|T|+1)/2$. Of course, it is very likely that there are more nodes in the network than observed tips because some individuals are never infected and/or some infected individuals are never sampled. The complexity of calculating $f(G \mid \theta)$ is obviously dependent on the particular model being investigated. For the BA model, we might have to sum over all possible orders in which the nodes could be added, and all assignments of edges to the nodes which generated them. However, even in the case that calculating $f(G \mid \theta)$ can be done in constant time, the sum (2.3) still has at least $O(2^{|T|^2})$ terms.

We have shown that both the normalizing constant $\int_\Theta f(T \mid \theta)\pi(\theta)\mathrm{d}\,\theta$ and the likelihood $f(T \mid \theta)$ are likely computationally prohibitive to calculate. If this is the case, the problem of fitting contact network models to phylogenies seems to be of the *doubly-intractable* type [142], which would imply that these models are amenable to neither ML nor Bayesian inference techniques. Although both methods are able to cope with an intractable normalizing constant (for example, by local search for ML or Bayesian MCMC), neither can avoid the intractable likelihood calculations. This justifies the use of ABC, which

is a likelihood-free method.

We have not proven here that eq. (2.3) is impossible to calculate in polynomial time - it could be possible to algebraically simplify the sum into a tractable expression. Furthermore, under certain models, a large proportion of $\mathcal{G}$ may have zero probability, which would enable the simplification of the outer sum on a model-specific basis. It should also be noted that extensions of Bayesian MCMC have been developed for doubly-intractable problems [142, 143], which might be adaptable to the problem at hand. These have not been as widely used as ABC, nor are they as easily parallelizable as SMC.

## 2.2 Analysis of Barabási-Albert model with synthetic data

### 2.2.1 Why study the Barabási-Albert model?

We developed *netabc* with the objective of extracting useful, quantitative information about network structures from viral phylogenies. An important aspect of "usefulness" is model specification and the biological or epidemiological interpretation of the parameters. We want the model to be realistic, but not so complicated that it becomes difficult to interpret. At least some of the parameters should be of interest from a theoretical or practical perspective, or there would be no point in estimating them. Since *netabc* is a phylodynamic method, intended to be used with viral sequence data, we would also like to choose parameters which may be difficult to estimate with more standard methods. Otherwise, our method provides no advantage. The Barabási-Albert (BA) model (section 1.3.2) satisfies these criteria, albeit some better than others. The purported realism of the model stems from the fat-tailed degree distributions it produces, which are similar to those observed in real world sexual networks [22–25, 144]. Moreover, the "rich get richer" phenomenon, where popular individuals attract new connections at an elevated rate, is intuitively reasonable for both sexual [105] and IDU [103] networks. However, the model is very simple, assuming that all nodes form the same number of links when added to the network and share the same preference for popular individuals.

In this thesis, we consider four parameters related to the BA model, denoted $N$, $m$, $\alpha$, and $I$ (see section 1.3.2). The first three of these parameterize the network structure, while $I$ is related to the simulation of transmission trees over the network. However, we will refer to all four as BA parameters. $N$ denotes the total number of nodes in the network, or equivalently, susceptible individuals in the population; $m$ is the number of new undirected edges added for each new vertex, or equivalently one-half of the average degree; $\alpha$ is the power of preferential attachment – new nodes are attached to existing nodes of degree $d$ with probability proportional to $d^\alpha + 1$; finally, $I$ is the number of infected individuals at the time when sampling occurs. The $\alpha$ parameter is unitless, while $m$ has units of edges or connections per vertex, and $N$ and $I$ both have units of nodes or individuals.

From a public health standpoint, all four parameters are of some interest. The prevalence $I$ can be used to estimate the resources required to combat an ongoing epidemic, while total susceptible population size $N$ provides a similar metric for preventative measures. The average degree of the network, in this case $2m$, is directly related to $R_0$, the basic reproductive number [96]. $R_0$ quantifies the average number of secondary infections ultimately caused by one infected individual; higher $R_0$ generally

indicates faster epidemic growth and/or larger eventual epidemic size [5]. In a homogeneously mixed population, the theoretically expected proportion of the population which must be vaccinated to control an epidemic (the *vaccination threshold*) can be expressed in terms of $R_0$ [145]. Although optimal vaccination strategies may differ in heterogeneous contact structures [10], it is reasonable to suppose that there would still be a relationship between $m$, $R_0$, and the vaccination threshold. The preferential attachment power $\alpha$ quantifies the degree to which high-degree nodes, also called superspreaders [100], characterize the network structure. Superspreaders have been hypothesized to play a disproportionately greater role in the spread of several diseases [37, 101]. If so, network-based interventions [28, 29] may be worth considering as part of an epidemic control strategy. $\alpha$ can also offer some insight into how the network would react to the removal of nodes. Dombrowski et al. [103] found evidence of preferential attachment in IDU networks, and suggested that as a consequence of this characteristic, the removal of random nodes (such as through a police crackdown) might inadvertently make it easier for epidemics to spread. When individuals with only one or two connections lose them, they might tend to seek out well-known (that is, high-degree) members of the community, thus increasing those individuals' connectivity even further. Although we do not consider a dynamic network in this work, it is not unreasonable that the network could be close to static over short periods of time, and this static structure might be informative of future dynamic behaviour.

Though it is theoretically possible to estimate all four BA parameters using more conventional approaches, the cost of doing so may be high, in addition to other situation-specific challenges. In theory, any network parameter can be estimated by explicitly constructing the contact network, although this is highly resource intensive and is hampered by misreporting and other challenges [42]. All parameters become more difficult to estimate when the infected population is "hidden" due to illegal or stigmatized behaviour, as is sometimes the case with HIV outbreaks among IDU, MSM, or sex workers. In such cases, phylodynamic methods may offer the advantage of providing information about the whole population from only a small sample. A survey, for example, will not tell us if there are large groups of the population that we simply have not sampled. Our hope is that estimating $N$ and $I$ phylogenetically might provide this additional information. The average degree of the network, $2m$, is also estimable by a survey, although individuals may be unwilling or unable to disclose how many contacts they have had. Sequence data also provide an advantage in this respect – they are objective and do not share the same biases as self-reported partner counts. The estimation of $\alpha$ is more complex, as this parameter is most strongly reflected in the connectivity of very high degree nodes, who are rare in the population. Locating them might require contact tracing, or respondent-driven sampling [90]. Even if the full degree distribution of a network is available, there are models other than preferential attachment which can produce scale-free networks [*e.g.* 146]. de Blasio, Svensson, and Liljeros [105] were able to estimate $\alpha$ by maximum likelihood using partner count data from several sequential time intervals, but they admit such detailed data are not usually available. Moreover, their dataset was constructed via a random survey, which would likely miss the few high-degree nodes characterizing a power law degree distribution. In summary, each of the BA parameters may be estimated without phylodynamics, but there are sufficient difficulties that we believe an alternative method using sequence data is warranted.

### 2.2.2 Using synthetic data to investigate identifiability and sources of estimation error

We have argued that the parameters of the BA model are interesting and worth estimating with phylodynamic methods. However, these estimates will only constitute "useful" information about the network if the parameters are identifiable from phylogenetic data. Roughly speaking, the identifiability of a parameter says how much information about that parameter can possibly be obtained from the observed data. If the parameters of the BA model do not influence tree shape at all, then we cannot possibly estimate them – the posterior distribution will exactly resemble the prior, no matter how accurate a representation of the posterior we are able to produce. Hence, before proceeding with a full validation of *netabc* on simulated data, we undertook two experiments designed to assess the identifiability of the BA parameters. These experiments only investigated one parameter of the BA model at a time while holding all others fixed, a strategy commonly used when performing sensitivity analyses of mathematical models. This allowed us to perform a fast preliminary analysis without dealing with the "curse of dimensionality" of the full parameter space. The experiments are motivated and described on a high level here, with more detail provided in the next section.

First, we simulated trees under three different values of each parameter, and asked how well we could tell the different trees apart. The better we are able to distinguish the trees, the more identifiability we might expect for the corresponding parameter when we attempt to estimate it with ABC. This experiment also had the secondary purpose of validating our choice of the tree kernel as a distance measure in ABC. To tell the trees apart, we used a classifier based on the tree kernel, but we also tested two other tree shape statistics: one which considers only the topology, and another which considers only branching times. Since the tree kernel incorporates both of these sources of information, we expected it to outperform the other two statistics. Finally, the tree kernel can be "tuned" by adjusting the values of the meta-parameters $\lambda$ and $\sigma$. The results of this experiment were used to select values for these meta-parameters to carry forward to the rest of the thesis, based on their accuracy in distinguishing the different trees.

A second experiment was designed to test whether we could actually estimate the parameters numerically, rather than just telling three different values apart, and also to assess how identifiability varied in different regions of the parameter space. An individual tree was compared to simulated trees on a one-dimensional grid of values of one BA parameter, to obtain a "distribution" of kernel score values. From this distribution, estimates and credible intervals of the parameter could be calculated. Repeating this experiment with trees located throughout the parameter space allowed us to better quantify the identifiability. Furthermore, doing marginal estimation (that is, estimating one parameter with all others fixed) can provide insight into any biases observed when doing joint estimation with ABC. If grid search is inaccurate, it indicates a lack of parameter identifiability. However, if the marginal grid search estimates are accurate but the estimates obtained with ABC are biased, this points to confounding between the parameters which could only be observed when they are all estimated jointly.

After these preliminary experiments, the strategy we used for testing *netabc* was a standard simulation-based validation. Transmission trees were simulated under several combinations of parameter values, and we tried to recover these values with *netabc*. We then used a multivariable analysis to investigate how accuracy of these estimates was influenced by the true parameter values.

In the previous section, we argued that the BA model parameters were worth investigating, and here we have presented several computational experiments designed to assess their identifiability. There is a final, more technical aspect of our method's "usefulness" to consider, which is the accuracy of the ABC approximation to the posterior. As discussed in section 1.5, ABC does not target the posterior distribution directly, but rather an approximate posterior derived from simulated data and a distance function. ABC *assumes* that this approximate posterior resembles the true posterior, and it is critical for our estimates' relevance that this assumption holds. There are two potential causes of an inaccurate ABC approximation [147]. First, the Monte Carlo approximation to the ABC target distribution may be poor, due to the settings used for ABC-SMC. Second, the ABC target distribution may not resemble the posterior, due to a poor choice of distance function. We designed two experiments to investigate the impact of these sources of error.

The Monte Carlo approximation error is fairly easily quantified by simply increasing the computing power used for SMC. We ran one simulation using a larger number of particles, more simulated datasets per particle, and a higher value for $\alpha_{\text{ESS}}$ (defined in section 2.1.3). A substantial improvement in accuracy resulting from these changes would likely indicate a high Monte Carlo error with the lower settings. The second issue, the resemblance of the ABC target distribution to the true posterior, is somewhat more difficult to investigate. We do not have access to the true posterior, even for simulations where the true parameter values are known. To address this source of error, we performed marginal parameter estimation with ABC by informing *netabc* of some of the true parameter values. Any inaccuracy or bias observed only in the joint estimation results, but not the marginal estimates, is most easily explained by interdependence between parameters in the true posterior. However, errors observed in both marginal and joint estimates could be due to either the shape of the true posterior or an inaccurate ABC approximation, and we have no way to distinguish one from the other. In other words, this experiment provided only an upper bound on the error due to an inaccurate ABC approximation.

### 2.2.3 Methods

For all simulations, 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). Nodes in our networks followed simple SI dynamics, meaning that they became infected at a rate proportional to their number of infected neighbours, and never recovered. 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, the removal rate of each node was set to 0, and all transmission trees' branch lengths were scaled by their mean. The *igraph* library's implementation of the BA model [136] was used to generate the graphs. The analyses were run on Westgrid (https://www.westgrid.ca/) and a local computer cluster. With the exception of our own C programs, all analyses were done in *R*, and all packages listed below are *R* packages. Code to run all experiments is freely available at https://github.com/rmcclosk/thesis.

**Classifiers for BA model parameters based on tree shape**

Our first computational experiment was designed as an exploratory analysis of the four BA model parameters defined above: $\alpha$, $I$, $m$, and $N$. The objective of this experiment was to determine whether any of the four parameters were identifiable from the shape of the transmission tree, as quantified by the tree kernel. Each of the BA model parameters was varied one at a time, while holding the other parameters at fixed, known values. Contact networks were generated according to each set of parameter values, and transmission trees were simulated over the networks. We then evaluated how well a classifier based on the tree kernel could differentiate the trees simulated under distinct parameter values. If the classifier's cross-validation accuracy was high, this could be taken as an indication that the parameter in question was identifiable in the range of values considered. A caveat of this preliminary analysis is that, since all parameters but one were held at known values, nothing could be said about the identifiability of *combinations* of parameters; this issue will be explored later by jointly estimating all parameters with ABC.

In addition to testing for identifiability, a secondary objective of this analysis was to validate the use of the tree kernel as a distance measure for ABC in our context. As discussed in section 1.5, the choice of distance function is extremely important for the accuracy of the ABC approximation to the posterior. Therefore, we evaluated two additional tree statistics in the same manner as we evaluated the tree kernel (that is, by constructing and testing a classifier). First, we considered Sackin's index [69], which measures the degree of imbalance or asymmetry in a phylogeny (see section 1.2.4). Sackin's index is widely used for characterizing phylogenies [148] and has been demonstrated to vary between transmission trees simulated under different contact network types [106]. Sackin's index does not take branch lengths into account, considering only the tree's topology. The other statistic we considered was the normalized lineages-through-time (nLTT) [72], which compares two trees based on normalized distributions of their branching times (see section 1.2.4). In contrast with Sackin's index, the nLTT does not explicitly consider the trees' topologies, but it does use their normalized branch lengths. While the nLTT is a newly developed statistic not yet in widespread use, the unnormalized LTT [34] was the basis of seminal early work extracting epidemiological information from phylogenies [41]. We expected the tree kernel to classify the BA parameters more accurately than either Sackin's index or the nLTT, since the tree kernel takes both topology and branch lengths into account.

This experiment involved a large number of variables that 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, $\lambda$ and $\sigma$ (section 1.2.4), will be referred to as *meta-parameters* to distinguish them from the parameters of the BA model.

The attachment power parameter $\alpha$ 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 $\alpha$. 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 [149]. Both the tree kernel and the nLTT are pairwise statistics, and the SVMs classifiers we used to investigate them operate on pairwise distance matrices. 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 $300 \times 300$ kernel matrix. Similarly, we computed the nLTT statistic between each pair of trees using our *treestat* binary, and stored them in a second $300 \times 300$ matrix.

To investigate the identifiability of $\alpha$ from tree shape, we constructed classifiers for $\alpha$ based on the three tree shape statistics discussed above. The *kernlab* package [150] was used to create a kernel support vector machine (kSVM) classifier using the computed kernel matrix, and the *e1071* package [151] was used to create ordinary SVM classifiers using the pairwise nLTT matrix and Sackin's index values. Each of these classifiers was evaluated with 1000 two-fold cross-validations with equally-sized folds. We also performed a kernel principal components analysis (kPCA) projection of the kernel matrix, and used it to visualize the separation of the different $\alpha$ values in the tree kernel's feature space. A schematic of this experiment is presented in fig. 2.3.

Similar experiments were performed with the values shown in table 2.1. The other three BA parameters, $N$, $m$, and $I$, were each varied while holding the others fixed. The experiments for $\alpha$, $m$, 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 $\sigma = \{1/8, 1/4, 1/2, 1, 2, 4, 8\}$.

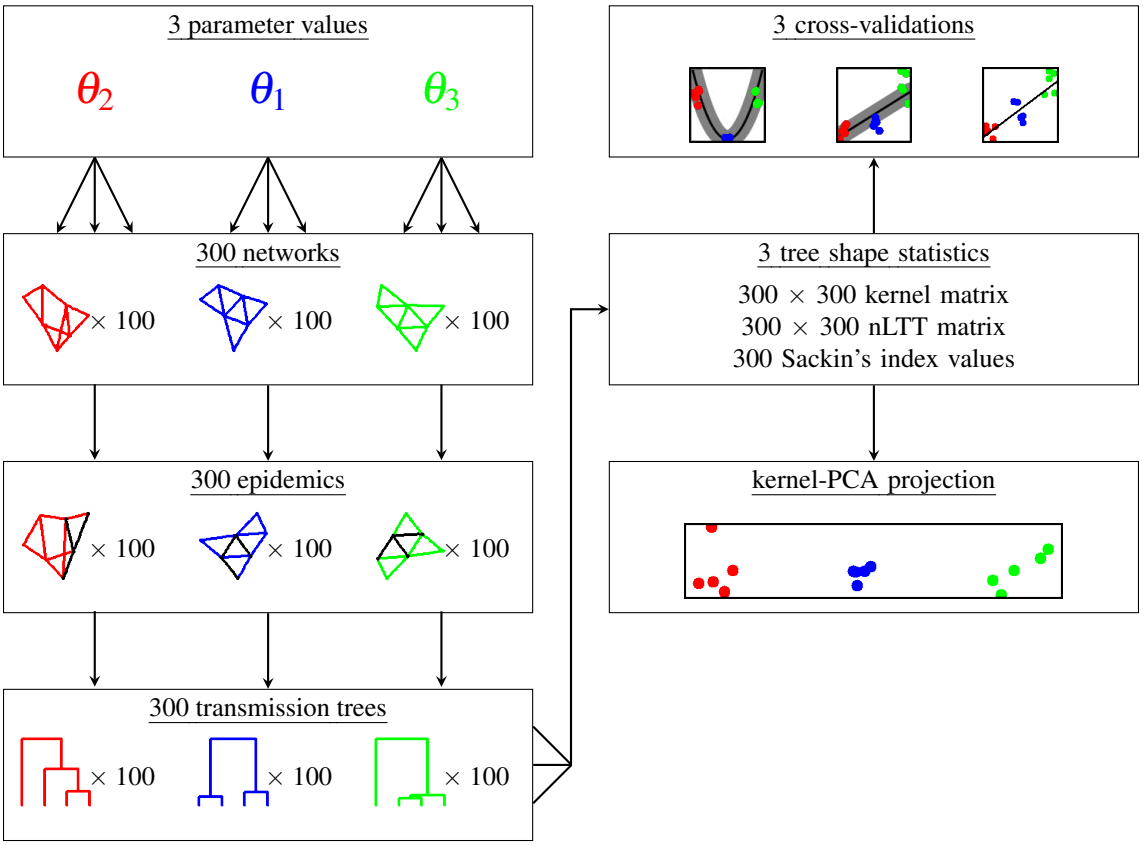


Figure 2.3: Schematic of classifier experiments investigating identifiability of BA model parameters from tree shapes. The parameters of the BA model were varied one at a time while holding all others fixed. Transmission trees were simulated under three different values of each parameter. Classifiers were constructed for each parameter based on three tree shape statistics, and their accuracy was evaluated by cross-validation. Kernel-PCA projections were used to visually examine the separation of the trees in the feature space defined by the tree kernel.

| varied parameter | $N$ | $\alpha$ | $m$ | $I$ | tips | $\lambda$ | $\sigma$ |
|---|---|---|---|---|---|---|---|
| $N$ | 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 |
| $\alpha$ | 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 meta-parameters used in classifier experiments to investigate identifiability of BA model parameters from tree shapes. 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 | $N$ | $\alpha$ | $m$ | $I$ | tips |
|---|---|---|---|---|---|---|---|
| $N$ | 1050, 1125, ..., 15000 | 1000, 3000, ..., 15000 | - | 1.0 | 2 | 1000 | 100, 500, 1000 |
| $\alpha$ | 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, 1000, 1500, 2000 | 5000 | 1.0 | 2 | - | 100, 500 |

Table 2.2: Values of parameters and meta-parameters used in grid search experiments to further investigate identifiability of BA model parameters. 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 each parameter, which were compared to the test values.

### Grid search

The previous experiment was an exploratory analysis intended to determine which of the BA parameters were identifiable, and whether the tree kernel could potentially be used to distinguish different parameter values when all others were held fixed. In this experiment, which was still of an exploratory nature, we continued to consider one parameter at a time while fixing the other three. However, rather than checking for identifiability, we were now interested in quantifying the accuracy and precision of kernel score-based estimates. This was done by examining the distribution of kernel scores on a grid of parameter values, when trees simulated according to those values were compared with a single simulated test tree.

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 $\alpha$ 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". 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 the known values $N = 5000$, $m = 2$, 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. For all parameters except $m$, 50% and 95% highest density intervals were obtained using the *hpd* function in the *TeachingDemos* package [152]. Since the *hpd* function assumes a continuous distribution, we implemented our own version for discrete distributions to use for $m$.

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. 2.4.

### Approximate Bayesian computation

Our final synthetic data experiment was designed to test the full ABC-SMC algorithm by jointly estimating the four parameters of the BA model. We used the standard validation approach of simulating transmission trees under the model with known parameter values and attempting to recover those values with *netabc*. The algorithm was not informed of any of the true parameter values for the main set of simulations. Despite the fact that the parameter values used to generate the simulated transmission trees were known, the true posterior distributions on the BA parameters were unknown. Therefore, any apparent errors or biases in the estimates could be due to either poor performance of our method, or to real features of the posterior distribution. The latter type of error reflects on the suitability of the model, but does not invalidate the use of our method in cases where the parameters are more identifiable. Two retrospective experiments were performed to disambiguate some of the observed errors: one where we ran a simulation with increased computational power to test for an increase in accuracy, and a second where we estimated parameters marginally to remove confounding from the other parameters in the

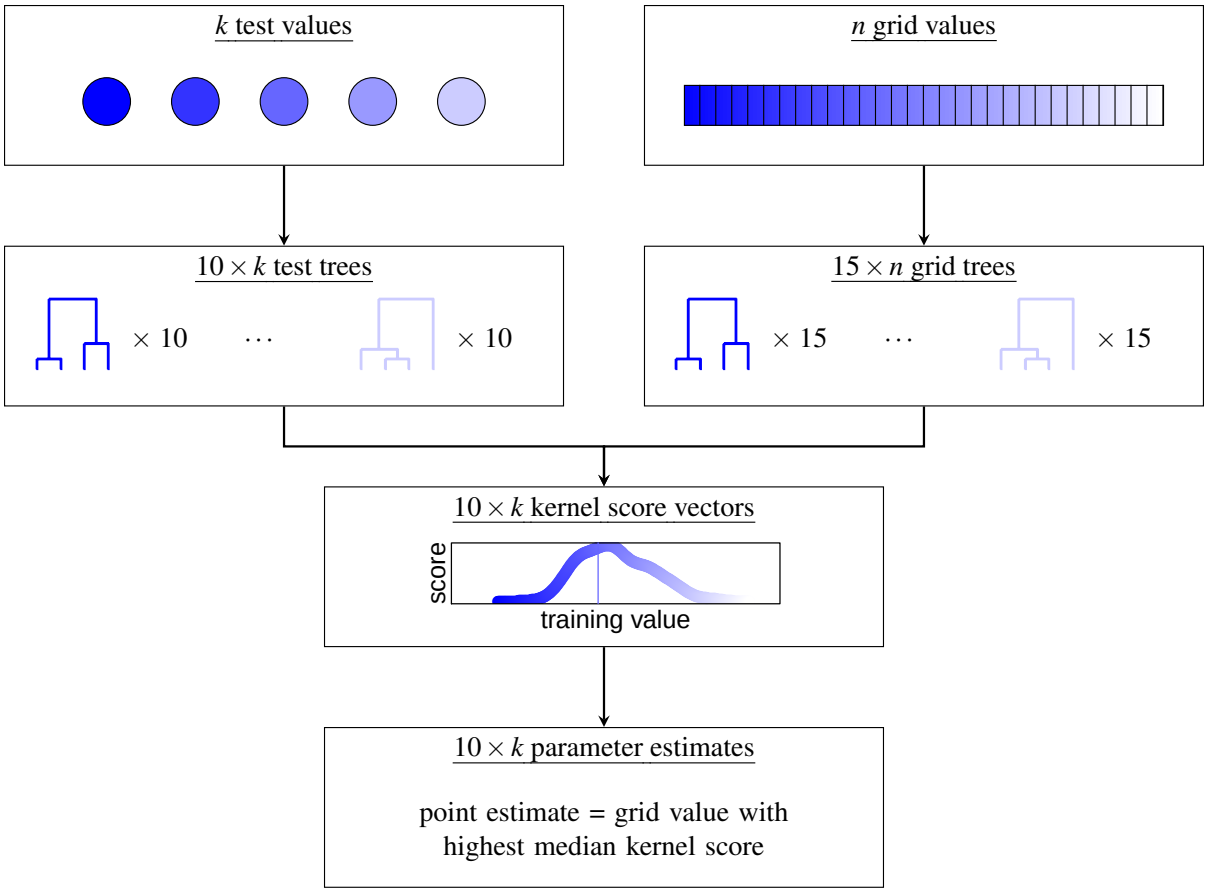


Figure 2.4: Schematic of grid search experiment to further investigate identifiability of BA model parameters from tree shapes. Trees were simulated along a narrowly spaced grid of values of one parameter ("grid trees") with all other parameters fixed to known values. Separate trees were simulated for a small subset of the grid values ("test trees"), also holding the other parameters fixed. Each test tree was compared to every grid tree using the tree kernel, and the resulting kernel scores were normalized to resemble a probability density from which the mode and 95% highest density interval were calculated.

joint posterior.

The parameter values and priors used are listed in table 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 $\alpha_{\text{ESS}}$ set to 0.95. The algorithm was stopped when the acceptance rate of the MCMC kernel (see section 2.1.3) dropped below 1.5%, the same criterion used by Del Moral, Doucet, and Jasra [20]. For visualization, approximate marginal posterior densities for each parameter were calculated using the *density* function in *R* applied to the final weighted population of particles. Posterior means obtained for each parameter using the *wtd.mean* function in the *Hmisc* package [153]. Credible intervals were obtained using the *hpd* function in the *TeachingDemos* package [152] for $\alpha$ $I$, and $N$, and using our own implementation for discrete distributions for $m$.

To evaluate the effects of the true parameter values on the accuracy of the posterior mean estimates, we analyzed the $\alpha$ and $I$ parameters individually using GLMs. The response variable was the error of

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

Table 2.3: Parameter values used in simulation experiments to test accuracy of BA model fitting with *netabc*. Trees were simulated under the test values, and *netabc* was used to estimate posterior distributions on the BA parameters for each simulated tree. *Netabc* was naïve to the true parameter values.

the point estimate, and the predictor variables were the true values of $\alpha$, $I$, and $m$. We did not test for differences across true values of $N$, because $N$ was not varied in these simulations. The distribution family and link function for the GLMs were chosen as Gaussian and inverse, respectively, by examination of residual plots and Akaike information criterion (AIC). The $p$-values of the estimated GLMs coefficients were corrected using Holm-Bonferroni correction [154] with $n = 6$ (two GLMs with three predictors each). Because there was clearly little to no identifiability of $N$ and $m$ with ABC (see results in next section), we did not construct GLMs for those parameters.

Two further simulations were performed to assess the possible impact of two types of model misspecification. To consider the effect of heterogeneity among nodes, we generated a network where half the nodes were attached with power $\alpha = 0.5$ and the other half with power $\alpha = 1.5$. The other parameters for this network were $N = 5000$, $I = 1000$, and $m = 2$. To investigate the effects of potential sampling bias [155], we simulated a transmission tree where the tips were sampled in a peer-driven fashion, rather than at random. That is, the probability to sample a node was twice as high if any of that node's network peers had already been sampled. The parameters of this network were $N = 5000$, $I = 2000$, $m = 2$, and $\alpha = 0.5$.

To assess the impact of the SMC settings on *netabc*'s accuracy, we ran *netabc* twice on the same simulated transmission tree. For the first run, the SMC settings were the same as in the other simulations: 1000 particles, 5 simulated transmission trees per particle, and $\alpha_{\text{ESS}} = 0.95$. The second run was performed with 2000 particles, 10 simulated transmission trees per particle, and $\alpha_{\text{ESS}} = 0.97$. To investigate the extent to which errors in the estimated BA parameters were due to true features of the posterior, rather than an inaccurate ABC approximation, we performed marginal estimation for one set of parameter values. Each combination of 1, 2, or 3 model parameters (14 combinations total) was fixed to their known values, and the remaining parameters were estimated with *netabc*. The parameter values were $\alpha = 0.0$, $m = 2$, $I = 2000$, and $N = 5000$.

### 2.2.4 Results

#### Classifiers for BA model parameters based on tree shape

Trees simulated under different values of $\alpha$ were visibly quite distinct (fig. 2.5). In particular, higher values of $\alpha$ 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 $\alpha$ values. None of the other three parameters produced trees that were as easily distinguished from each other (figs. B.4 to B.6). Sackin's index, which measures tree imbalance, was significantly correlated with all four parameters (for $\alpha$, $I$, $m$, 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 $\alpha$ and $I$, and positively correlated with $m$ and $N$ (Spearman's rho $-0.8$, $-0.69$, 0.09, 0.17; all $p < 10^{-5}$).

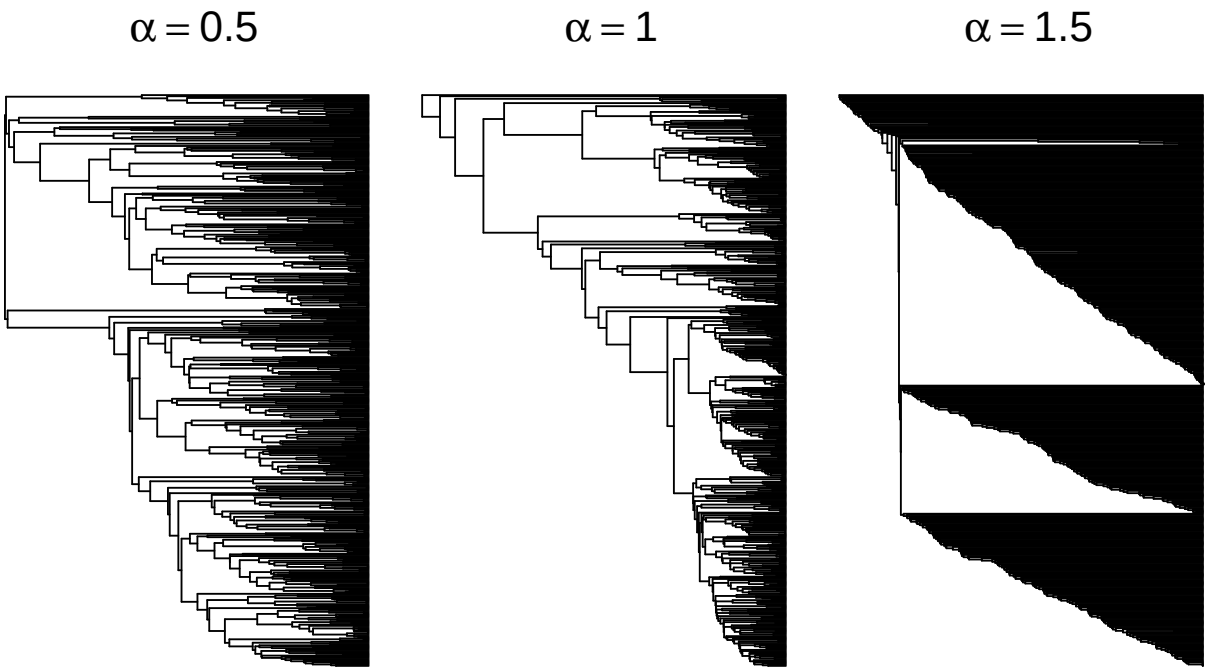


Figure 2.5: Simulated transmission trees under three different values of preferential attachment power ($\alpha$) parameter of BA model. Epidemics were simulated on BA networks of 5000 nodes, with $\alpha$ equal to 0.5, 1.0, or 1.5, until 1000 individuals were infected. Transmission trees were created by randomly sampling 500 infected nodes. Higher $\alpha$ values produced networks with a small number of highly-connected nodes, resulting in highly unbalanced, ladder-like trees.

Figure 2.6 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 $\alpha$ and $I$ values considered are well separated from each other in the feature space mapped to by the tree kernel. 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. B.11 to B.14). The values of $I$ and $N$ separated more clearly with larger numbers of tips, and in the case of $N$, with larger epidemic sizes (figs. B.12 and B.14).

The accuracy of each classifier, which is the proportion of trees assigned the correct parameter value, is shown in fig. 2.7. Since there were three possible values, random guessing would produce an accuracy of 0.33. Classifiers based on the nLTT and Sackin's index generally exhibited worse performance than the tree kernel, although the magnitude of the disparity varied between the parameters (fig. 2.7, centre

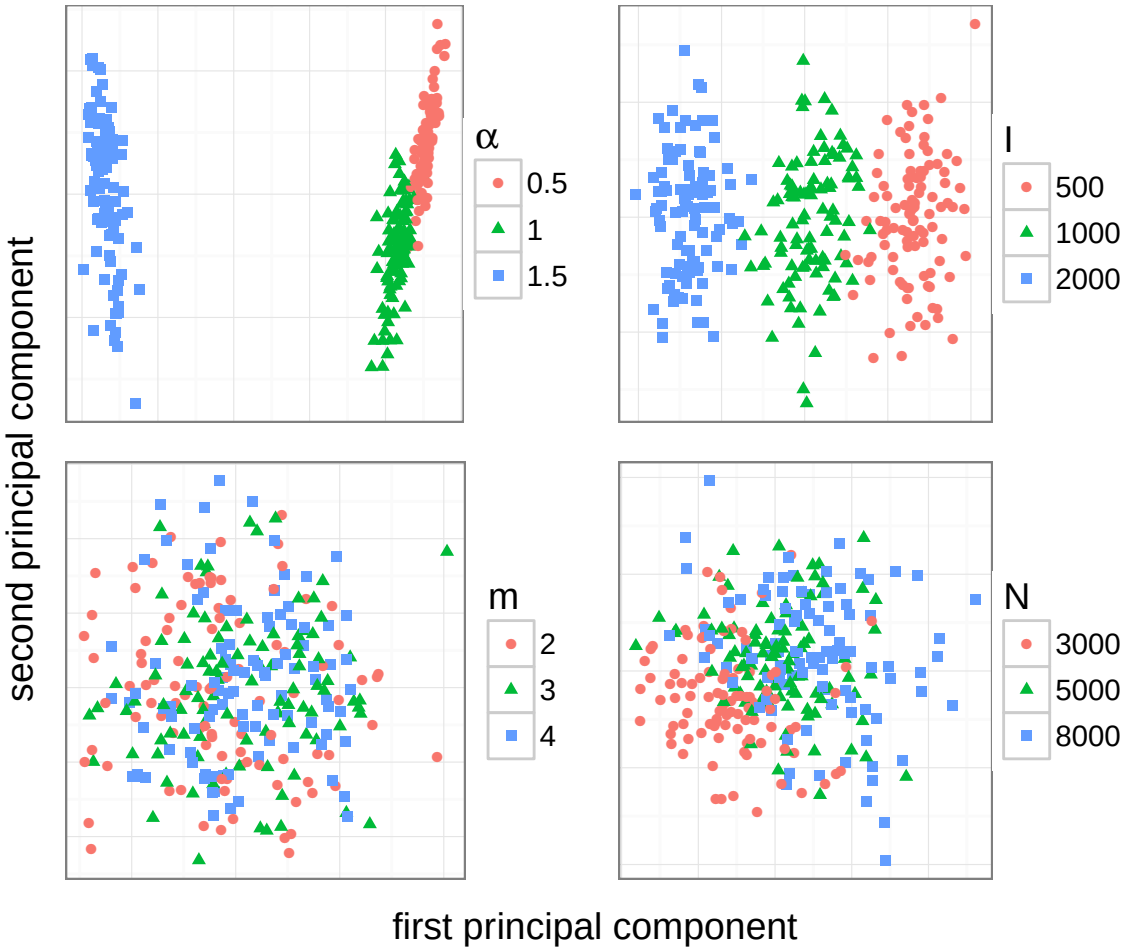


Figure 2.6: Each parameter of the BA model was individually varied to produce 300 simulated trees with 500 tips each. Kernel matrices were formed from all pairwise kernel scores among each set of 300 trees. The trees were projected onto the first two principal components of the kernel matrix calculated using kPCA. The other parameters, which were not varied, were set to $\alpha = 1$, $I = 1000$, $m = 2$, and $N = 5000$. The tree kernel meta-parameters were $\lambda = 0.3$ and $\sigma = 4$.

and right). The results were largely robust to variations in the tree kernel meta-parameters $\lambda$ and $\sigma$, although accuracy varied between different epidemic and sampling scenarios (figs. B.7 to B.10). For all parameters except $m$, the absolute number of tips in the tree had a much greater impact on accuracy than the proportion of infected individuals these tips represented. However, for $m$, both the number and proportion of sampled tips had a strong impact or the accuracy of the kSVM (fig. B.9).

The kSVM classifier for $\alpha$ had an average accuracy of 0.92, compared to 0.6 for the nLTT-based SVM, and 0.77 for Sackin's index. There was little variation about the mean for different tree and epidemic sizes. No classifier could accurately identify $m$ in any epidemic scenario, with average accuracy values of 0.35 for kSVM, 0.32 for the nLTT, and 0.38 for Sackin's index. Again, there was little variation in accuracy between epidemic scenarios, although the accuracy of the kSVM was slightly higher on 1000-tip trees (average 0.33, 0.35, 0.38 for 100, 500, and 1000 tips respectively).

The accuracy of classifiers for $I$ appeared to be strongly influenced by the number of tips in the

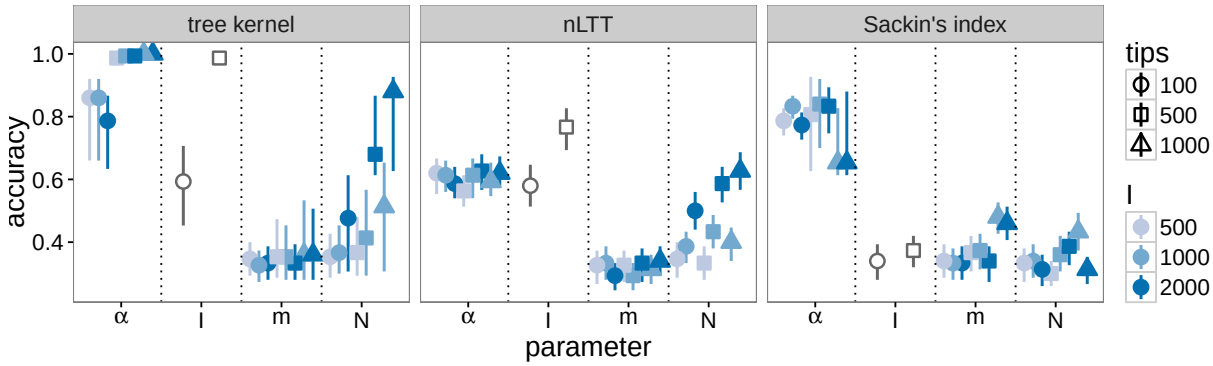


Figure 2.7: Cross-validation accuracy of kernel-SVM classifier (left), SVM classifier using nLTT (centre), and SVM using Sackin's index (right) for BA model parameters. Kernel meta-parameters were set to $\lambda = 0.3$ and $\sigma = 4$. Each point was calculated based on 300 simulated transmission trees over networks with three different values of the parameter being tested, assuming perfect knowledge of the other parameters. Vertical lines are empirical 95% confidence intervals based on 1000 two-fold cross-validations. The classifiers for $I$ were not evaluated with 1000-tip trees, because one of the tested $I$ values was 500, and it is not possible to sample a tree of size 1000 from 500 infected individuals.

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

### Marginal parameter estimates with grid search

Figure 2.8 shows point estimates and 50% and 95% highest density intervals for each of the BA parameters, for one replicate experiment with 500-tip trees. Plots showing the point estimates for all replicates can be found in figs. B.19 to B.22. For all parameters except $m$, the error of point estimates was negatively correlated with the number of sampled tips in the tree (for $\alpha$, $I$, and $N$ respectively: Spearman's $\rho$ = $-0.22$, $-0.51$, $-0.16$; $p$-values $4\times10^{-4}$, $<10^{-5}$, 0.01). The 95% highest density intervals obtained for all parameters were extremely wide, occupying >75% of the grid in all cases (fig. 2.8).

Across all replicates, $R^2$ values for the correlations between the estimated and true values were 0.91, 0.91, 0.25, and 0.54 for $\alpha$, $I$, $m$, and $N$ respectively. The mean absolute errors of the point estimates were 0.14, 310, 1.31, and 2419, representing 7%, 8%, 26%, and 17% of the respective grids. Note that fig. 2.8 contains only one replicate data point per parameter value, out of ten total. For $I$ and $N$, relative errors may be more appropriate to consider. An overestimate of 100 individuals would be very misleading if the true population was only of size 100, but almost negligible in a population of size 5000. The mean relative errors for $I$ and $N$ respectively were 18% and 31%.

Qualitatively, $\alpha$ and $I$ exhibited weak identifiability within particular sections of the grid (fig. 2.8).

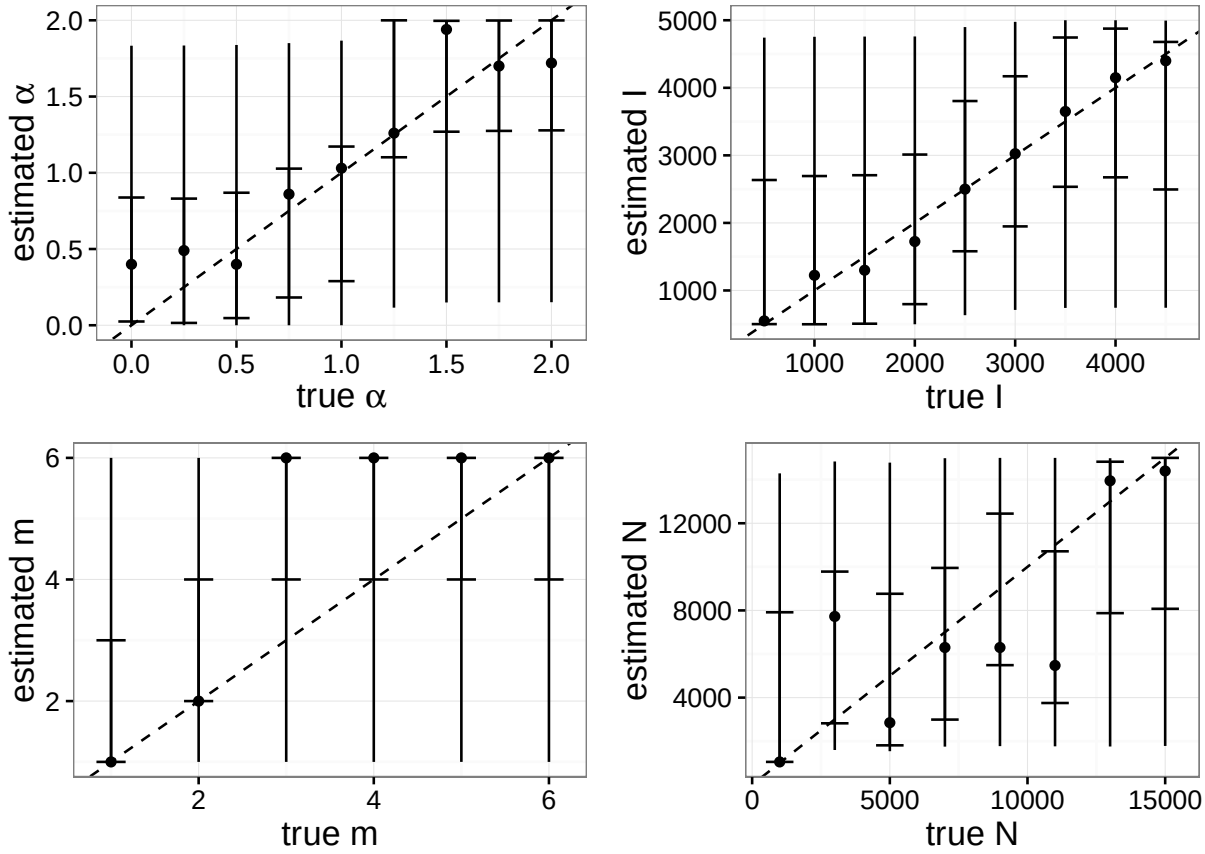


Figure 2.8: Grid search estimates of BA model parameters for one replicate experiment with trees of size 500. Point estimates (dots), 95% HDIs (lines), and 50% HDIs (notches) for each BA model parameter, obtained using grid search. Networks and transmission trees were simulated over a grid of values for each parameter while holding the others fixed to known values. 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.

That is, the posterior distributions for several different grid values were extremely similar. The 50% HDIs for $\alpha$ were similar for the values $\alpha \leq 0.5$ (on average 0.02 - 0.84) and for $\alpha \geq 1.5$ (1.26 - 2). For $I$, similar 50% HPD were observed for $I \leq 1500$ (average [650 - 2846]) and for $I > 3000$ ([2596 - 4802]). No similar patterns were observed for $m$ or $N$ (although the 50% HDIs appear to be identical for $m > 2$ in fig. 2.4, this was not consistent across replicates).

Figures B.15 to B.18 show kernel score distributions of kernel scores along the grid for each parameter. The distributions for some values, such as $\alpha = 1.25$, $I = 500$ and 4500, $m = 1$, and $N = 1000$, exhibited distinct peaks around the true value. This indicates that these values produce distinctively shaped trees that can be identified with the tree kernel, when the other parameter values are fixed and known. However, for the majority of values of each parameter, the score distributions were fairly flat around the true value. This means there is a range of values which produce similarly shaped trees, and the parameter is less identifiable within that range. The exception was $I$, whose score distributions

exhibited a more or less rounded shape with the highest point near the true value.

### Joint parameter estimates with *netabc*

Figure 2.9 shows stratified posterior mean point estimates of the BA model parameters $\alpha$ and $I$ obtained with ABC on simulated data. The parameters $m$ and $N$ were not identifiable with ABC for any parameter combinations (fig. B.23). Average boundaries of 95% HPD intervals are given in table 2.4.

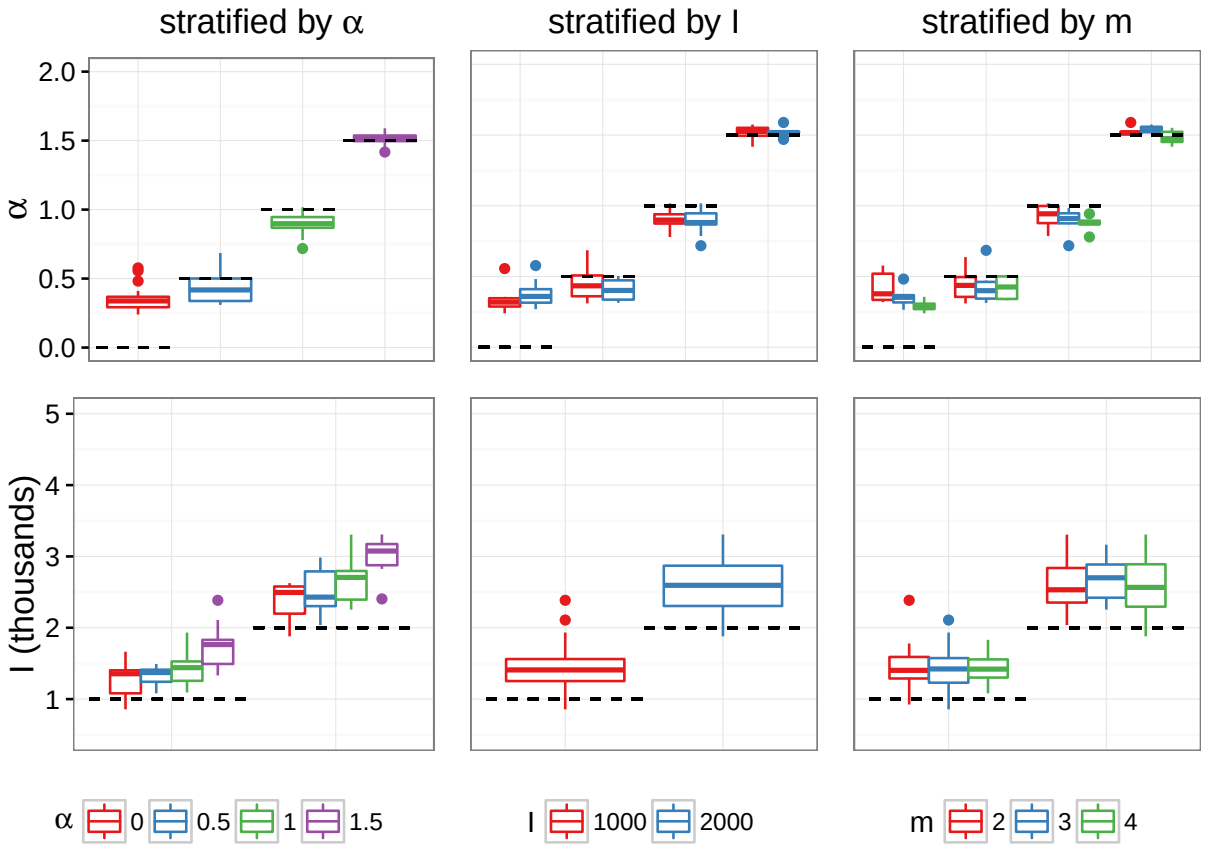


Figure 2.9: Posterior mean point estimates for BA model parameters $\alpha$ and $I$ obtained by running *netabc* on simulated data, stratified by true parameter values. First row of plots contains true versus estimated values of $\alpha$; second row contains true versus estimated values of $I$. Columns are stratified by $\alpha$, $I$, and $m$ respectively. Dashed lines indicate true values.

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

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

Table 2.4: Average posterior mean point estimates and 95% HPD interval widths for BA model parameter estimates obtained with *netabc* on simulated data. Three transmission trees were simulated under each combination of the listed parameter values, and the parameters were estimated with ABC without training.

To analyze the effects of the true parameter values on the accuracy our estimates of $\alpha$ and $I$, we fitted one GLM for each of these two parameters, with error rate as the dependent variable and the true parameter values as independent variables. Since the estimates of $m$ and $N$ were roughly equal across all simulations (fig. B.23), GLMs were not fitted for these parameters. The estimated coefficients are shown in tables 2.5 and 2.6. The GLM were fitted using the inverse link function. That is, if $p$ is the true value of the parameter and $\hat{p}$ is a random variable representing our estimate of the parameter, the GLM posits a relationship of the form

$$\mathbb{E}(|p-\hat{p}|) = (\beta_0 + \beta_\alpha \alpha + \beta_I I + \beta_m m)^{-1},$$

where the $\beta$'s are coefficients to be fitted. If the true value $\alpha$, say, is increased by one, the *inverse* of the expected absolute error will increase by $\beta_\alpha$. If $\beta_\alpha$ is positive, it means that the absolute error decreases as the true value of $\alpha$ increases.

| Parameter | Estimate | Standard error | $p$-value |
|---|---|---|---|
| (Intercept) | 2 | 0.6 | 0.01 |
| $\alpha$ | 10 | 2 | $<10^{-5}$ |
| $I$ | $-3\times10^{-4}$ | $2\times10^{-4}$ | 0.7 |
| $m$ | 0.5 | 0.2 | 0.01 |

Table 2.5: Parameters of a fitted GLM relating error in estimated $\alpha$ to true values of BA parameters. GLM was fitted with a Gaussian distribution and inverse link function. Coefficients are interpretable as additive effects on the inverse of the mean error.

The GLM analysis indicated that the error in estimates of $\alpha$ decreased with larger true values of $\alpha$ ($p<10^{-5}$) and $m$ ($p=0.01$) but was not significantly affected by $I$ (table 2.5). Qualitatively, $\alpha$ seemed to be only weakly identifiable between the values of 0 and 0.5 (fig. 2.9). The error in the estimated prevalence $I$ was slightly lower for smaller values of $\alpha$ ($p<10^{-5}$) and $I$ ($p=0.05$), but was not significantly

| Parameter | Estimate | Standard error | $p$-value |
|---|---|---|---|
| (Intercept) | 0.004 | $5\times10^{-4}$ | $<10^{-5}$ |
| $\alpha$ | $-0.001$ | $2\times10^{-4}$ | $<10^{-5}$ |
| $I$ | $-4\times10^{-7}$ | $2\times10^{-7}$ | 0.05 |
| $m$ | $-7\times10^{-5}$ | $8\times10^{-5}$ | 1 |

Table 2.6: Parameters of a fitted GLM relating error in estimated $I$ to true values of BA parameters. GLM was fitted with a Gaussian distribution and inverse link function. Coefficients are interpretable as additive effects on the inverse of the mean error.

affected by the true value of $m$ (table 2.6).

The dispersion of the ABC approximation to the posterior also varied between the parameters (table 2.4). HPD intervals around $\alpha$ and $I$ were often narrow relative to the region of nonzero prior density, whereas the intervals for $m$ and $N$ were more widely dispersed. Figures 2.10 and 2.11 show one- and two-dimensional marginal distributions for a simulation with relatively low error. Each parameter was chosen based on its mean error rate across all simulations. The parameters for this simulation were $\alpha = 1.5$, $I = 1000$, $m = 4$, and $N = 5000$. Figures 2.12 and 2.13 show the equivalent marginals for a different simulation with relatively high error rates for each individual parameter. The parameters were $\alpha = 0$, $I = 2000$, $m = 2$, and $N = 5000$. The two-dimensional marginals indicate some dependence between pairs of parameters, particularly $I$ and $N$ which show a diagonally shaped region of high posterior density.

To test the effect of model misspecification, we simulated one network where the nodes exhibited heterogeneous preferential attachment power (half 0.5, the other half 1.5), with $m = 2$, $N = 5000$, and $I = 1000$. The posterior mean [95% HPD] estimates for each parameter were: $\alpha$, 1.03 [0.67 - 1.18]; $I$, 1474 [511 - 2990]; $m$, 3 [1 - 5]; $N$, 9861 [3710- 14977]. To test the effect of sampling bias, we sampled one transmission tree in a peer-driven fashion, where the probability to sample a node was twice as high if one of its peers had already been sampled. The parameters for this experiment were $N = 5000$, $m = 2$, $\alpha = 0.5$, and $I = 2000$. The estimated values were: $\alpha$, 0.3 [0 - 0.63]; $I$, 2449 [1417 - 3811]; $m$, 3 [2 - 5]; $N$, 9132 [2852 - 14780]. Both of these results were in line with estimates obtained on other simulated datasets (table 2.4), although the estimate of peer-driven sampling for $\alpha$ was somewhat lower than typical.

Figure 2.14 shows the effect of performing marginal ABC estimation of each of the BA parameters on the same simulated transmission tree. The estimates of $m$ were apparently unaffected by marginalizing out the other parameters, corroborating the previous experiments' findings that $m$ is not an identifiable parameter from scaled tree shapes. Compared to allowing all parameters to vary, estimates of $\alpha$, $I$, and $N$ were improved by 41%, 59%, and 46% when all other parameters were fixed. Figure 2.15 shows the impact of increasing the number of particles, simulated datasets, and $\alpha_{\mathrm{ESS}}$ parameter on the accuracy of a single simulation. The results of the two simulations were similar, but surprisingly, the results with higher SMC settings were slightly worse (by 10%, 8%, and 11% for $\alpha$, $I$, and $N$ respectively). However, the 50% HPD interval for $I$ was closer to the true value of 2000 with the improved settings (2338 - 3423, vs. 2810 - 3767 with basic settings). The estimate of $m$, 3 in both cases, was unaffected by the settings.

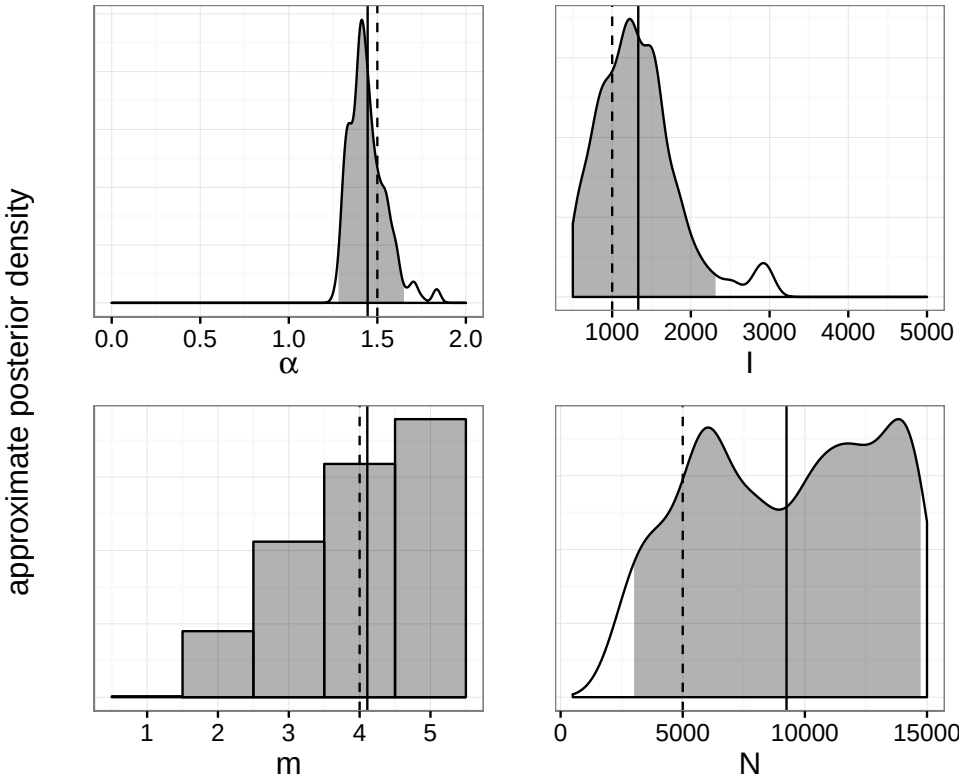


Figure 2.10: One-dimensional marginal posterior distributions of BA model parameters estimated with low error by *netabc* from a simulated transmission tree. Dashed lines indicate true values, solid lines indicate posterior means, and shaded areas show 95% highest posterior density intervals.

## 2.3 Application to real world HIV data

### 2.3.1 Methods

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. To identify datasets fitting these criteria, we used the *Entrez* module in the *BioPython* library [156] to identify all studies which were linked to at least 150 sequences of the same HIV gene in GenBank (635 at the time when the data was collected). We manually curated a subset of these articles which, based on their title and abstract, appeared to have sampled one sequence per individual in a group likely to be epidemiologically related. For example, studies were excluded if they were investigating response to a particular drug, pregnant women or pediatric HIV patients, intra-host evolution, or a multi-region or multi-country cohort. We acknowledge that our perusal of the complete set of articles was rather cursory, often limited to reading the title, and it is quite possible that we failed to identify other studies which would have been suitable to include. Each potential dataset was revisited, and those without sampling time annotation in GenBank were excluded. We also excluded studies where all sequences were sampled at the same timepoint, which was necessary because the method we used to time scale the tree requires

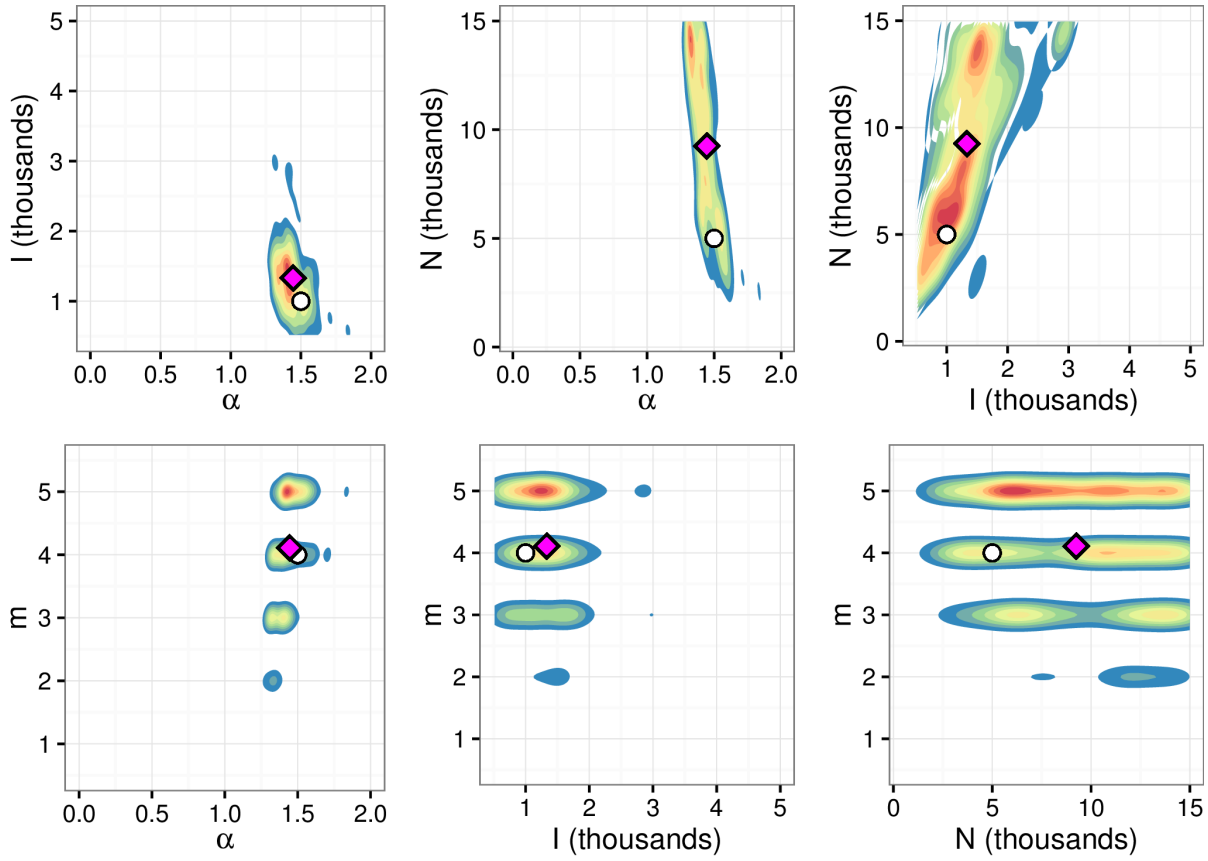


Figure 2.11: Two-dimensional marginal posterior distributions of BA model parameters estimated with low error by *netabc* from a simulated transmission tree. White circles indicate true values, magenta diamonds indicate posterior means.

non-contemporaneous tips. The datasets are summarized in table 2.7. In addition to the published data, we analyzed an in-house dataset sampled from HIV-positive individuals in British Columbia, Canada. For clarity, we will refer to each dataset by its risk group and location of origin in the text. For example, the Zetterberg et al. [157] data will be referred to as IDU/Estonia.

We downloaded all sequences associated with each published study from GenBank. For the IDU/Romania data, only sequences from IDU (whose sequence identifiers included the letters "DU") were included in the analysis. Kao et al. [163] (MSM/Taiwan) found a strong association in their study population between subtype and risk group - subtype B was most often associated with MSM, whereas IDU were usually infected with a circulating recombinant form. Since there were many more subtype B sequences in their data than sequences of other subtypes, we restricted our analysis to the subtype B sequences and labelled this dataset as MSM. Two datasets (HET/Uganda and HET/Malawi) included both *env* and *gag* sequences. Each gene was analyzed separately to assess the robustness of *netabc* to the particular HIV gene sequence used to estimate a transmission tree. The IDU/Estonia data also sequenced both genes, but the highly variable coverage and high homology of the *gag* sequences made it impossible to obtain a sufficiently large block of non-identical sequences to analyze. Therefore, we analyzed only *env* for this

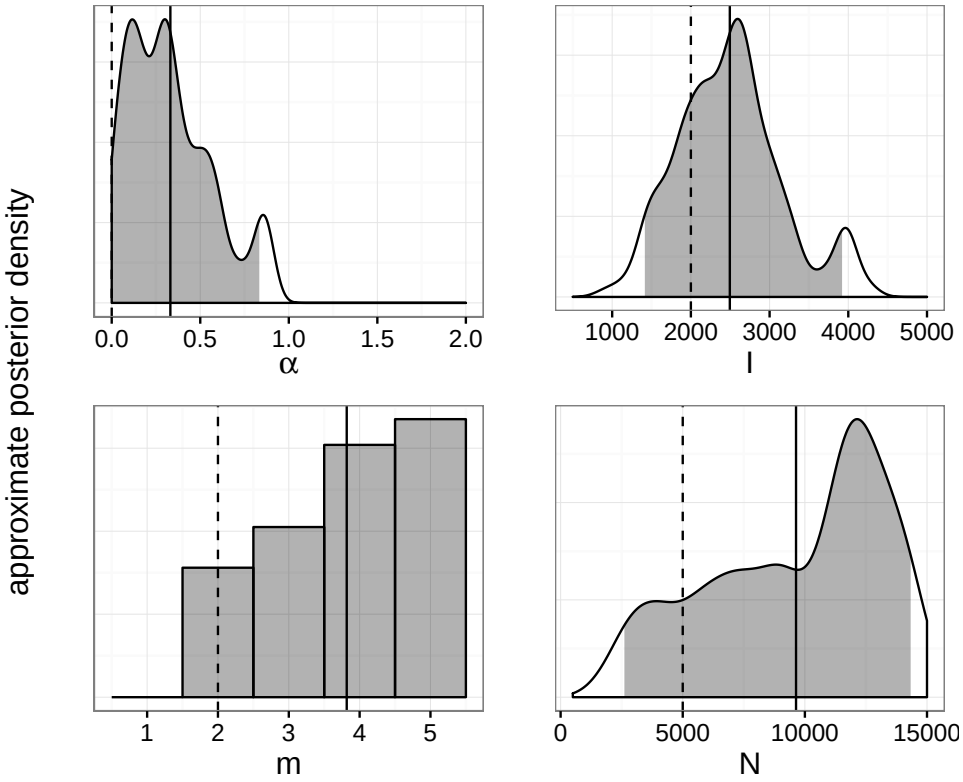


Figure 2.12: One-dimensional marginal posterior distributions of BA model parameters estimated with high error by *netabc* from a simulated transmission tree. Dashed lines indicate true values, solid lines indicate posterior means, and shaded areas show 95% highest posterior density intervals.

dataset.

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+ [156]. Sequences were multiply aligned using *MUSCLE* version 3.8.31 [166], and alignments were manually inspected with *Seaview* version 4.4.2 [167]. Duplicated sequences were removed with *BioPython*. Phylogenies were constructed from the nucleotide alignments by approximate maximum likelihood using *FastTree2* version 2.1.7 [54] with the generalized time-reversible (GTR) model [168]. 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 [169]) as described previously [21]. Due to the removal of duplicated sequences, all estimated transmission trees were fully binary.

To check if our results were robust to the choice of phylogenetic reconstruction method, we built and reanalyzed phylogenies for the datasets with the lowest and highest estimated $\alpha$ values (mixed/Spain and IDU/Estonia) with *RAxML* [55] with the GTR+$\Gamma$ model of sequence evolution and rate heterogeneity. The trees were rooted and time-scaled with *Least Square Dating* [*LSD*, 59]. For expediency, the analysis was run with the prior $m \sim \text{DiscreteUniform}(2,5)$, which defines a smaller total search space than the prior allowing $m = 1$.

Four of the datasets (MSM/Shanghai, HET/Botswana, HET/Uganda, and MSM/USA) were initially

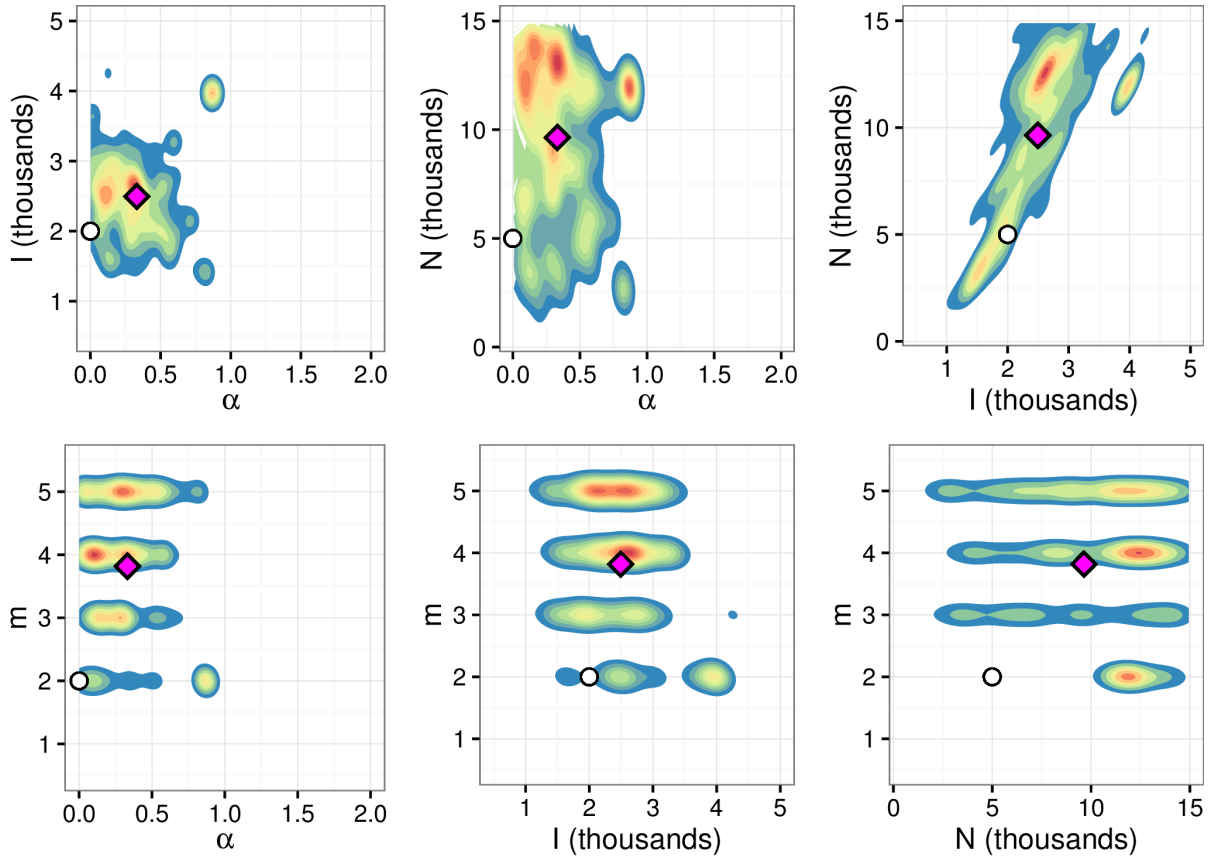


Figure 2.13: Two-dimensional marginal posterior distributions of BA model parameters estimated with high error by *netabc* from a simulated transmission tree. White circles indicate true values, magenta diamonds indicate posterior means.

larger than the others, containing 1265, 1299, 1026/915 (*env*/*gag*), and 648 sequences respectively. To ensure that our results would not be affected by the dataset size, we reduced these to a number of sequences similar to the smaller datasets. For the MSM/Shanghai dataset, we detected a cluster of size 280 using a patriotic distance cutoff of 0.02 as described previously [91]. Only sequences within this cluster were carried forward. For the HET/Uganda, HET/Botswana, and MSM/USA datasets, no large clusters were detected using the same cutoff, so we analyzed subsets of size 225, 180, and 180 respectively. The subset of the HET/Uganda data was chosen by eye such that the individuals were monopolistic in both the *gag* and *env* trees. The other subsets were arbitrarily chosen subtrees from phylogenies of the complete datasets.

For all datasets, we used the priors $\alpha \sim \text{Uniform}(0,2)$ and $N$ and $I$ jointly uniform on the region $\{n \leq N \leq 10000, n \leq I \leq 10000, I \leq N\}$, where $n$ is the number of tips in the tree (see table 2.7). Since the value $m = 1$ produces networks with no cycles, which we considered fairly implausible, we ran one analysis with the prior $m \sim \text{DiscreteUniform}(1,5)$, and one with the prior $m \sim \text{DiscreteUniform}(2,5)$. The other parameters to the SMC algorithm were the same as used for the simulation experiments, except that we used 10000 particles instead of 1000 to increase the accuracy of the estimated posterior.

| Reference | Sequences ($n$) | Location | Risk group | Gene |
|---|---|---|---|---|
| Zetterberg et al. [157] | 171 | Estonia | IDU | *env* |
| Niculescu et al. [158] | 136 | Romania | IDU | *pol* |
| unpublished | 399 | British Columbia, Canada | IDU | *pol* |
| Novitsky et al. [159]<br>Novitsky et al. [160] | 180 | Mochudi, Botswana | HET | *env* |
| McCormack et al. [161] | 141/154 | Karonga District, Malawi | HET | *env*/*gag* |
| Grabowski et al. [162] | 225 | Rakai District, Uganda | HET | *env*/*gag* |
| Wang et al. [29] | 173 | Beijing, China | MSM | *pol* |
| Kao et al. [163] | 275 | Taiwan | MSM | *pol* |
| Little et al. [28] | 180 | San Fransisco, USA | MSM | *pol* |
| Li et al. [164] | 280 | Shanghai, China | MSM | *pol* |
| Cuevas et al. [165] | 287 | Basque Country, Spain | mixed | *pol* |

Table 2.7: Characteristics of published HIV datasets analyzed with *netabc*. Abbreviations: MSM, men who have sex with men; HET, heterosexual; IDU, injection drug users. The HET data were sampled from a primarily heterosexual risk environment but did not explicitly exclude other risk factors. The number of sequences column indicates how many sequences were included in our analysis; there may have been additional sequences linked to the study which we excluded for various reasons (see methods).

This was computationally feasible due to the small number of runs required for this analysis.

Empirical studies of contact networks often report the exponent $\gamma$ of the power law degree distribution [22–25, 95, 144]. Although the BA model does not produce networks with true power law degree distributions except when $\alpha = 1$, the power law still provides a reasonable approximation to the slope when $\alpha \neq 1$ (fig. B.25). To compare our results to existing network literature we simulated 100 networks each according to the posterior mean parameter estimates obtained for each investigated dataset. Although the BA model does not produce power law networks except when $\alpha = 1$, simulations show that the power law fit still captures the slope of the degree distribution reasonably well (fig. B.25). $\gamma$ was calculated for each network using the *fit_power_law* function in *igraph*, with the 'R.mle' implementation. The median of the 100 $\gamma$ values was taken as a point estimate for the associated dataset.

### 2.3.2 Results

Posterior mean point estimates and 50% and 95% HPD intervals for each parameter are shown in fig. 2.16. Figure B.26 shows point estimates and HPD intervals obtained when the value $m = 1$ was disallowed by the prior. Since the results indicated that $m = 1$ was the most credible value for several datasets, all results discussed henceforth are for the prior $m \sim \text{DiscreteUniform}(1,5)$ unless otherwise stated.

Posterior mean point estimates for the preferential attachment power $\alpha$ were all sub-linear, ranging from 0.83 for the IDU/Estonia data to 0.27 for the mixed/Spain data. When aggregated by risk group, the average estimates were 0.73 for IDU, 0.41 for primarily heterosexual risk, and 0.37 for MSM. These values were obtained with *gag* for the datasets where both *gag* and *env* were sequenced, but the estimates for $\alpha$ did not change appreciably between the two genes (fig. 2.17). There was a large

amount of uncertainty associated with these estimates. 95% HPD intervals were very wide for most datasets, often encompassing almost the entire range from 0 to 1 (fig. 2.16). For all the datasets except HET/Malawi and MSM/Beijing, the posterior mean was either outside, or very close to the border of, the 50% HPD intervals. This indicates that the posterior distributions were diffuse with heavy tails, rather than having most of their mass around the mode.

For all but the HET/Botswana data, the posterior mean estimates for the prevalence $I$ were between 373 (IDU/Estonia) and 1391 (MSM/Shanghai). The HET/Botswana data had a much higher estimated $I$ value (5432) than the other datasets, with a very wide 95% HPD interval covering almost the entire prior region (fig. 2.16). There was no significant correlation between the number of sequences in the tree and the estimated prevalence (Spearman correlation, $p = 0.7$), indicating that the higher prevalence estimates were not simply due to increased sampling density. When both *gag* and *env* sequences were analyzed, the estimates from the *env* data were higher (HET/Uganda, 939 for *gag* vs. 1615 for *env*; HET/Malawi, 724 for *gag* vs. 845 for *env*). As with $\alpha$, many of the posterior means were outside the 50% HPD intervals, suggesting diffuse posterior distributions.

The posterior means of $m$ were equal to one for zero of the datasets analyzed. The widths of the 95% HPD intervals varied from 0 (all the mass on the estimated value) to 5 (the entire prior region). Estimates of $N$ were mostly uninformative, with very similar estimates for all datasets (mean 6212, range 5881 - 6882). This was similar to the pattern observed for the synthetic data, where the posterior mean always fell around the upper two-thirds mark of the region allowed by the prior (fig. B.23).

When the value $m = 1$ was disallowed by the prior, the separation in $\alpha$ between the IDU datasets and the others became more striking (fig. B.26). All three IDU datasets had estimated $\alpha$ values at or above 1. The estimate for the MSM/Beijing data was slightly lower (0.85) and the estimates for the seven remaining non-IDU datasets were bounded above by 0.58. The values of $I$ were fairly robust to the choice of prior (compare figs. 2.16 and B.26), although the 95% HPD intervals were slightly narrower (average width 2244 for $m \geq 1$ and 1848 for $m \geq 2$). The posterior means of $m$ for all but the HET/Botswana data took on the value 3 with this prior, with the HPD intervals spanning the entire prior region. This is very similar to the results observed for $m$ on simulated data (table 2.4), and suggests that $m$ is not identifiable from these data with this prior. The results for $N$ did not change appreciably between the two choices of prior.

For the two datasets we reanalyzed using RAxML [55] and LSD [59], $\alpha$ was relatively robust to the choice of method (**??**, posterior means 0.48 vs. 0.48 for mixed/Spain and 1.02 vs. 1.12 for IDU/Estonia). However, the estimates of $I$ were about twice as high when RAxML was used instead of FastTree to reconstruct the trees (228 vs. 437 for IDU/Estonia, 816 vs. 1949 for mixed/Spain).

To compare our results to the existing network literature, we estimated values of the power law exponent $\gamma$ for each of the datasets investigated, using the posterior median estimates. The results are summarized in table 2.8. All of the estimated exponents were in the range $2 \leq \gamma \leq 2.5$, which is on the lower end of the range $2 \leq \gamma \leq 4$ reported in the literature. Because the estimates of $N$ were uninformative, we also simulated another set of networks with the estimated $I$ value, rather than $N$, as the number of nodes. However, none of the estimated $\gamma$ values were changed by this operation (not shown).

| Dataset | $\gamma$ |
|---|---|
| mixed/Spain (Cuevas et al. 2009) | 2.09 |
| MSM/Shanghai (Li et al. 2015) | 2.10 |
| MSM/USA (Little et al. 2014) | 2.10 |
| MSM/Taiwan (Kao et al. 2011) | 2.42 |
| MSM/Beijing (Wang et al. 2015) | 2.12 |
| HET/Uganda (Grabowski et al. 2014) | 2.09 |
| HET/Malawi (McCormack et al. 2002) | 2.32 |
| HET/Botswana (Novitsky et al. 2013 & 2014) | 2.44 |
| IDU/Canada (unpublished) | 2.12 |
| IDU/Romania (Niculescu et al. 2015) | 2.14 |
| IDU/Estonia (Zetterberg et al. 2004) | 2.41 |

Table 2.8: Estimated power law exponents for six HIV datasets based on posterior mean point estimates of BA model parameters. 100 networks were simulated using posterior mean parameter estimates obtained with *netabc*. The power law exponent $\gamma$ was estimated for each, and the median of those estimates was used as a point estimate for the corresponding dataset.

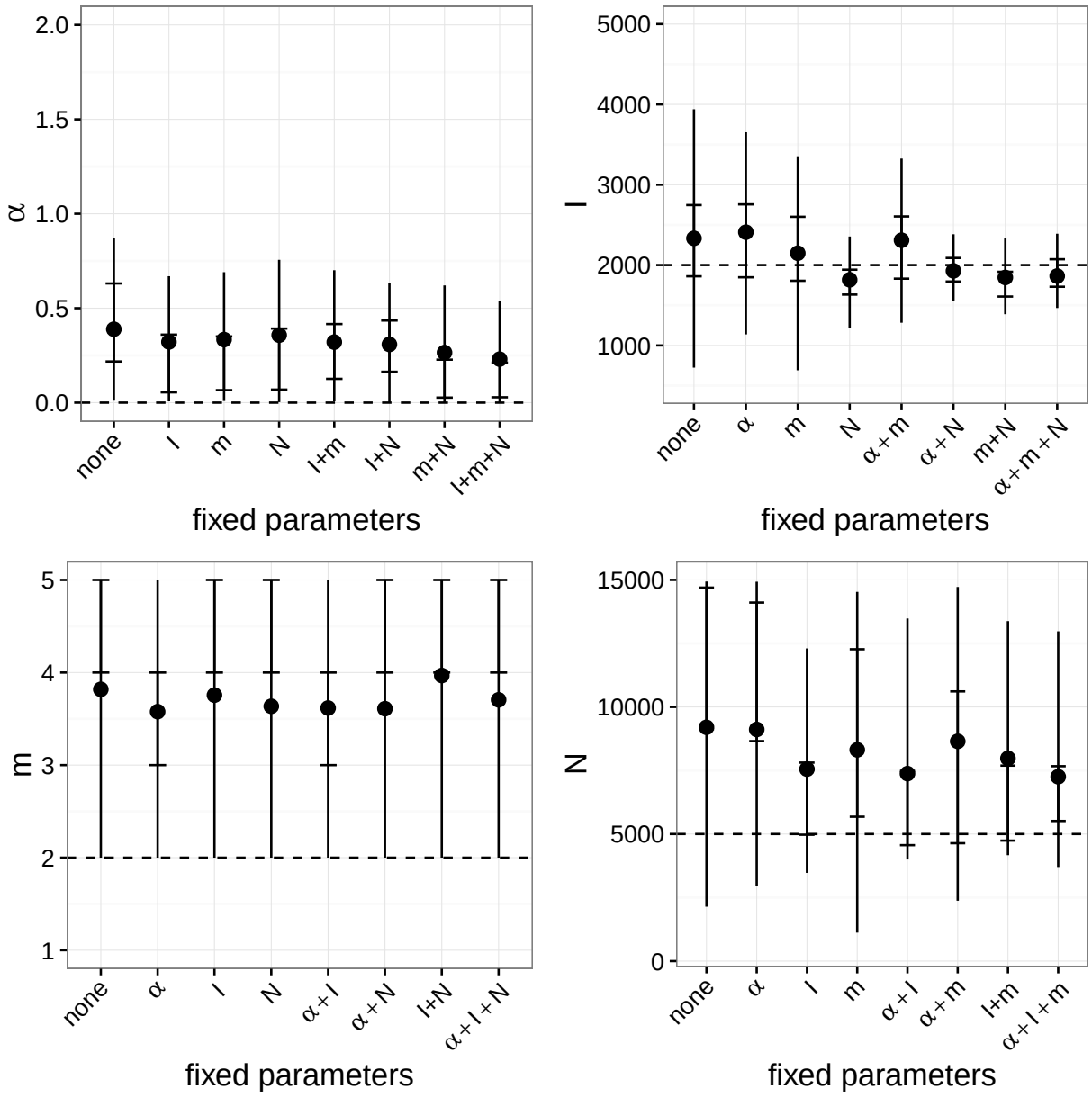


Figure 2.14: Marginal posterior mean estimates (points), 50% HPD intervals (notches), and 95% HPD intervals (lines) of BA model parameters when some parameters ($x$-axis) were fixed. Parameters were fixed by specifying a Dirac-delta prior at the true value. True values are indicated by horizontal dashed lines.

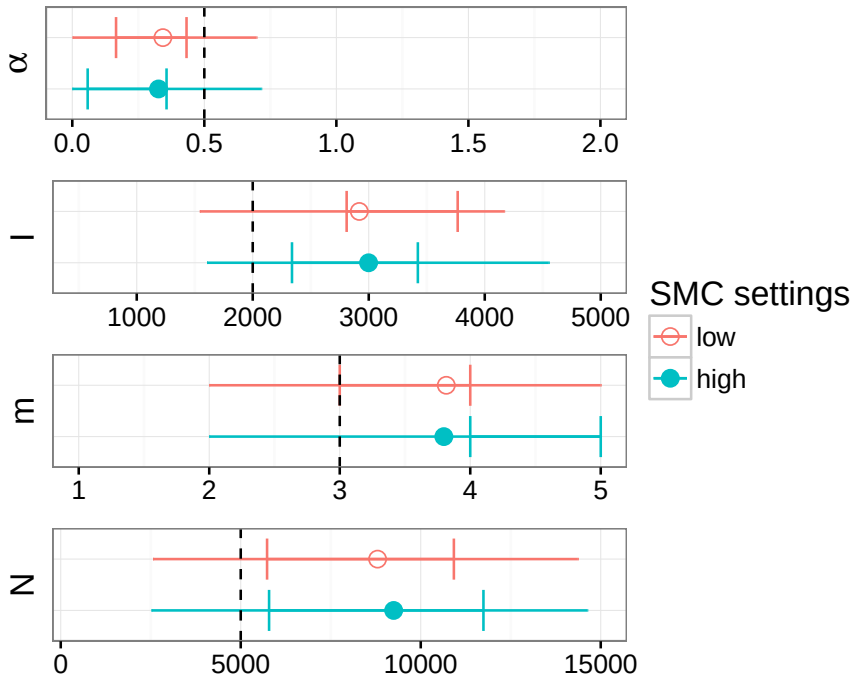


Figure 2.15: Comparison of BA parameter estimates obtained with *netabc* on the same dataset with different SMC settings. Points are posterior means, notches are 50% HPD intervals, and lines are 95% HPD intervals.

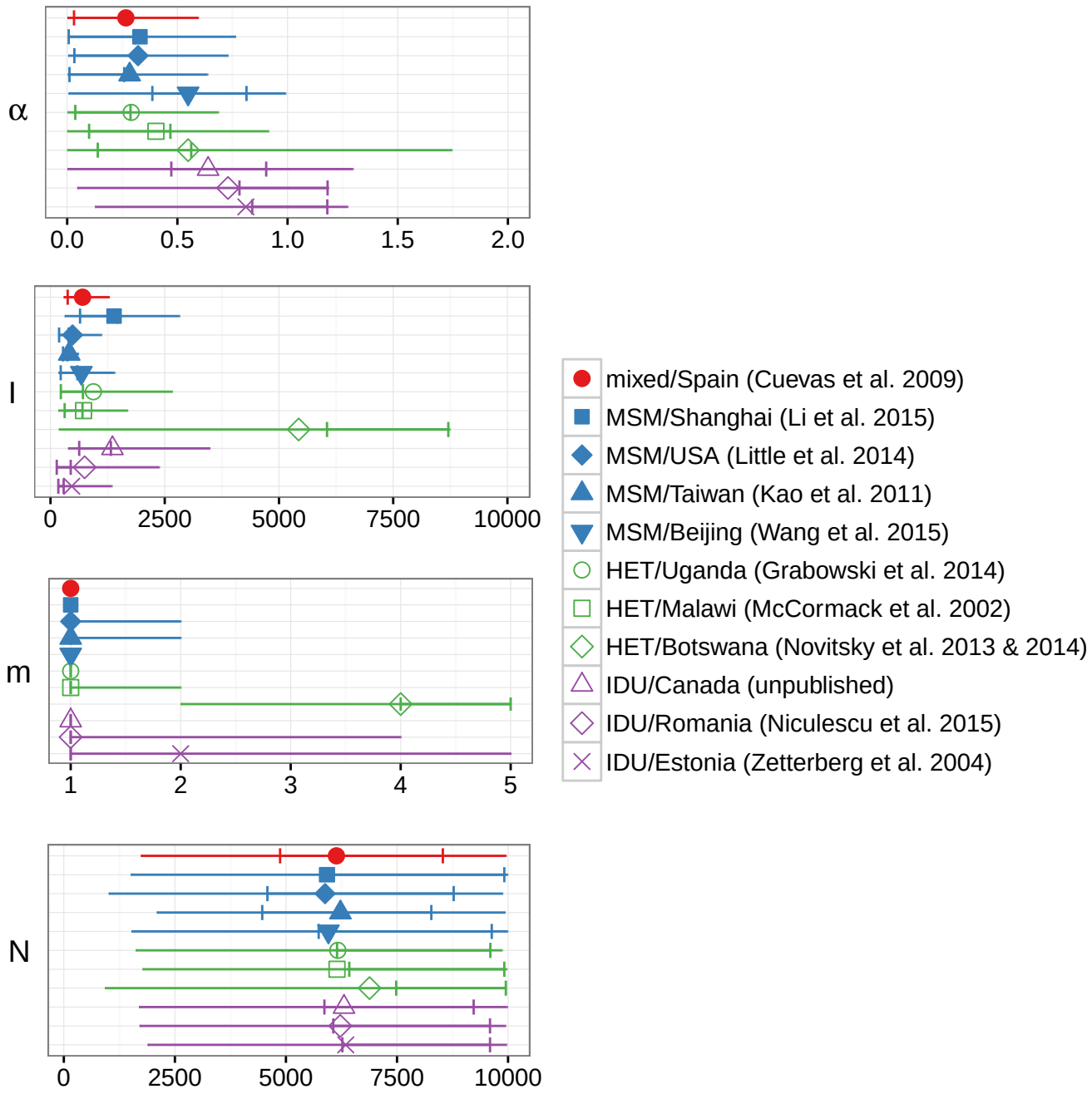


Figure 2.16: Posterior means (points), 50% HPD intervals (notches), and 95% HPD intervals (lines) for parameters of the BA network model, fitted to eleven HIV datasets with *netabc*. Legend labels indicate risk group and country of origin. Abbreviations: IDU, injection drug users; MSM, men who have sex with men; HET, heterosexual. Note that posterior means can fall outside of HPD intervals if the distribution is diffuse.

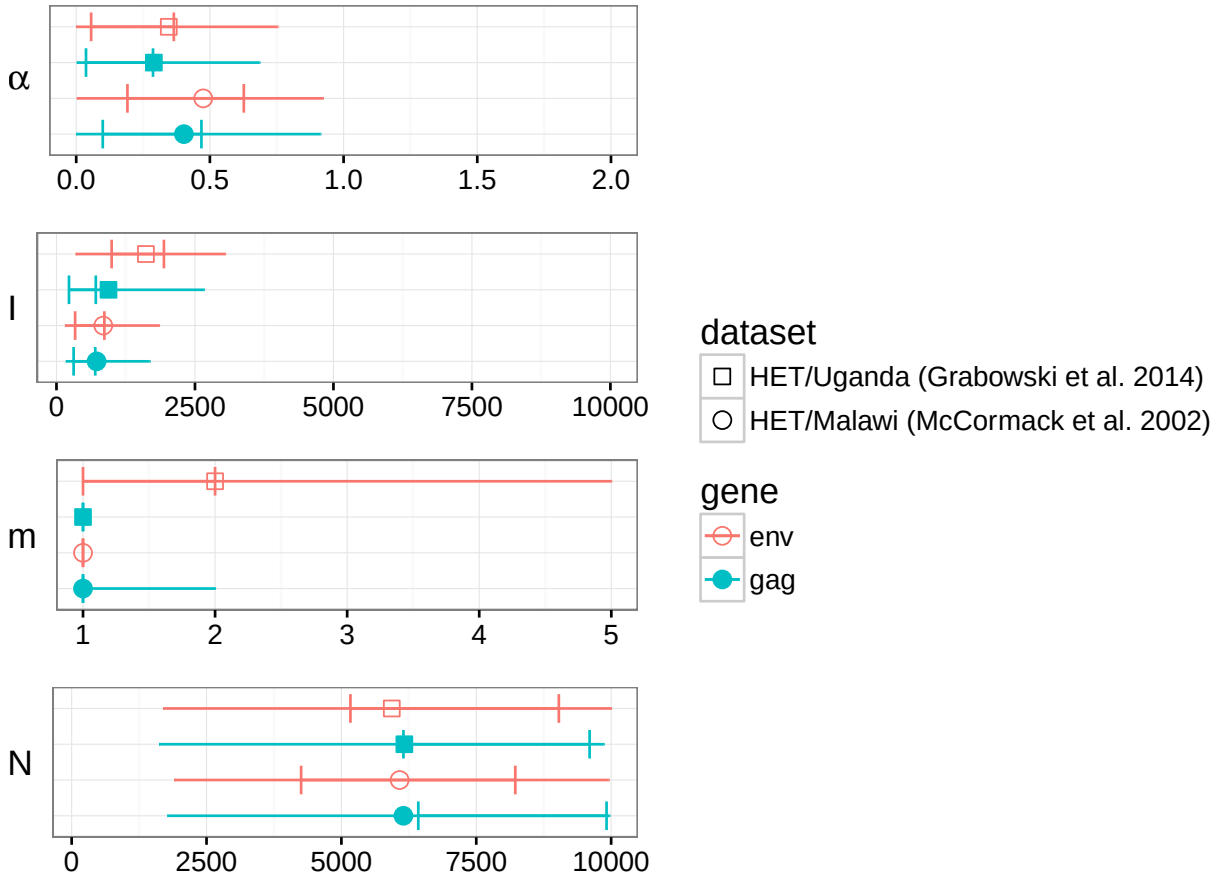


Figure 2.17: Posterior means (points), 50% HPD intervals (notches), and 95% HPD intervals (lines) for parameters of the BA network model, fitted to two HIV datasets where both *gag* and *env* genes were sequenced.

## 2.4 Discussion

### 2.4.1 *Netabc*: uses, limitations, and possible extensions

Contact networks can have a strong influence on epidemic progression, and are potentially useful as a public health tool [28, 29]. Despite this, few methods exist for investigating contact network parameters in a phylodynamic framework [although see 86, 95, 97, 106, 170, for related work]. Kernel-ABC is a model-agnostic method which can be used to investigate any quantity that affects tree shape [21]. In this work, we developed *netabc*, a method based on kernel-ABC to infer the parameters of a contact network model. The method can be used to infer parameters of any network model, as long as it allows simulated networks to be easily generated. We have included generators for the BA model discussed here, as well as the Erdős-Rényi (ER) [171] and Watts-Strogatz (WS) [87] 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.

Several alternative network models and modelling frameworks have been developed which may provide useful future targets for ABC. Waring models [26, 172] are a more flexible type of preferential attachment model which permit a subset of attachments to be formed non-preferentially. These models were used by Leigh Brown et al. [95] to characterize the transmission network in the United Kingdom. Exponential random graph models (ERGMs) [173] are a flexible and expressive parameterization of contact networks in terms of statistics of network features such as pairs and triads. Goodreau [112] evaluated the effect of several different ERGM parameterizations on transmission tree shape and effective population size, and suggested the use of ERGMs as a general framework for estimation of epidemiological quantities related to HIV transmission. Except for a few special cases, simulating a network according to an ERGM generally requires MCMC, which would be too computationally intensive to integrate into *netabc* as it currently stands. To fit ERGM with ABC, one possibility would be to consider the network itself as a parameter to be modified by the MCMC kernel. Other network modelling frameworks include the partnership-centric formulation developed by Eames and Keeling [174] and the log-linear adjacency matrix parameterization applied by Morris [77].

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 false, as people make and break partnerships on a regular basis. Addressing the impact of this simplifying assumption is outside the scope of this work. However, the same assumption is made by most studies using contact network models in an epidemiological context [6, 39]. In principle, ABC could be adapted to dynamic contact networks by using a method such as that developed by Robinson, Cohen, and Colijn [175] to simulate such a network, while concurrently simulating the spread of an epidemic.

It is important to note that *netabc* takes a transmission tree as input, rather than a viral phylogeny. In reality, true transmission trees are not available and must be estimated; these estimates are often based on the viral phylogeny. Although this has been demonstrated to be a fair approximation [e.g. 62], and is

frequently used in practice [e.g. 37], the topologies of a viral phylogeny and transmission tree can differ significantly [36, 53] due to within-host evolution and the sampling process. We have left the estimation of a transmission tree up to the user. 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. Indeed, progress has very recently been demonstrated on this front in a talk by Giardina [176], who have independently developed in a method similar to ours to fit contact network models to phylogenetic data that additionally incorporates a within-host evolutionary model. 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. In addition, there are a number of different methods available for inferring transmission trees [31, 53, 64, 65, 67], 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 method of constructing a transmission tree.

Our implementation of SMC uses a simple multinomial scheme to sample particles from the population according to their weights. Several other sampling strategies have been developed [124], and it is possible that the use of a more sophisticated technique might increase the algorithm's accuracy. Finally, the ABC-SMC algorithm is computationally intensive, taking about a day when run on 20 cores in parallel with the settings described in the methods section. Implementing parallelization using message passing interface (MPI), rather than POSIX threads as we have done here, would allow the program to be run over a larger number of cores on multiple CPUs in parallel.

### 2.4.2 Analysis of Barabási-Albert model with synthetic data

The preferential attachment power $\alpha$ had a very strong influence on tree shape in the range of values we considered (figs. 2.5 and 2.6). Although the tree kernel was the most effective classifier for $\alpha$, a Sackin's index of tree imbalance performed nearly as well (fig. 2.7). This result was intuitive: high $\alpha$ 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 (fig. 2.5). There appeared to be weaker identifiability for $\alpha < 1$ than for $\alpha \geq 1$ (fig. 2.9 and table 2.4), which may be partially explained by the relationship between $\alpha$ and the power law exponent $\gamma$ (fig. B.24). Although the degree distributions do not truly follow a power law for $\alpha \neq 1$, the fitted exponent still captures the slope of the degree distribution reasonably well (fig. B.25), and distributions with similar fitted $\gamma$ values appear qualitatively similar in other respects. The $\gamma$ values fitted to $\alpha = 0$ and $\alpha = 0.5$ are nearly identical (about $2.28 for \alpha = 0$ and $2.33 for \alpha = 0.5$ with $N = 5000$ and $m = 2$). In other words, the degree distributions of networks with $\alpha < 1$ are similar to each other, which may result in similarity of corresponding transmission trees as well.

The $I$ parameter, representing the prevalence at the time of sampling, was also generally identifiable, although it was slightly over-estimated for both cases we considered with ABC. Sackin's index was better able to capture the effect of $I$ on tree shape than the nLTT (figs. 2.7 and B.8), suggesting that this parameter impacts the distribution of branching times in the tree more than the topology. In a homogeneously-mixed population, branching times can be modelled by the coalescent process [46], in

our case under the SI model [52]. Although our populations are not homogeneously mixed, the forces which affect the distributions of branching times still apply. In our simulations, we assumed that all discordant edges shared the same transmission rate, so that the waiting time until the next transmission in the entire network was always inversely proportional to the number of discordant edges. In the initial phase of the epidemic, when $I$ is small, each new transmission results in many new discordant edges. Hence, there is an early exponential growth phase, producing many short branches near the root of the tree. As the epidemic gets closer to saturating the network, the number of discordant edges decays, causing longer waiting times.

The number of nodes in the network, $N$, exhibited the most variation in terms of its effect on tree shape. There was almost no difference between trees simulated under different $N$ values when the number of infected nodes $I$ was small (figs. B.10 and B.14). 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. In fact, this is why the degree distributions of these networks are called "scale-free." 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 relative to $N$, the coalescent dynamics discussed above also apply. The waiting times until the next infection increase, resulting in longer coalescence times toward the tips. The relative accuracy of the nLTT in these situations (figs. 2.7 and B.10) corroborates this hypothesis, as the nLTT uses only information about the coalescence times. When all BA parameters were simultaneously estimated with ABC, $N$ was nearly always over-estimated by approximately a factor of two (fig. B.23 and table 2.4). One factor which may have contributed to this bias was our choice of prior distribution. Since the prior for $I$ and $N$ was jointly uniform on a region where $I \leq N$, more prior weight was assigned to higher $N$ values. Another contributing factor relates to the dynamics of the SI model and the coalescent process and is discussed below.

$I$ and $N$ were both systematically over-estimated by *netabc*, although the bias was more severe for $N$ than for $I$. The number of infected individuals follows a logistic growth curve under the SI model. This kind of growth curve has three qualitative phases: a slow ramp-up, an exponential growth phase, and a slow final phase when the susceptible population is almost depleted. The waiting times until the next transmission, which determine the coalescence times in the tree, are dependent on the growth phase of the epidemic. Therefore, we hypothesize that it is the growth phase at the time of sampling which most affects tree shape, rather than the specific values of $I$ or $N$. To investigate this hypothesis, we simulated transmission trees over networks on a grid of $I$ and $N$ values in the region of uniform prior density. We fit logistic growth curves to the proportion of infected individuals over time, and calculated the first and second derivatives of these curves (with respect to time) at the time of transmission tree sampling. These derivatives give us an indication of the growth rates of the epidemics at the time of sampling. As shown in fig. 2.18, there are bands along which both derivatives are similar which contain the values we tested. These bands span mostly higher values of $N$ and $I$ than the true values. Therefore, if $N$ and $I$ are free to vary (as is the case in ABC), and our hypothesis is true, both parameters will tend to be overestimated

due to being less identifiable within their own band. However, when $N$ is fixed at 5000, the derivatives vary substantially along the $I$-axis, which explains why the grid search estimates of $I$ were accurate and unbiased (figs. 2.8 and B.20). We also note the resemblance of the contour surface of fig. 2.18 to the two-dimensional marginal posterior distributions on $I$ and $N$ obtained with simulated data (figs. 2.11 and 2.13). In addition, both the accuracy and precision of the ABC estimates of $I$ improved substantially when $N$ was marginalized out (fig. 2.14).

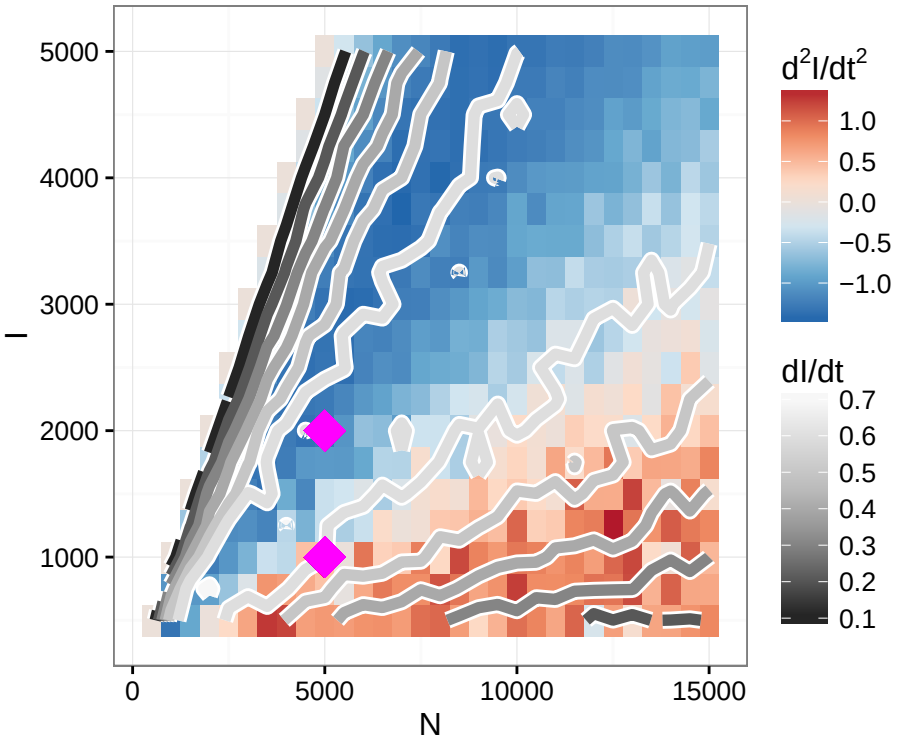


Figure 2.18: First and second time derivatives of epidemic growth curves at time of sampling for various values of $I$ and $N$. Networks were simulated under the BA model with $\alpha = 1.0$, $m = 2$, and $N$ varied along the values shown on the $x$-axis. Transmission trees were sampled at the time when $I$ nodes were infected ($y$-axis). Logistic growth curves were fit to epidemic trajectories derived from the transmission trees, and their first and second derivatives were calculated at the time of sampling. Contours show first derivatives, while colours indicate second derivatives. Values of $I$ and $N$ used in simulation experiments with ABC are indicated by diamonds.

The $m$ parameter, which controls the number of connections added to the network per vertex, did not have a measurable impact on tree shape and was not identifiable with ABC, nor with grid search when the other parameters were held fixed. The exception to this was the value $m = 1$, which produces networks without cycles whose associated trees were more easily distinguished (figs. 2.8 and B.17). 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 $m$. It was also pointed out to us by an anonymous reviewer that for a fixed $I$, an

infected node may only end up transmitting along a fraction of its outgoing edges, which could mask the presence of the extra edges associated with higher $m$.

In addition to the tree height, many summary statistics have been developed to capture particular details of tree shape [177]. 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 [126, 178], rather than a kernel-based similarity score as we have done here. Methods have been developed to select an optimal combination of summary statistics for a given inference task [147]. 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 BA 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 [160].

For $\alpha$ and $I$, the credible intervals attained from the marginal ABC target distributions were narrower than those obtained through grid search. 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. Our method of finding credible intervals from kernel scores along the grid, 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.

As noted by Lintusaari et al. [179], uniform priors on model parameters may translate to highly informative priors on quantities of interest. We observed a non-linear relationship between the preferential attachment power $\alpha$ and the fitted power law exponent $\gamma$ (fig. B.24, although see the discussion in section 1.3.2). Therefore, placing a uniform prior on $\alpha$ between 0 and 2 is equivalent to placing an informative prior that $\gamma$ is close to 2. If we were primarily interested in $\gamma$ rather than $\alpha$, a more sensible choice of prior might have a shape informed by fig. B.24 and be bounded above by approximately $\alpha = 1.5$. This would bound $\gamma$ in the region $2 \leq \gamma \leq 4$ commonly reported in the network literature [22–25, 95]. We note however that Jones and Handcock [180] estimated $\gamma$ values greater than four for some datasets, in one case as high as 17, indicating that consideration of a wider range of permitted $\gamma$ values may be warranted.

The combination of method, model, and priors we employed did not produce perfect estimates of any of the parameters. The estimates of $\alpha$ were the most accurate, although the variance of the estimates was high and the confidence intervals were wide for $\alpha < 1$ (fig. 2.9 and table 2.4). The estimates of $N$ and $I$ were both upwardly biased, and the estimates of $m$ were largely uninformative. Despite these limitations, a major result of our investigation is that some contact network parameters have a measurable impact on tree shape which can be used to perform statistical inference. Further refinements to *netabc*, as well as the use of more sophisticated network models, may improve the accuracy and precision of these estimates.

### 2.4.3 Application to real world HIV data

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

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

For all datasets we examined, the posterior mean estimates for $\alpha$ were sub-linear, ranging from 0.27 to 0.83. The sub-linearity is consistent with the results of de Blasio, Svensson, and Liljeros [105], who developed a statistical inference method to estimate the parameters of a more sophisticated preferential attachment model incorporating heterogeneous node behaviour. When used to analyze population-level longitudinal partner count data, they found $\alpha$ values ranging from 0.26 to 0.62 depending on the gender and time period considered. Our estimates of $\alpha$ for the IDU/Romania was above this range under both priors, as were the estimates for the MSM/Beijing data and the BC data when $m = 1$ was disallowed by the prior. The dataset investigated by de Blasio, Svensson, and Liljeros [105] was derived from a survey of a random sample of the Norwegian population, whereas our investigation focused on datasets from known phylogenetic or geographic clusters of HIV-infected persons. It is therefore unsurprising that we detected stronger preferential attachment dynamics in some cases. For instance, random sampling is much less likely to discover the high-degree nodes characterizing the tail of the degree distribution, simply because those individuals are rare in the general population. In addition, it is plausible that HIV-positive individuals are more likely to be highly connected in their sexual networks, as the odds of acquiring HIV increase with the number of unprotected sexual contacts.

Both de Blasio, Svensson, and Liljeros [105] and the HET/Botswana data studied populations whose primary risk factor for HIV infection was heterosexual contact. de Blasio, Svensson, and Liljeros explicitly excluded reported homosexual contacts; Novitsky et al. did not, but noted that heterosexual contact is the primary mode of transmission in Botswana where the study was done. In the first of the two papers describing the Botswana study [159], the authors noted that their sample was gender-biased, being composed of approximately 75% women. Our estimate of $\alpha$ for these data was 0.55 or 0.53, depending on the prior on $m$. Similarly, de Blasio, Svensson, and Liljeros [105] estimated 0.54, 0.57, and 0.29 for 3-year, 5-year, and lifetime partnership networks respectively for the female portion of their sample.

For both choices of prior on $m$, the datasets derived from IDU populations had a higher estimated preferential attachment power than the other datasets (figs. 2.16 and B.26). This finding is in line with Dombrowski et al. [103], who reanalyzed a network of IDUs in Brooklyn, USA, collected between 1991 and 1993 [181]. They found that the IDU network resembled a BA network much more closely than other social and sexual networks, and offered sociological explanations for the apparent preferential attachment dynamics in this population. Importantly, from a public health perspective, the authors asserted that the removal of *random* individuals from IDU networks may have the paradoxical effect of increasing the network's epidemic susceptibility. When low-degree nodes are removed, as would occur during a police crackdown, their network neighbours may turn to well-known community members for advice or supplies, thus increasing the connectivity of these high-degree nodes.

Unfortunately, the sub-linear region for $\alpha$ identified by both de Blasio, Svensson, and Liljeros [105] and *netabc* is also the region of poorest identifiability (figs. 2.8 and 2.9). This was reflected in the high level of uncertainty in the estimates, with most 95% HPD intervals covering the majority of the range [0, 1]. The value $\alpha = 0.5$ was contained in the 95% HPD interval for every dataset; consequently, it is not possible to say with high confidence that any of the $\alpha$ values are different from each other. In synthetic data, the confidence intervals around $\alpha$ narrowed when other parameters were marginalized out (fig. 2.14). Thus, it is possible that estimates of $\alpha$ could be made more precise by specifying either exact values or informative priors on the other BA parameters. We argued informally in section 2.2.1 that $I$, $m$, and $N$ can sometimes be estimated by standard epidemiological methods, although the difficulty may depend on the specific epidemic under consideration.

**Power law exponent is within reported range, but hard to interpret**

In order to compare our results to existing literature on networks and distributions of partner counts, we have reported estimated values for the power law exponent $\gamma$ of the real data sets we evaluated. However, the posterior means for $\alpha$ for all six datasets were less than one; the degree distributions in this parameter range are stretched exponential, not power law [104]. As we show in fig. B.25, the power law fit does capture the slope of the degree distribution fairly well, but the results should still be interpreted cautiously. Krapivsky, Redner, and Leyvraz [104] showed that the power law distribution can be maintained, with $\gamma$ tuned to any desired value, by a straightforward modification of the BA model. The authors define the "connection kernel" $A_k$ the probability of a new connection to a node of degree $k$, up to a normalizing constant. In the BA model as we have presented it here, $A_k = k^\alpha + 1$. Taking $A_k$ as any asymptotically linear function will result in a power law distribution, with the exponent $\gamma$ determined by the properties of $A_k$. Implementing such a model would be straightforward and seems a natural next step toward improving the realism of the BA model.

Table 2.9 shows some estimated $\gamma$ values reported in the network literature. Our estimates are most similar to those of Liljeros et al. [23] and Rothenberg and Muth [144], but interpretation of the exact values is challenging. Estimates from the literature mainly fall between 2 and 4, although substantial variation within this range exists, even for networks with apparently similar demographics. For example, Schneeberger et al. [24] estimate higher $\gamma$ values for heterosexual females than heterosexual males, while

Clémençon et al. [25] find the opposite. Some of these differences are likely due to the differing time intervals considered in each study – some investigate the distribution of lifetime partners, while others consider only recent partnerships, such as in the past year. However, the effect of the time period on $\gamma$ is not readily obvious from the available estimates.

| reference | risk group | observation period | estimated $\gamma$ |
|---|---|---|---|
| Colgate et al. [22] | MSM | 1 year | 3 |
| Clémençon et al. [25] | MSM | 20 years | 3.02 |
| Schneeberger et al. [24] | MSM | 1 year | 1.57-1.75 |
| | | six years | 1.85 |
| | | lifetime | 1.64-1.75 |
| Leigh Brown et al. [95] | MSM | 7 years | 2.7 |
| Clémençon et al. [25] | HET, females | 20 years | 2.71 |
| | HET, males | 20 years | 3.36 |
| Schneeberger et al. [24] | HET, females | 1 year | 3.10-3.68 |
| | | lifetime | 2.99-3.09 |
| | HET, males | 1 year | 2.48-2.85 |
| | | lifetime | 2.46-2.47 |
| Liljeros et al. [23] | mixed sexual, females | 1 year | 2.54 |
| | | lifetime | 2.1 |
| | mixed sexual, males | 1 year | 2.31 |
| | | lifetime | 1.6 |
| Rothenberg and Muth [144] | mixed sexual | 1-15 years | 2.12-2.65 |
| Jones and Handcock [180] | mixed sexual, females | not stated | 3.84-17.04 |
| | mixed sexual, males | not stated | 3.03-5.43 |
| Latora et al. [182] | mixed sexual, females | 1 year | 3.9 |
| | mixed sexual, males | 1 year | 2.93 |
| Rothenberg and Muth [144] | IDU | 1-2 years | 2.08-2.87 |
| | mixed sexual and IDU | 3-5 years | 2.20-4.74 |
| Dombrowski et al. [103] | IDU | 2 years | 1.8 |

Table 2.9: Values of power law exponent $\gamma$ previously reported in the literature. In the risk group column, "mixed sexual" indicates participants were not selected on the basis of sexual orientation; these values likely reflect primarily heterosexual dynamics. Leigh Brown et al. [95] studied a phylogenetically estimated transmission network, rather than a contact network. The values presented by Jones and Handcock [180] were in the context of arguing that the power law distribution is not suitable for use with network data.

### HIV prevalence estimates are often discordant with epidemiological information

The true HIV prevalence in a population can be difficult to estimate for several reasons. HIV-infected individuals may be asymptomatic for months or years, possibly delaying their awareness of their status. In many contexts, the risk factors for acquisition of HIV are illegal or stigmatized, which may represent a barrier to testing, treatment, and/or disclosure of status. Several of the studies whose data we analyzed report the number of new infections over the study period. We expect these numbers to be lower, perhaps substantially, than the true prevalence. We also note that our simulation study showed that the

estimates of $I$ obtained with *netabc* were upwardly biased, the severity increasing with the true value of $I$. In addition, our initial exploratory analysis showed that the identifiability of $I$ decreases with the number of sampled tips (figs. B.8 and B.12). In real world studies, the proportion of infected individuals sampled is usually low, which may impede our ability to estimate $I$. In the following, we discuss any information available about the HIV prevalence in the analyzed study populations, and relate this to our estimates. However, due to the aforementioned challenges, this information cannot be taken as absolute "ground truth". Discrepancies between our results and the available data may indicate that the estimates produced by *netabc* are inaccurate, but they may also indicate that the available epidemiological data was not sufficient to estimate the true prevalence. Since the estimates of $I$ were largely robust to the choice of prior on $m$ (figs. 2.16 and B.26), we discuss here only the results obtained with the prior $m \sim \text{DiscreteUniform}(1,5)$.

The prevalence estimates for two IDU epidemics in Eastern Europe, IDU/Romania and IDU/Estonia, were 747 and 373. The authors of the IDU/Romania study [158] reported that 494 new HIV infections were diagnosed among the studied population (IDUs in Bucharest) during the study period, and this value was contained in our 95% HPD interval (136 - 2379). On the other hand, the IDU/Estonia data were collected from several locations in the country. Using their reported incidence estimates, there were approximately 2000 new infections in the country during the study period, a value outside our 95% HPD interval (171 - 1012). The authors indicated that their sample was "convenience based"; it is possible that other, harder-to-reach individuals had no contacts with those sampled, reducing the apparent network size. In contrast, the IDU/Canada data was collected primarily from IDU in Vancouver's downtown east side, a population that has been intensively targeted by public health interventions. Combined with the routine clinical genotyping performed in British Columbia for all new HIV infections, it is not unreasonable to suppose that the sampled individuals (399) comprise the majority of HIV-positive network members. This value was the lower bound of our estimated 95% HPD interval (384 - 3484), but it seems likely that the posterior mean of 1356 is an overestimate.

The prevalence estimates for two MSM populations in major Chinese cities (MSM/Beijing and MSM/Shanghai) were 676 and 1391. The individuals sampled for the MSM/Beijing study were recruited by following a prospective cohort of size 2000. 179 new infections were identified in the cohort during the study period, which is much lower than the estimated prevalence obtained with *netabc*. However, the authors did not claim to have sampled the entire sexual network. In a nationwide survey, Wu et al. [183] estimated that there were 24,198 MSM living in Beijing, of whom 5.7%, or 1379, were HIV-infected, which is close to the upper limit of the 95% confidence interval of our estimate of $I$ (175 - 1401). The same survey estimated the size of the MSM population in Shanghai 14,511, with an HIV prevalence of 6.8% or approximately 987 individuals. This was within the 95% confidence interval of the estimate obtained with *netabc* for the MSM/Shanghai data (311 - 2822). It is worth noting that the authors of the IDU/Shanghai study [164] claimed that there were 80,000 MSM living in Shanghai, which is a significantly different estimate than that obtained by Wu et al. [183]. The authors of the MSM/Taiwan study [163] stated that there were 18,378 HIV-infected individuals in Taiwan, of which approximately 44% (8086 individuals) were MSM. This value is well outside of our 95% HPD (268 - 607), suggesting

that the entire MSM population does not comprise a single network. In fact, our estimate of $I$ for these data (410), was closer to the number of MSM enrolled in the study (301) than to the epidemiologically estimated prevalence. The study focused exclusively on newly diagnosed individuals. On the other hand, our estimated prevalence for the MSM/USA data (482) was clearly an underestimate. The study enrolled almost 648 HIV-infected individuals, which should have been a lower bound on $I$, although the value was contained in the 95% HPD interval (180 - 1112).

The HET/Uganda, HET/Malawi, and HET/Botswana data all originated from populations in sub-Saharan Africa, where the dominant mode of transmission is believed to be heterosexual. The HET/Uganda study population included 1434 seropositive individuals. Since the majority of inhabitants of the study district were enrolled in the study, it is likely that this number represents the majority of HIV-infected persons. Our estimate was slightly lower (939), but the value 1434 was contained within the 95% HPD interval (226 - 2661). The HET/Malawi data represents possibly the oldest population-level HIV sequencing effort, carried out on samples collected for other purposes between 1981 and 1989. Every individual in the study district was sampled; among these, a total of 200 were HIV-positive. Although the HET/Malawi data had the lowest estimated $I$ value among the heterosexual datasets, the posterior mean (724), was much higher than this value. Due to the thorough population-level screening, it is most likely that the posterior mean represents an overestimate, especially considering the upward bias of ABC estimates of $I$ observed on simulated data (fig. B.23). It is also possible that the true sexual network extended outside the study's geographic area to include a larger number of infected individuals. The authors of the HET/Botswana data [159, 160] estimated that there were 1731 HIV-positive individuals in the study area. HIV sequences were obtained from approximately 70% of these individuals. The estimated prevalence we obtained was much higher (5432), with a 95% HPD interval spanning nearly the entire prior region. The authors noted that the town where the study took place was located proximally to the country's capital city, and suggested that frequent travel between the two locations may have facilitated linking of their sexual networks. Thus, the high estimate of $I$ we obtained for these data may include a larger network component based in the capital city.

Cuevas et al. [165] reported that 620 newly infected and 1500 chronically infected patients had been sampled during the study period, for a total of 2120. This is substantially greater than the point estimate of 702 we obtained with ABC. This error may in part be due to model misspecification – it is highly unlikely that a mixed risk group population would form a single, well-connected contact network of the type generated by the BA model.

### Modelling assumptions

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 ABC to fit a model which relaxes some or all of these assumptions, which may be a fruitful avenue for future investigation. Despite the possible misspecification, our estimates of the power law exponent $\gamma$ were within the range of values reported in the literature (table 2.8), and our estimates of $\alpha$ were in rough agreement with a previous study using survey-based epidemiology [105].

# Chapter 3

# Conclusion

Due to the rapid advancement of nucleotide sequencing technology, viral sequence data have become increasingly feasible to collect on a population level. Through phylodynamic methods, these data offer a window into epidemiological processes that would otherwise be virtually impossible to study on a realistic scale.

This thesis developed *netabc*, a computer program implementing a statistical inference method for contact network parameters from viral phylogenetic data. *Netabc* brings together the areas of viral phylodynamics and network epidemiology, which have only intersected in a very limited fashion thus far [39]. The use of kernel-ABC, a likelihood-free method, makes it possible to fit network models to phylogenies without calculating intractable likelihoods.

Although phylodynamic methods have been developed to fit a wide variety of epidemiological models to phylogenetic data assuming homogeneous mixing [47, 184], our method is able to fit models not requiring this assumption. We believe this capability will be of broad interest to the molecular evolution and epidemiology community, as it widens the field of epidemiological parameters that may be investigated through viral sequence data. In addition, the characterization of local contact networks could be valuable from a public health perspective, such as for investigating optimal vaccination strategies [8–10, 89]. This information could assist in curtailing current epidemics, as well as preventing future epidemics of different diseases over the same contact network.

The particular model we have investigated uses a preferential attachment mechanism to generate scale-free networks resembling real-world social and sexual networks [22–24]. Of the four parameters we considered, the preferential attachment power $\alpha$ was the most readily estimable. Estimating $\alpha$ with traditional epidemiological methods is challenging due to the requirement of sampling the high-degree nodes making up the tail of the power law distribution, although approaches such as respondent-driven sampling [90] or collection of longitudinal partner count data [105] may be effective.

In closing, *netabc* combines phylodynamics, contact network epidemiology, approximate Bayesian computation, and sequential Monte Carlo to provide a source of insight into network structures complementary to traditional epidemiology.

# 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] JAP Heesterbeek. *Mathematical Epidemiology of Infectious Diseases: Model Building, Analysis and Interpretation*. Vol. 5. John Wiley & Sons, 2000.

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

[6] 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.

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

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

[9] Xiao-Long Peng et al. "Vaccination intervention on epidemic dynamics in networks". In: *Physical Review E* 87.2 (2013), p. 022813.

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

[11] 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.

[12] Erik M Volz, Katia Koelle, and Trevor Bedford. "Viral phylodynamics". In: *PLoS Computational Biology* 9.3 (2013), e1002947.

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

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

[15] Yun-Xin Fu and Wen-Hsiung Li. “Estimating the age of the common ancestor of a sample of DNA sequences”. In: *Molecular Biology and Evolution* 14.2 (1997), pp. 195–199.

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

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

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

[19] Trevelyan McKinley, Alex R Cook, and Robert Deardon. “Inference in epidemic models without likelihoods”. In: *The International Journal of Biostatistics* 5.1 (2009), pp. 1–40.

[20] 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.

[21] 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.

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

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

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

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

[26] 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.

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

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

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

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

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

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

[33] 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.

[34] 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.

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

[36] 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.

[37] 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).

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

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

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

[41] 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.

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

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

[44] David G Kendall et al. "On the generalized "birth-and-death" process". In: *The Annals of Mathematical Statistics* 19.1 (1948), pp. 1–15.

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

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

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

[48] Masatoshi Nei and Sudhir Kumar. *Molecular Evolution and Phylogenetics*. Oxford University Press, 2000.

[49] Thomas Leitner. *The Molecular Epidemiology of Human Viruses*. Springer Science & Business Media, 2002.

[50] Eben Kenah et al. “Molecular infectious disease epidemiology: survival analysis and algorithms linking phylogenies to transmission trees”. In: *PLOS Computational Biology* 12.4 (2016), e1004869.

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

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

[53] Matthew Hall, Mark Woolhouse, and Andrew Rambaut. “Epidemic reconstruction in a phylogenetics framework: transmission trees as partitions of the node set”. In: *PLoS Computational Biology* 11.12 (2015), e1004613.

[54] 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.

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

[56] 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), p. 10489.

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

[58] 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.

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

[60] 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.

[61] Ethan Romero-Severson et al. “Timing and order of transmission events is not directly reflected in a pathogen phylogeny”. In: *Molecular biology and evolution* (2014), msu179.

[62] 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.

[63] Dimitrios Paraskevis et al. “Phylogenetic reconstruction of a known HIV-1 CRF04_cpx transmission network using maximum likelihood and Bayesian methods”. In: *Journal of molecular evolution* 59.5 (2004), pp. 709–717.

[64] 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.

[65] 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.

[66] Marco J Morelli et al. “A Bayesian inference framework to reconstruct transmission trees using epidemiological and genetic data”. In: *PLoS Computational Biology* 8.11 (2012), e1002768.

[67] 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.

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

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

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

[71] 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.

[72] 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.

[73] Christopher JC Burges. “A tutorial on support vector machines for pattern recognition”. In: *Data Mining and Knowledge Discovery* 2.2 (1998), pp. 121–167.

[74] 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 of the Association for Computational Linguistics*. Association for Computational Linguistics. 2002, pp. 263–270.

[75] Alessandro Moschitti. “Making tree kernels practical for natural language learning”. In: *European Chapter of the Association for Computational Linguistics*. Vol. 113. 120. 2006, p. 24.

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

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

[78] Jacob L Moreno. *Who Shall Survive?* Beacon House Inc., 1934.

[79] JA Barnes. “Class and Committees in a Norwegian Island Parish”. In: *Human Relations* 7.1 (1954), pp. 39–58.

[80] Stanley Wasserman and Katherine Faust. *Social Network Analysis: Methods and Applications*. Vol. 8. Cambridge University Press, 1994.

[81] Rebecca F Baggaley, Richard G White, and Marie-Claude Boily. “HIV transmission risk through anal intercourse: systematic review, meta-analysis and implications for HIV prevention”. In: *International journal of epidemiology* (2010), dyq057.

[82] John A Jacquez et al. “Modeling and analyzing HIV transmission: the effect of contact patterns”. In: *Mathematical Biosciences* 92.2 (1988), pp. 119–199.

[83] 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.

[84] David A Rolls et al. “A simulation study comparing epidemic dynamics on exponential random graph and edge-Triangle configuration type contact network models”. In: *PloS ONE* 10.11 (2015), e0142181.

[85] Tanja Stadler et al. “Insights into the early epidemic spread of Ebola in Sierra Leone provided by viral sequence data”. In: *PLoS Currents Outbreaks* 10 (2014).

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

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

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

[89] Julie Rushmore et al. “Network-based vaccination improves prospects for disease control in wild chimpanzees”. In: *Journal of The Royal Society Interface* 11.97 (2014), p. 20140349.

[90] Douglas D Heckathorn. “Respondent-driven sampling: a new approach to the study of hidden populations”. In: *Social problems* (1997), pp. 174–199.

[91] 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.

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

[93] 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.

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

[95] 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.

[96] 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.

[97] 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.

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

[99] Albert-László Barabási et al. “Evolution of the social network of scientific collaborations”. In: *Physica A: Statistical mechanics and its applications* 311.3 (2002), pp. 590–614.

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

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

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

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

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

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

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

[107] 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.

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

[109] 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.

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

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

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

[113] 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.

[114] Christian Robert and George Casella. *Monte Carlo Statistical Methods*. Springer Science & Business Media, 2004.

[115] Arnaud Doucet, Simon Godsill, and Christophe Andrieu. "On sequential Monte Carlo sampling methods for Bayesian filtering". In: *Statistics and Computing* 10.3 (2000), pp. 197–208.

[116] 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.

[117] Jun S Liu. *Monte Carlo Strategies in Scientific Computing*. Springer Science & Business Media, 2008.

[118] 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.

[119] Olav Kallenberg. *Foundations of Modern Probability*. Springer Science & Business Media, 2006.

[120] Neil J Gordon, David J Salmond, and Adrian FM Smith. "Novel approach to nonlinear/non-Gaussian Bayesian state estimation". In: *Radar and Signal Processing, IEE Proceedings F*. Vol. 140. 2. IET. 1993, pp. 107–113.

[121] Adrian Smith et al. *Sequential Monte Carlo Methods in Practice*. Springer Science & Business Media, 2013.

[122] 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.

[123] Jun S Liu and Rong Chen. “Blind deconvolution via sequential imputations”. In: *Journal of the American Statistical Association* 90.430 (1995), pp. 567–576.

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

[125] Mark A Beaumont. “Approximate Bayesian computation in evolution and ecology”. In: *Annual Review of Ecology, Evolution, and Systematics* 41 (2010), pp. 379–406.

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

[127] Simon Aeschbacher, Mark A Beaumont, and Andreas Futschik. “A novel approach for choosing summary statistics in approximate Bayesian computation”. In: *Genetics* 192.3 (2012), pp. 1027–1047.

[128] Michael GB Blum et al. “A comparative review of dimension reduction methods in approximate Bayesian computation”. In: *Statistical Science* 28.2 (2013), pp. 189–208.

[129] Mark M Tanaka et al. “Using approximate Bayesian computation to estimate tuberculosis transmission parameters from genotype data”. In: *Genetics* 173.3 (2006), pp. 1511–1520.

[130] Paul Marjoram and Simon Tavaré. “Modern computational approaches for analysing molecular genetic variation data”. In: *Nature Reviews Genetics* 7.10 (2006), pp. 759–770.

[131] 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.

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

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

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

[135] 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.

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

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

[138] Brian Gough. *GNU Scientific Library Reference Manual*. Third edition. Network Theory Ltd., 2009.

[139] Blaise Barney. *POSIX Threads Programming*. 2009. URL: https://computing.llnl.gov/tutorials/pthreads/ (visited on 09/05/2016).

[140] Tal Pupko et al. “A fast algorithm for joint reconstruction of ancestral amino acid sequences”. In: *Molecular Biology and Evolution* 17.6 (2000), pp. 890–896.

[141] Frank Harary and Edgar M Palmer. *Graphical Enumeration*. Elsevier, 2014.

[142] Iain Murray, Zoubin Ghahramani, and David MacKay. “MCMC for doubly-intractable distributions”. In: *arXiv preprint arXiv:1206.6848* (2012).

[143] Faming Liang. “A double Metropolis–Hastings sampler for spatial models with intractable normalizing constants”. In: *Journal of Statistical Computation and Simulation* 80.9 (2010), pp. 1007–1022.

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

[145] Roy M Anderson and Robert M May. “Directly transmitted infections diseases: control by vaccination”. In: *Science* 215.4536 (1982), pp. 1053–1060.

[146] Ravi Kumar et al. “Stochastic models for the web graph”. In: *Foundations of Computer Science, 2000. Proceedings. 41st Annual Symposium on*. IEEE. 2000, pp. 57–65.

[147] 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.

[148] Simon DW Frost and Erik M Volz. “Modelling tree shape and structure in viral phylodynamics”. In: *Philosophical Transactions of the Royal Society of London B: Biological Sciences* 368.1614 (2013).

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

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

[151] 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.

[152] Greg Snow. *TeachingDemos: Demonstrations for Teaching and Learning*. R package version 2.10. 2016.

[153] Frank E Harrell Jr, with contributions from Charles Dupont, and many others. *Hmisc: Harrell Miscellaneous*. R package version 3.17-4. 2016. URL: https://CRAN.R-project.org/package=Hmisc.

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

[155] Michael D. Karcher et al. “Quantifying and mitigating the effect of preferential sampling on phylodynamic inference”. In: *PLoS Computational Biology* 12.3 (Mar. 2016), pp. 1–19.

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

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

[158] 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.

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

[160] 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.

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

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

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

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

[165] 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.

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

[167] 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.

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

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

[170] Gili Greenbaum, Alan R. Templeton, and Shirli Bar-David. “Inference and analysis of population structure using genetic data and network theory”. In: *Genetics* 202.4 (2016), pp. 1299–312.

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

[172] Joseph Oscar Irwin. “The place of mathematics in medical and biological statistics”. In: *Journal of the Royal Statistical Society* 126.Pt. 1 (1963), pp. 1–41.

[173] Garry Robins et al. “An introduction to exponential random graph (p*) models for social networks”. In: *Social networks* 29.2 (2007), pp. 173–191.

[174] Ken TD Eames and Matt J Keeling. “Modeling dynamic and network heterogeneities in the spread of sexually transmitted diseases”. In: *Proceedings of the National Academy of Sciences* 99.20 (2002), pp. 13330–13335.

[175] 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.

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

[177] Nicolas Bortolussi et al. “apTreeshape: statistical analysis of phylogenetic tree shape”. In: *Bioinformatics* 22.3 (2006), pp. 363–364.

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

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

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

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

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

[183] Z Wu et al. “HIV and syphilis prevalence among men who have sex with men: a cross-sectional survey of 61 cities in China”. In: *Clinical Infectious Diseases* 57.2 (2013), p. 298.

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

[185] Nicholas Metropolis et al. “Equation of state calculations by fast computing machines”. In: *The journal of chemical physics* 21.6 (1953), pp. 1087–1092.

[186] W Keith Hastings. “Monte Carlo sampling methods using Markov chains and their applications”. In: *Biometrika* 57.1 (1970), pp. 97–109.

# Appendix A

# Mathematical models, likelihood, and Bayesian inference

This appendix reviews some fundamental statistical concepts that are referred to throughout the body of this thesis. A *mathematical model* is a formal description of a hypothesized relationship between some observed data, $y$ and covariates $x$. A *parametric* model defines a family of possible relationships between data and outcomes, parameterized by one or more numeric parameters $\theta$. A *statistical* model describes the relationship between data and outcomes in terms of probabilities. Statistical models define, either explicitly or implicitly, the probability of observing $y$ given $x$ and, if the model is parametric, $\theta$. Note that it is entirely possible to have no data $x$, 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 $\beta$ is the slope of the linear relationship, and $\varepsilon_i$ is the error associated with measurement $i$. We can make this model a statistical one by hypothesizing a distribution for the error terms $\varepsilon_i$; most commonly, it is assumed that they are normally distributed with variance $\sigma$. 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 fof observing outcome $y_i$ given the parameters,

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

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

For a general model, the probability density of $y$ given the parameters $\theta$ is also known as the *likeli-*

*hood*, written $\mathcal{L}$, of $\theta$. 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 $\theta$ such that the probability density of our observed outcomes $y$ is as high as possible. The parameters that 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 $\theta$ are examined, $\mathcal{L}(\theta \mid y)$ is calculated for each, and the parameters that 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 multi-modal.

ML estimation makes use only of the data and outcomes to estimate the model parameters $\theta$. However, it is frequently the case that the investigator has some additional information or belief about what $\theta$ 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 $\theta$,

$$\pi(\theta \mid y) = \frac{f(y \mid \theta)\pi(\theta)}{\int f(y \mid \theta)\pi(\theta)\mathrm{d}\,\theta}. \tag{A.1}$$

In principle, $\pi(y \mid \theta)\pi(\theta)$ 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, $\theta$ 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. (A.1), 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 on the space of possible model parameters. The transition density from parameters $\theta_1$ to $\theta_2$ is taken to be

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

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 $\theta$. Therefore, if a long enough random walk is performed on the chain, the distribution of states visited will be a Monte Carlo approximation of $\pi(\theta \mid y)$, from which we can calculate statistics of interest. Actually performing this

random walk is straightforward and can be accomplished via the Metropolis-Hastings algorithm [185, 186] (algorithm A.1).

**Algorithm A.1** 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 $\min\left(1, \frac{f(y \mid \theta')\pi(\theta')q(\theta', \theta)}{f(y \mid \theta\ )\pi(\theta\ )q(\theta, \theta')}\right)$
**end loop**

# Appendix B

# Additional plots

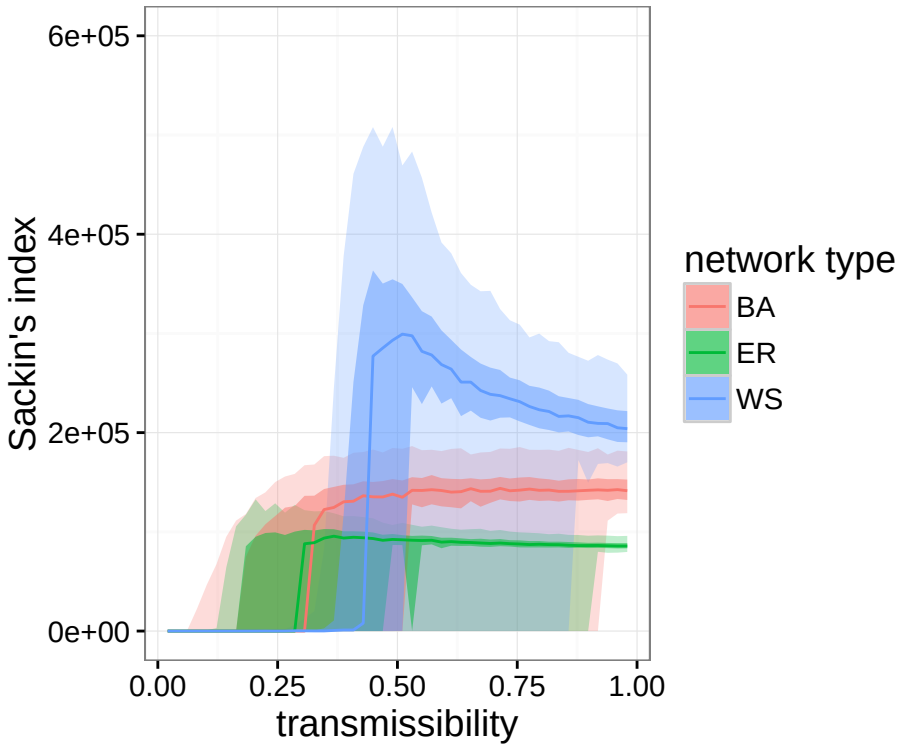


Figure B.1: Reproduction of Figure 1A from Leventhal *et al.* (2012) used to check the accuracy of our implementation of Gillespie simulation. Transmission trees were simulated over three types of network, with pathogen transmissibility varying from 0 to 1. Sackin's index was calculated for each simulated transmission tree. Lines indicate median Sackin's index values, and shaded areas are interquartile ranges.

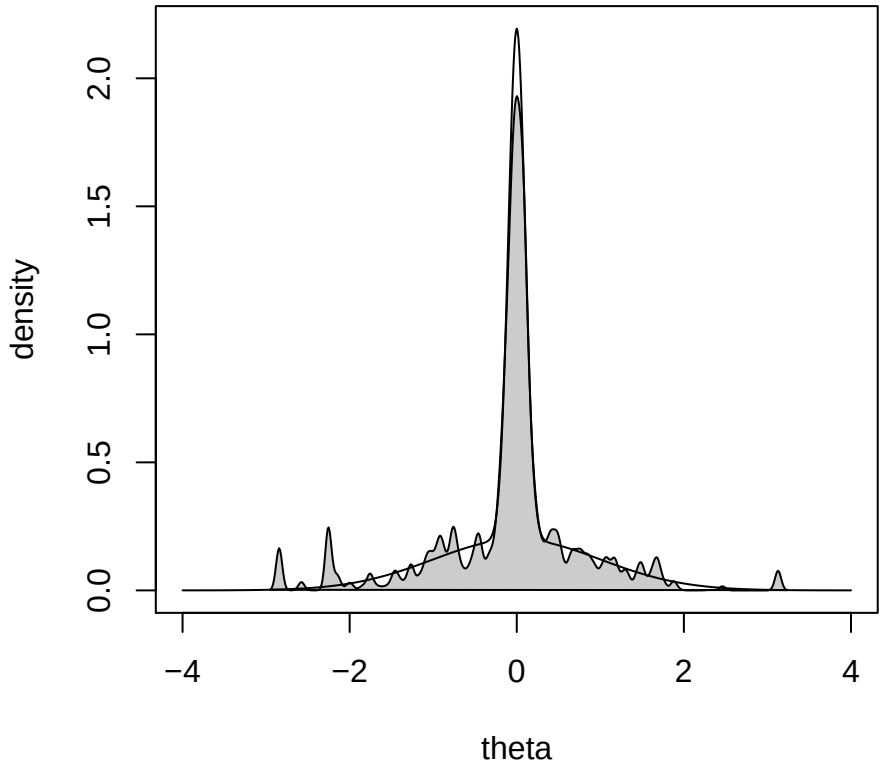


Figure B.2: Approximation of mixture of Gaussians used by Del Moral *et al.* (2012) and Sisson *et al.* (2009) to test SMC. Solid black line indicates true distribution. Grey shaded area shows ABC approximation obtained with our implementation of adaptive ABC-SMC, using 10000 particles with one simulated data point per particle.

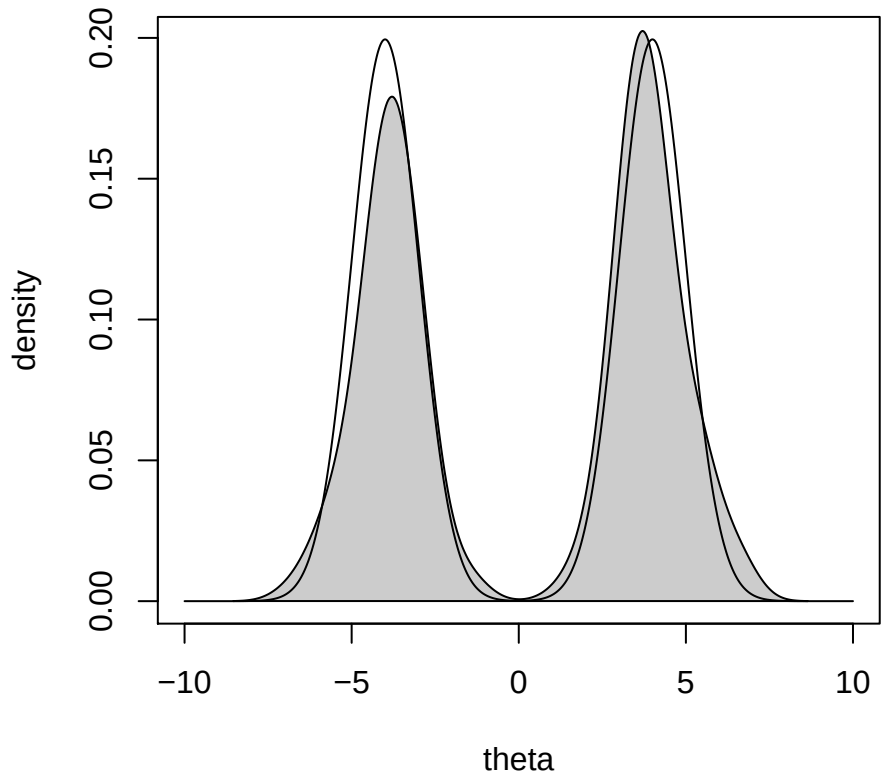


Figure B.3: Approximation of mixture of two Gaussians used to test convergence of SMC algorithm to a bimodal distribution. Solid black line indicates true distribution. Grey shaded area shows ABC-SMC approximation obtained with our implementation, using 10000 particles with one simulated data point per particle.

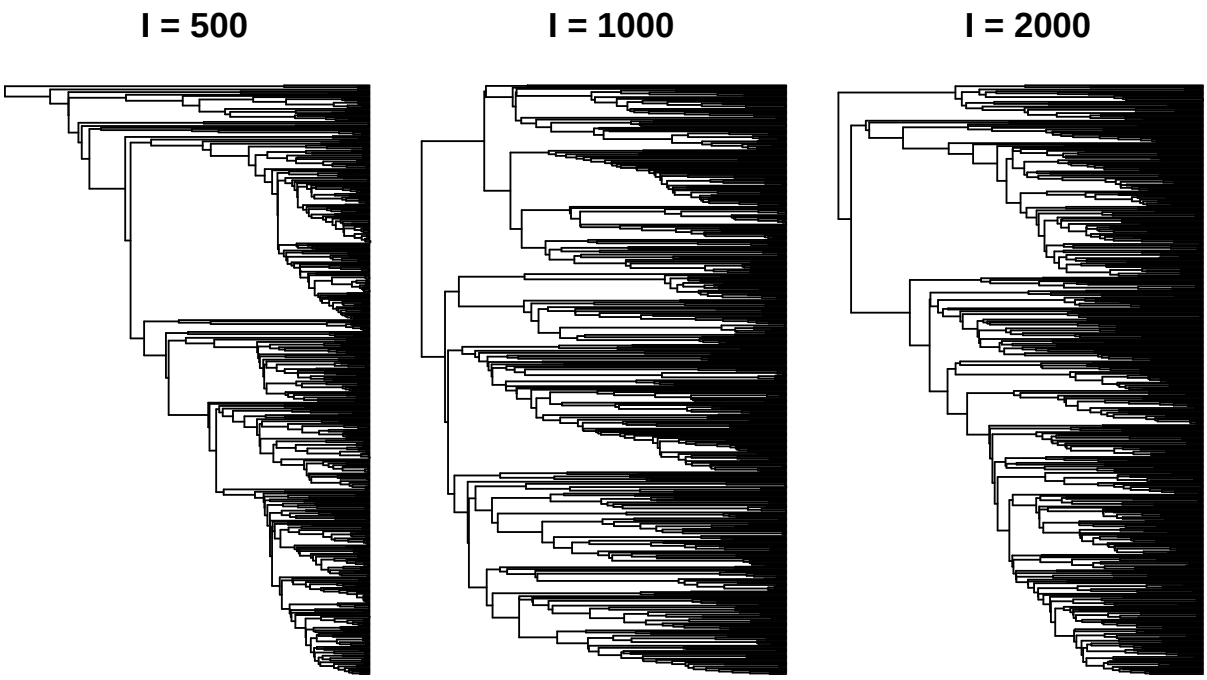


Figure B.4: Simulated transmission trees under three different values of BA parameter $I$. Epidemics were simulated on BA networks with parameters $\alpha = 1.0$, $m = 2$, and $N = 5000$. Epidemics were simulated until $I = 500$, 1000, or 2000 nodes were infected. Transmission trees were created by sampling 500 infected nodes. For higher $I$ values, the network was closer to saturation at the time of sampling, resulting in longer terminal branches as the waiting time until the next transmission increased.

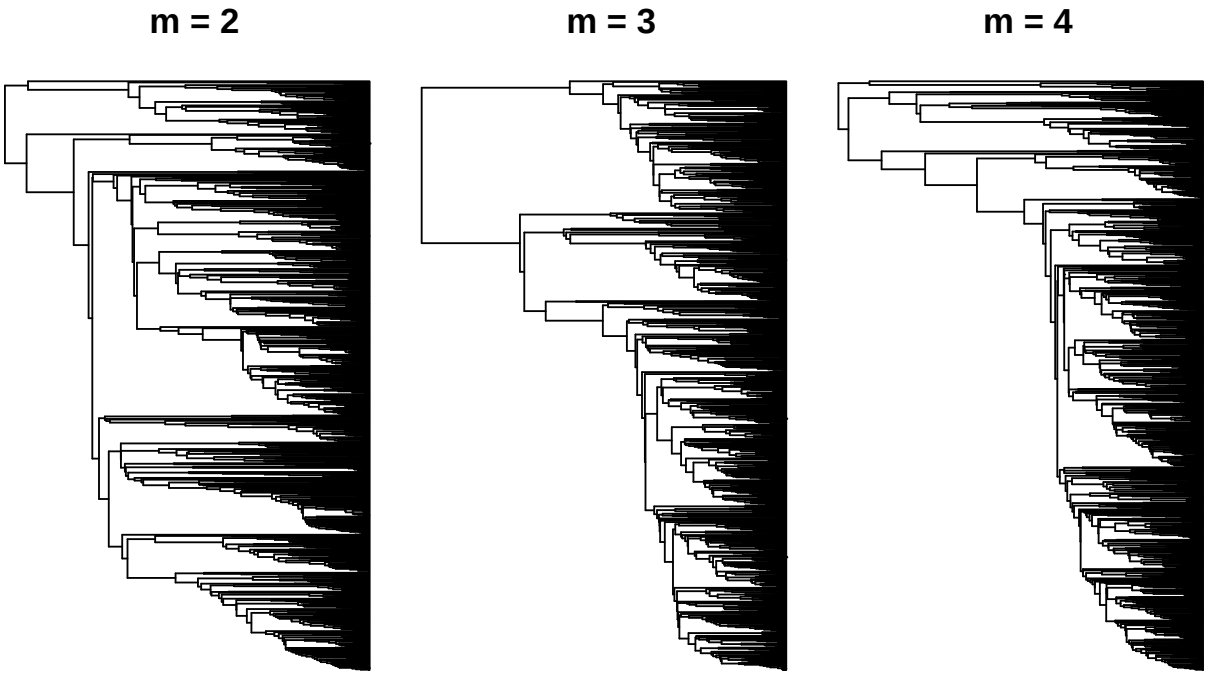


Figure B.5: Simulated transmission trees under three different values of BA parameter $m$. Epidemics were simulated on BA networks with parameters $\alpha = 1.0$, $n = 5000$, and $m = 2$, 3, or 4. Epidemics were simulated until $I = 1000$ nodes were infected. Transmission trees were created by sampling 500 infected nodes.

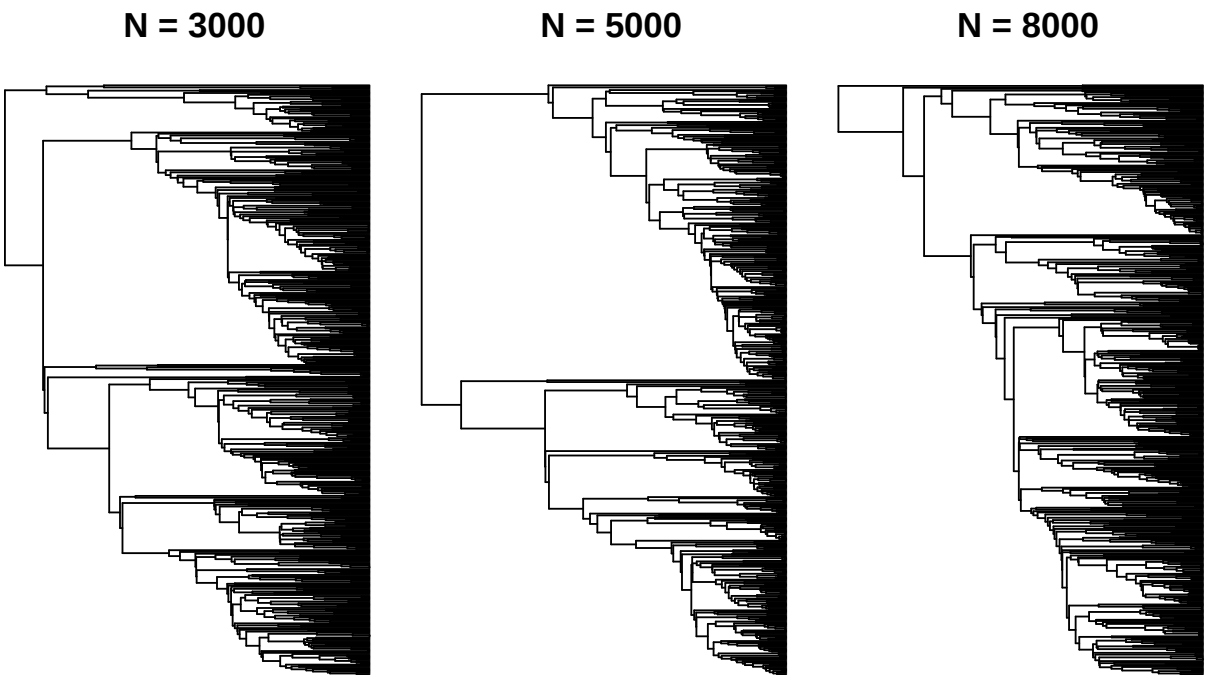


Figure B.6: Simulated transmission trees under three different values of BA parameter $N$. Epidemics were simulated on BA networks with parameters $\alpha = 1.0$, $m = 2$, and $N$ = 3000, 5000, or 8000. Epidemics were simulated until $I = 1000$ nodes were infected. Transmission trees were created by sampling 500 infected nodes. For lower $N$ values, the network was closer to saturation at the time of sampling, resulting in longer waiting times until the next transmission and longer terminal branch lengths.

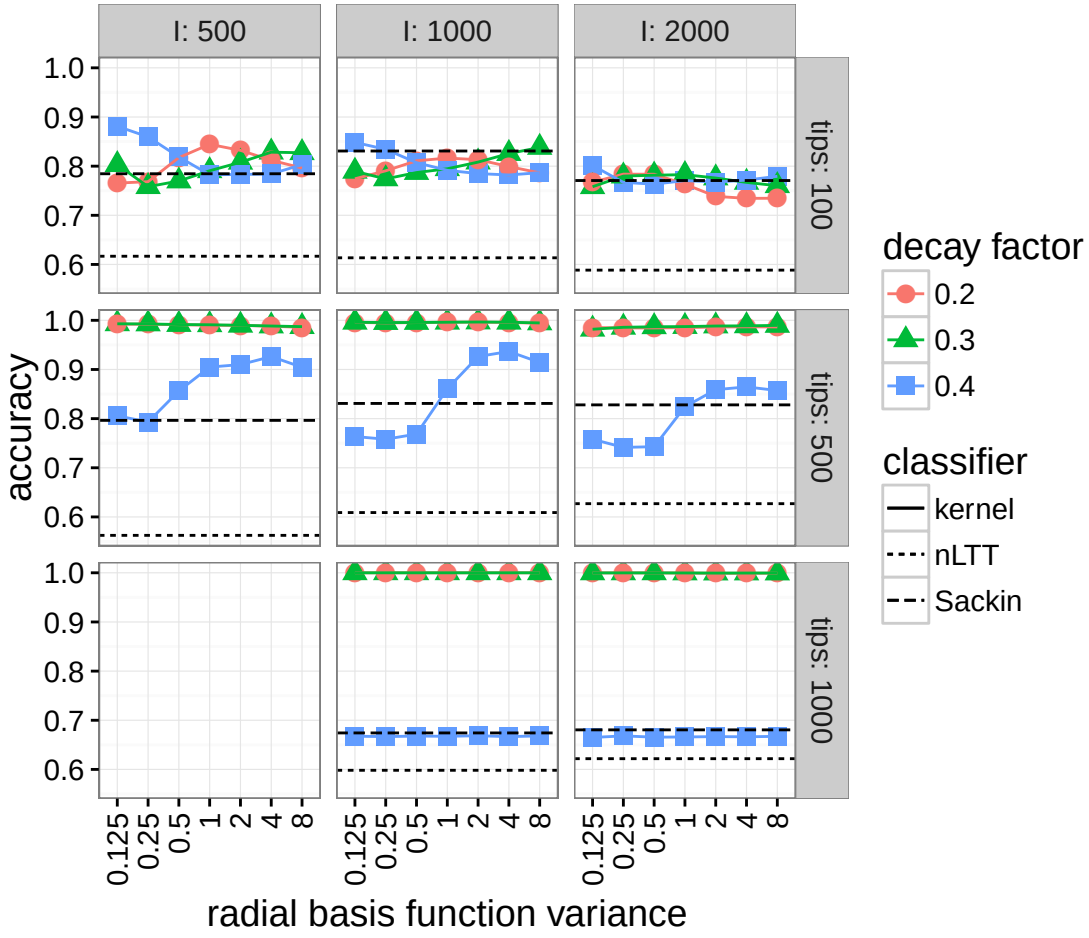


Figure B.7: Cross validation accuracy of classifiers for BA model parameter $\alpha$ for eight epidemic scenarios. Solid lines and points are $R^2$ of tree kernel kSVR under various kernel meta-parameters. Dashed and dotted lines are $R^2$ of linear regression against Sackin's index, and SVR using nLTT.

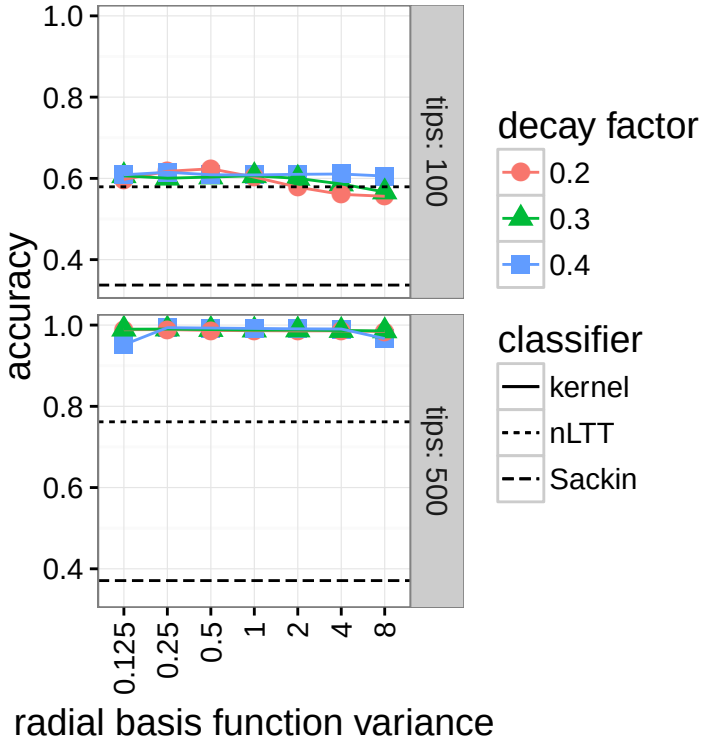


Figure B.8: Cross validation accuracy of classifiers for BA model parameter $I$ for eight epidemic scenarios. Solid lines and points are $R^2$ of tree kernel kSVR under various kernel meta-parameters. Dashed and dotted lines are $R^2$ of linear regression against Sackin's index, and SVR using nLTT.

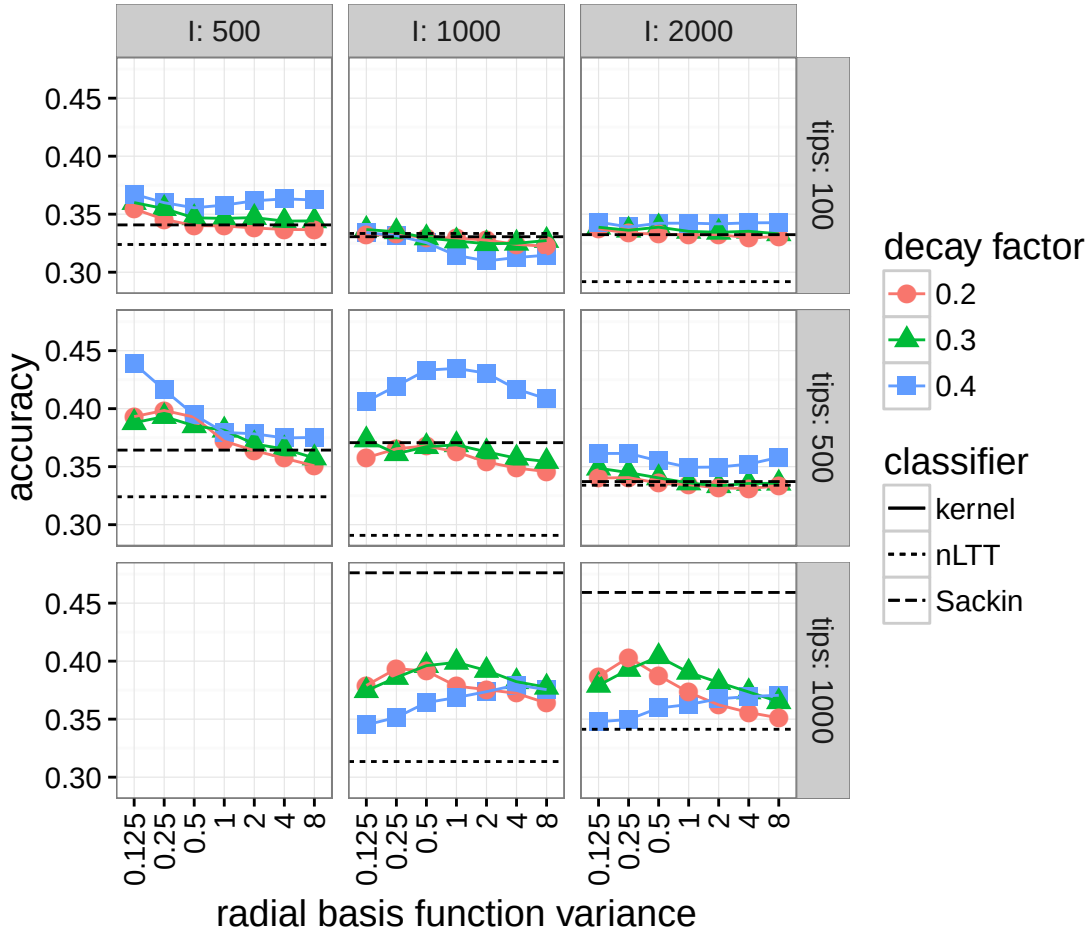


Figure B.9: Cross validation accuracy of classifiers for BA model parameter $m$ for eight epidemic scenarios. Solid lines and points are $R^2$ of tree kernel kSVR under various kernel meta-parameters. Dashed and dotted lines are $R^2$ of linear regression against Sackin's index, and SVR using nLTT.

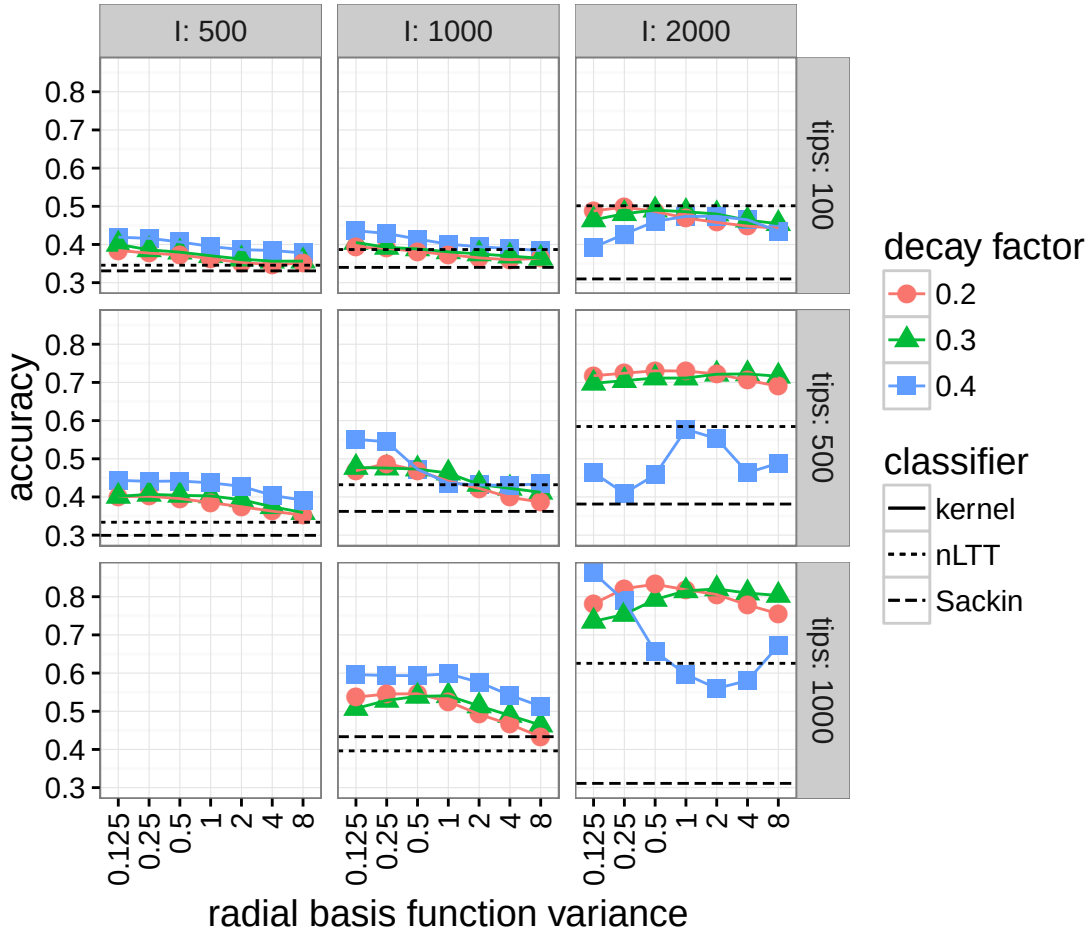


Figure B.10: Cross validation accuracy of classifiers for BA model parameter $N$ for eight epidemic scenarios. Solid lines and points are $R^2$ of tree kernel kSVR under various kernel meta-parameters. Dashed and dotted lines are $R^2$ of linear regression against Sackin's index, and SVR using nLTT.

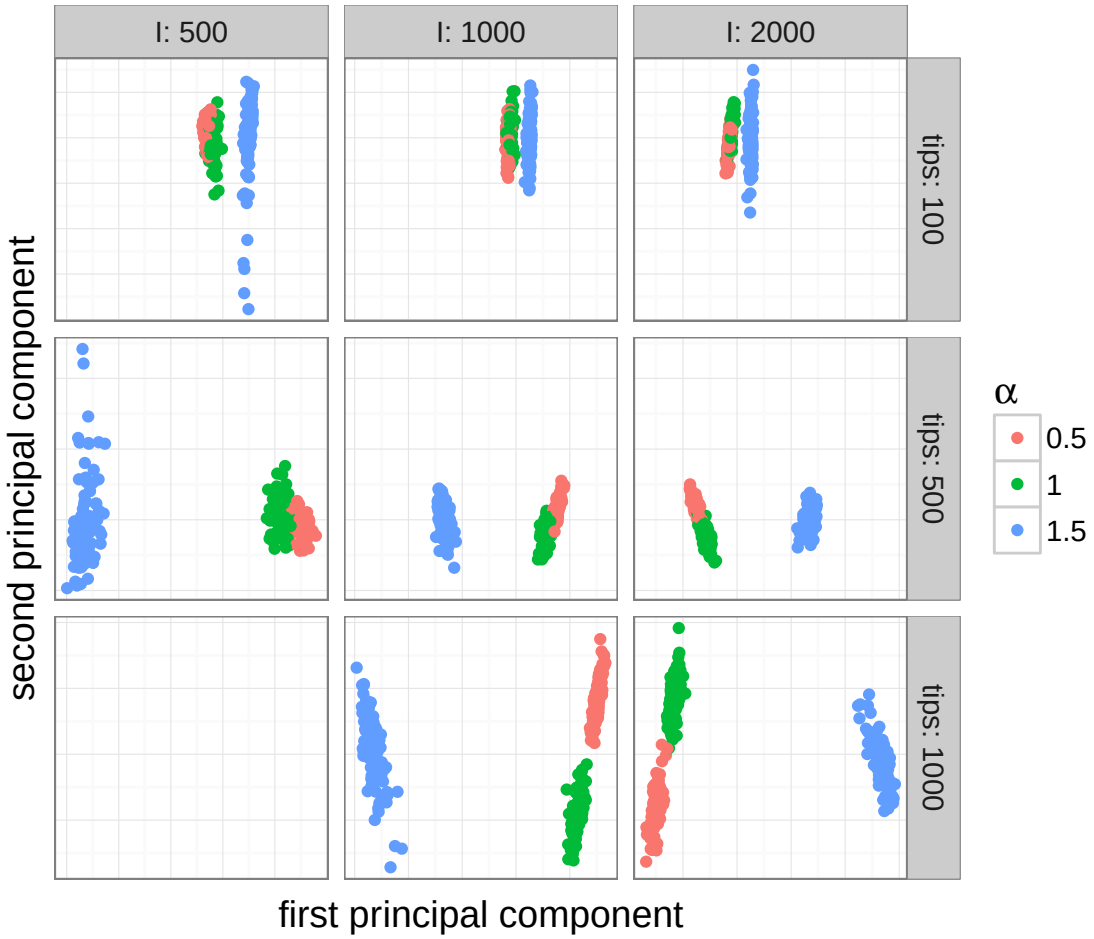


Figure B.11: Kernel principal components projection of trees simulated under three different values of BA parameter $\alpha$, for eight epidemic scenarios.

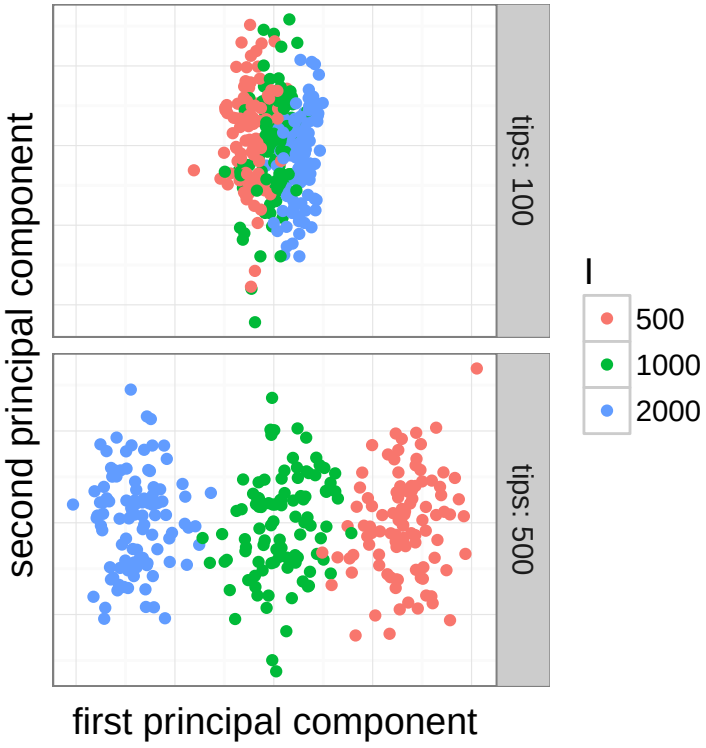


Figure B.12: Kernel principal components projection of trees simulated under three different values of BA parameter $I$, for eight epidemic scenarios.

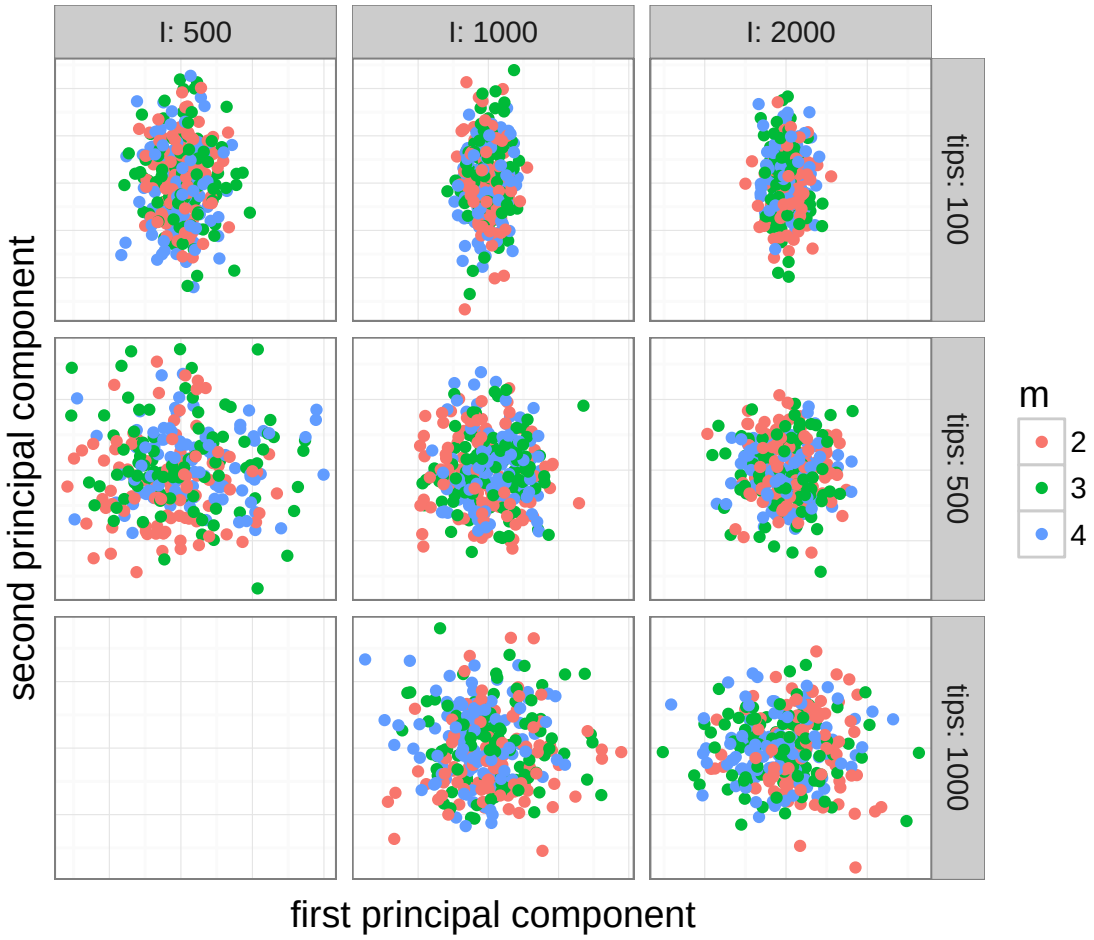


Figure B.13: Kernel principal components projection of trees simulated under three different values of BA parameter $m$, for eight epidemic scenarios.

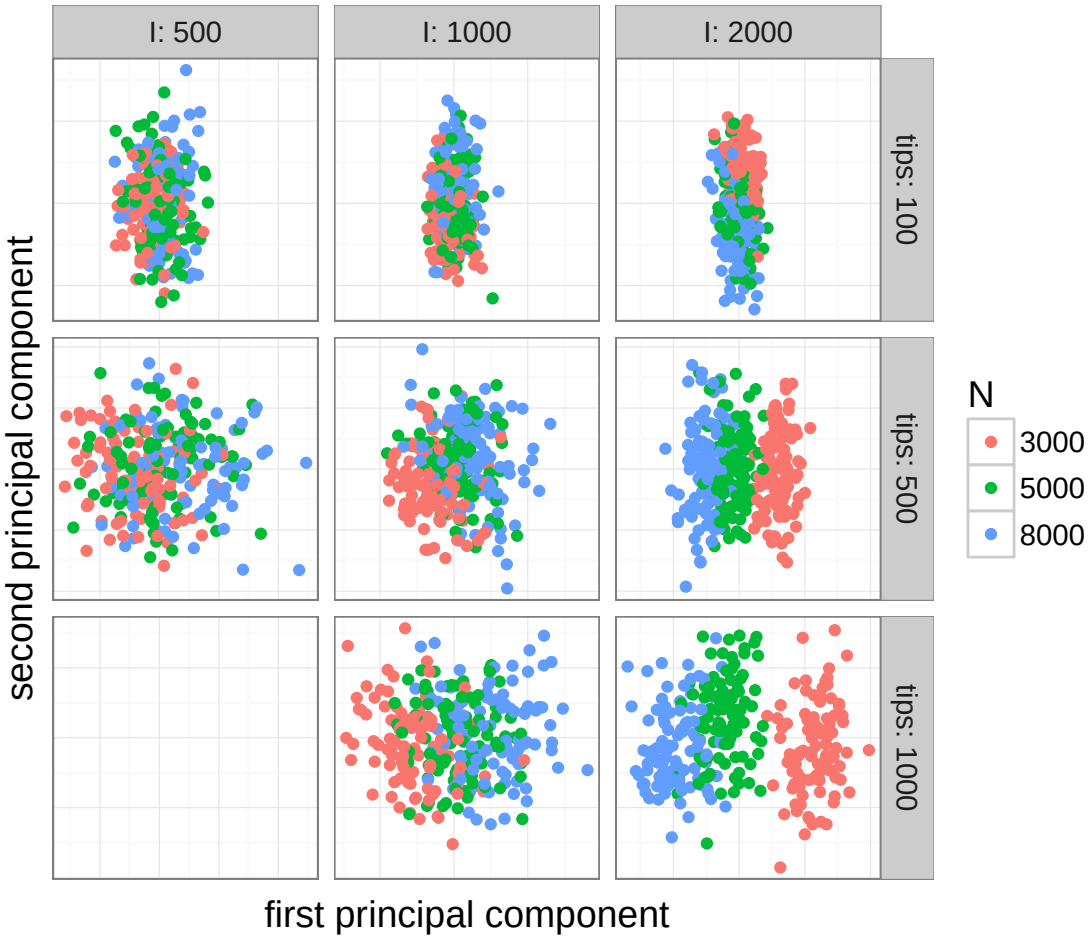


Figure B.14: Kernel principal components projection of trees simulated under three different values of BA parameter $N$, for eight epidemic scenarios.

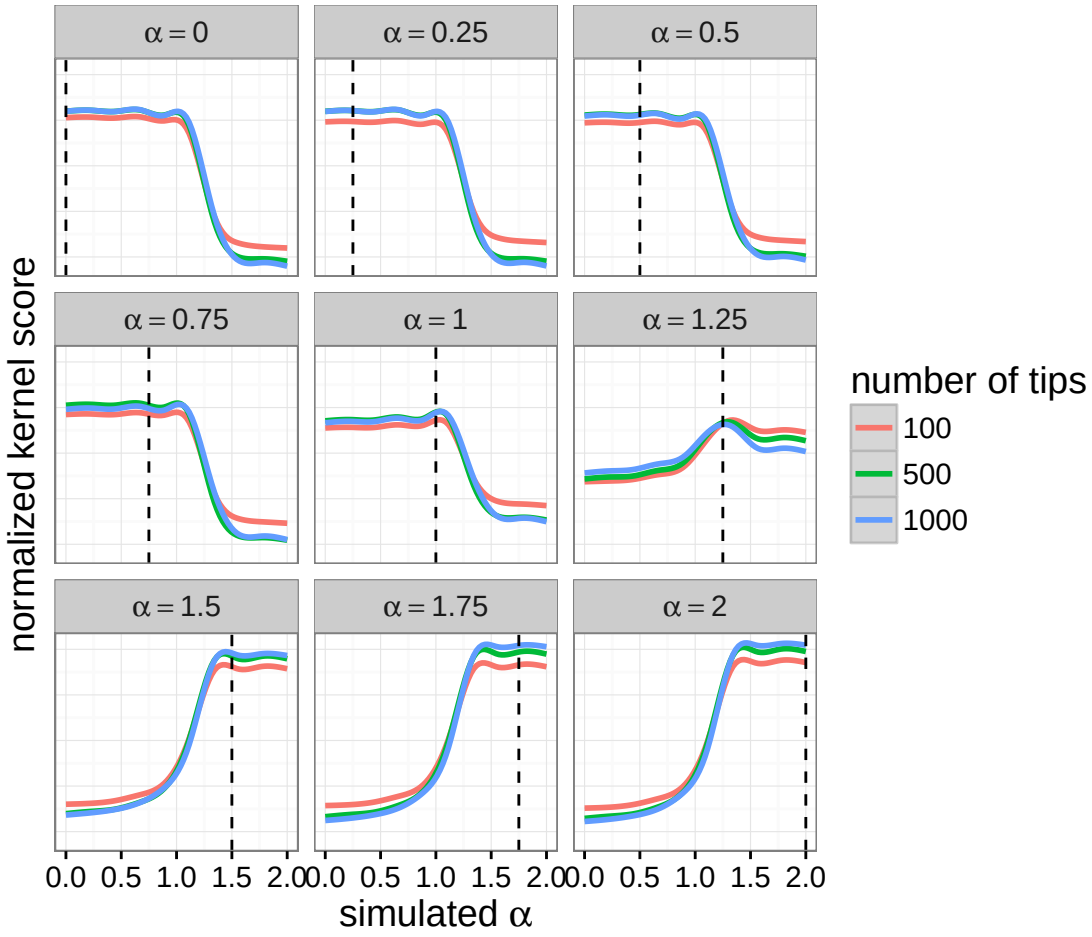


Figure B.15: Grid search kernel scores for testing trees simulated under various $\alpha$ values. The other BA parameters were fixed at $I = 1000$, $N = 5000$, and $m = 2$.

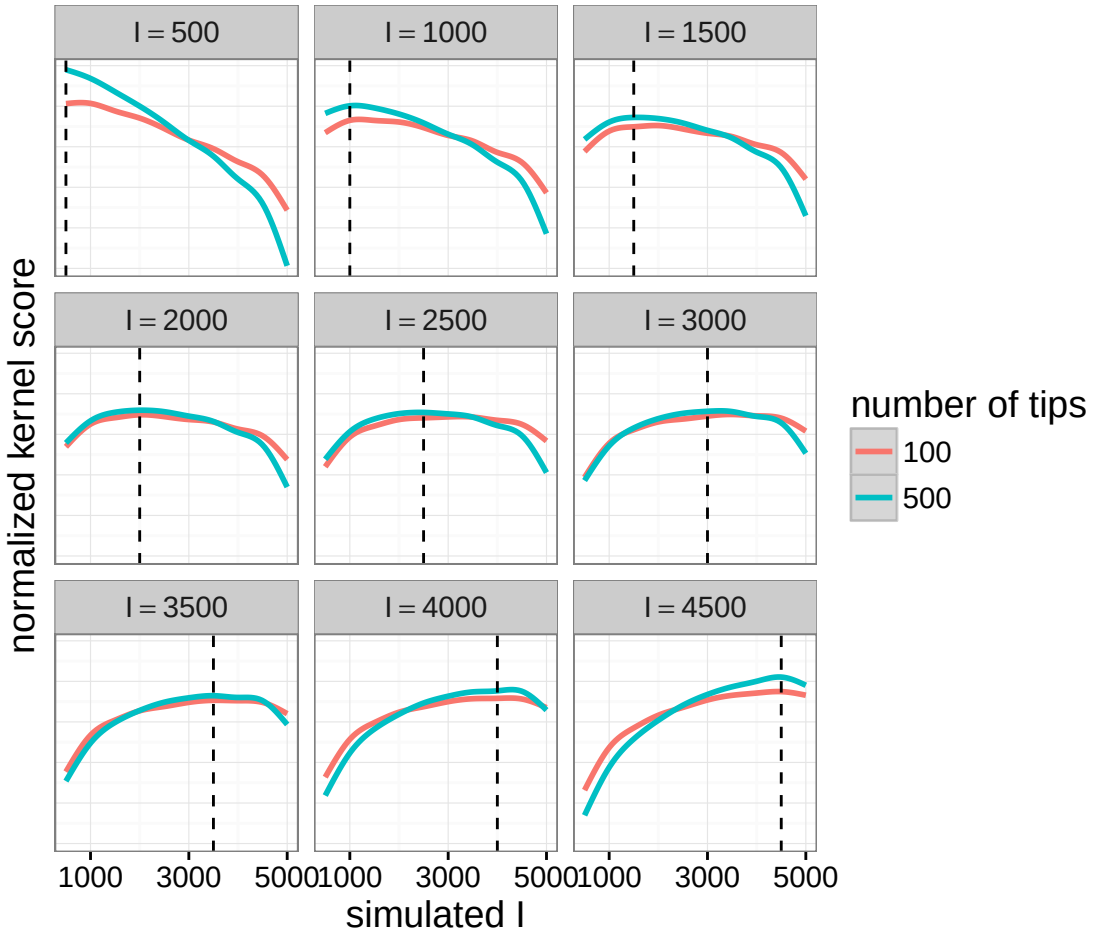


Figure B.16: Grid search kernel scores for testing trees simulated under various $I$ values. The other BA parameters were fixed at $\alpha = 1.0$, $N = 5000$, and $m = 2$.

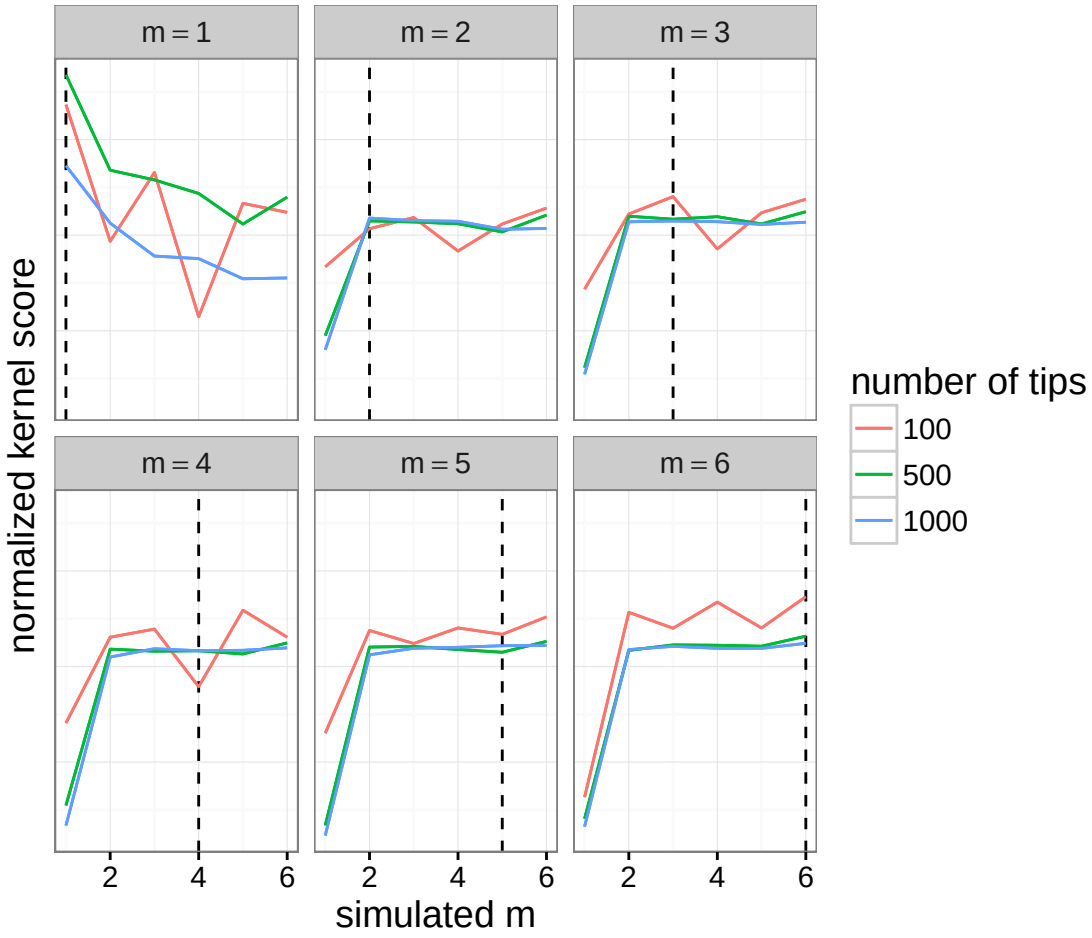


Figure B.17: Grid search kernel scores for testing trees simulated under various $m$ values. The other BA parameters were fixed at $\alpha = 1.0$, $I = 1000$, and $N = 5000$.

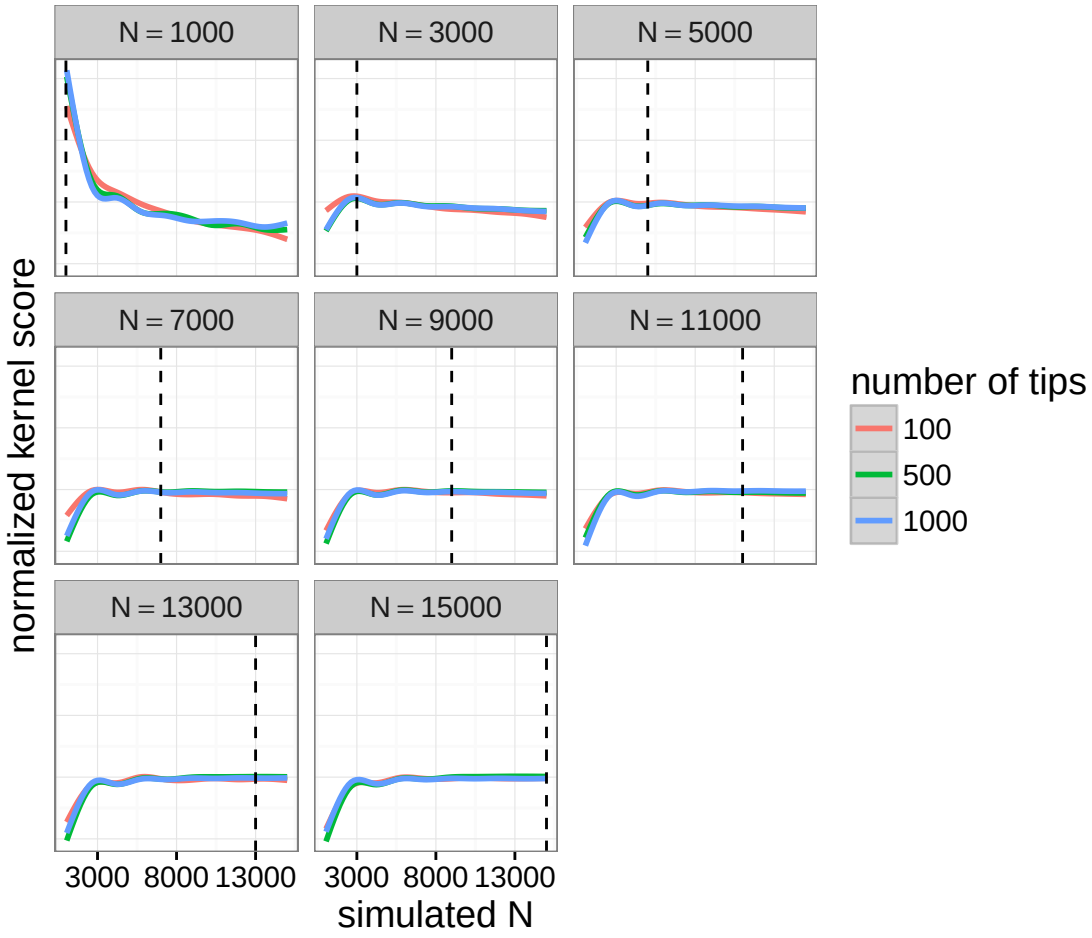


Figure B.18: Grid search kernel scores for testing trees simulated under various $N$ values. The other BA parameters were fixed at $\alpha = 1.0$, $I = 1000$, and $m = 2$.

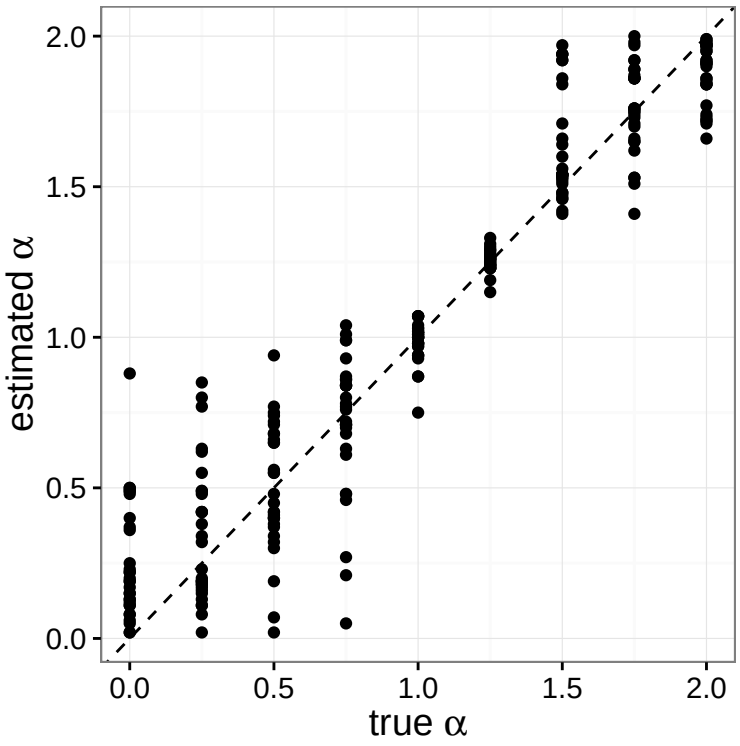


Figure B.19: Point estimates of preferential attachment power $\alpha$ of Barabási-Albert network model, obtained on simulated trees with kernel-score-based grid search. Test trees were simulated according to several values of $\alpha$ ($x$-axis) with other model parameters fixed at $m = 2$, $N = 5000$, and $I = 1000$. The test trees were compared to trees simulated along a narrowly spaced grid of $\alpha$ values using the tree kernel, with the same values of the other parameters. The grid value with the highest median kernel score was taken as a point estimate for $\alpha$ ($y$-axis).

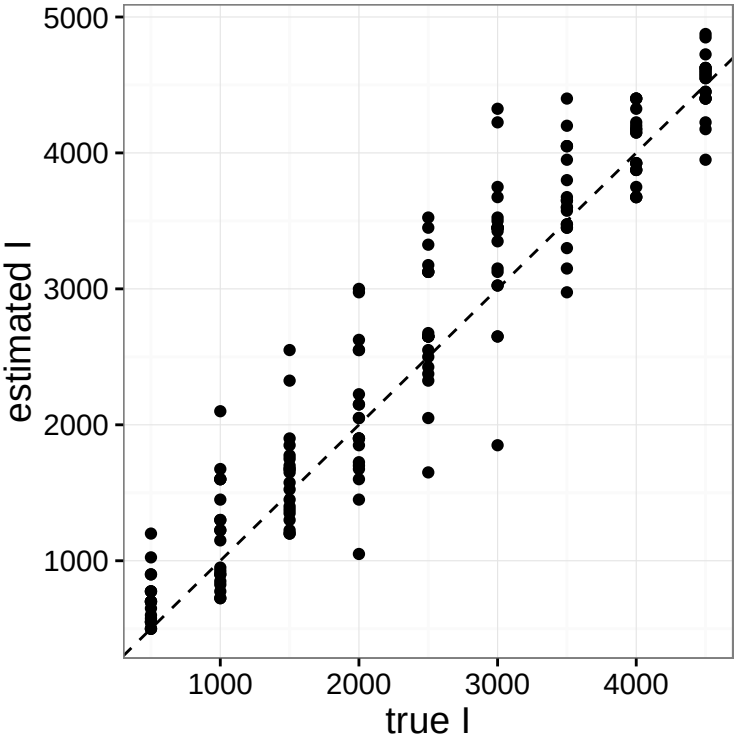


Figure B.20: Point estimates of prevalence at time of sampling $I$ of BA network model, obtained on simulated trees with kernel-score-based grid search. Test trees were simulated according to several values of $I$ ($x$-axis) with other model parameters fixed at $\alpha = 1$, $m = 2$, and $N = 5000$. The test trees were compared to trees simulated along a narrowly spaced grid of $I$ values using the tree kernel, with the same values of the other parameters. The grid value with the highest median kernel score was taken as a point estimate for $I$ ($y$-axis).

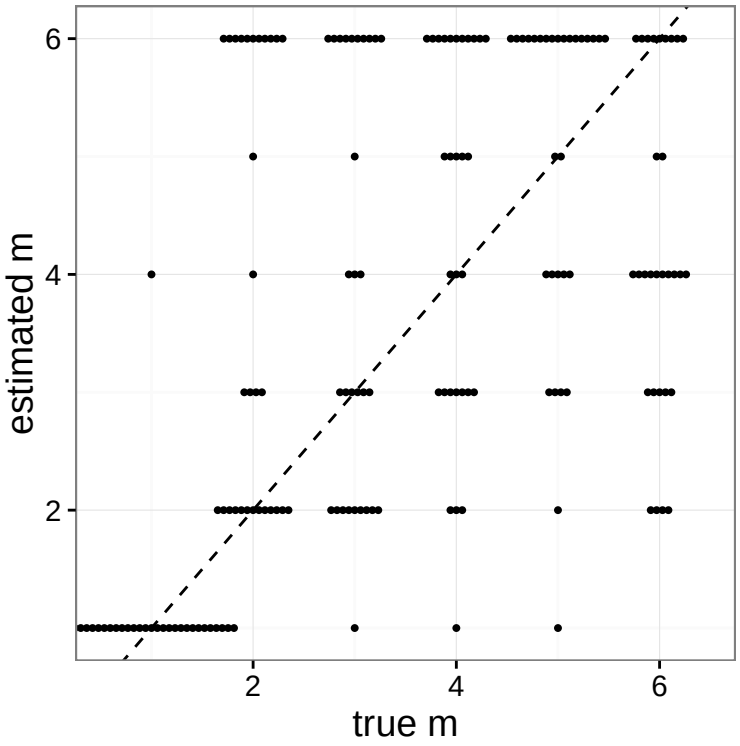


Figure B.21: Point estimates of number of edges per vertex $m$ of BA network model, obtained on simulated trees with kernel-score-based grid search. Test trees were simulated according to several values of $m$ ($x$-axis) with other model parameters fixed at $\alpha = 1$, $I = 1000$, and $N = 5000$. The test trees were compared to trees simulated along a narrowly spaced grid of $m$ values using the tree kernel, with the same values of the other parameters. The grid value with the highest median kernel score was taken as a point estimate for $m$ ($y$-axis).

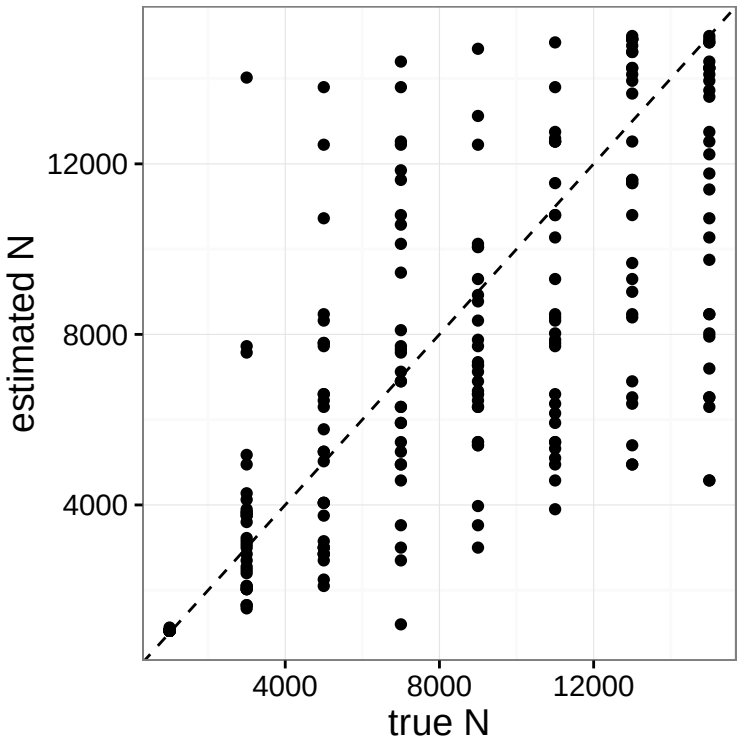


Figure B.22: Point estimates of number of edges per vertex $N$ of BA network model, obtained on simulated trees with kernel-score-based grid search. Test trees were simulated according to several values of $N$ ($x$-axis) with other model parameters fixed at $\alpha = 1$, $m = 2$, and $I = 1000$. The test trees were compared to trees simulated along a narrowly spaced grid of $N$ values using the tree kernel, with the same values of the other parameters. The grid value with the highest median kernel score was taken as a point estimate for $m$ ($y$-axis).

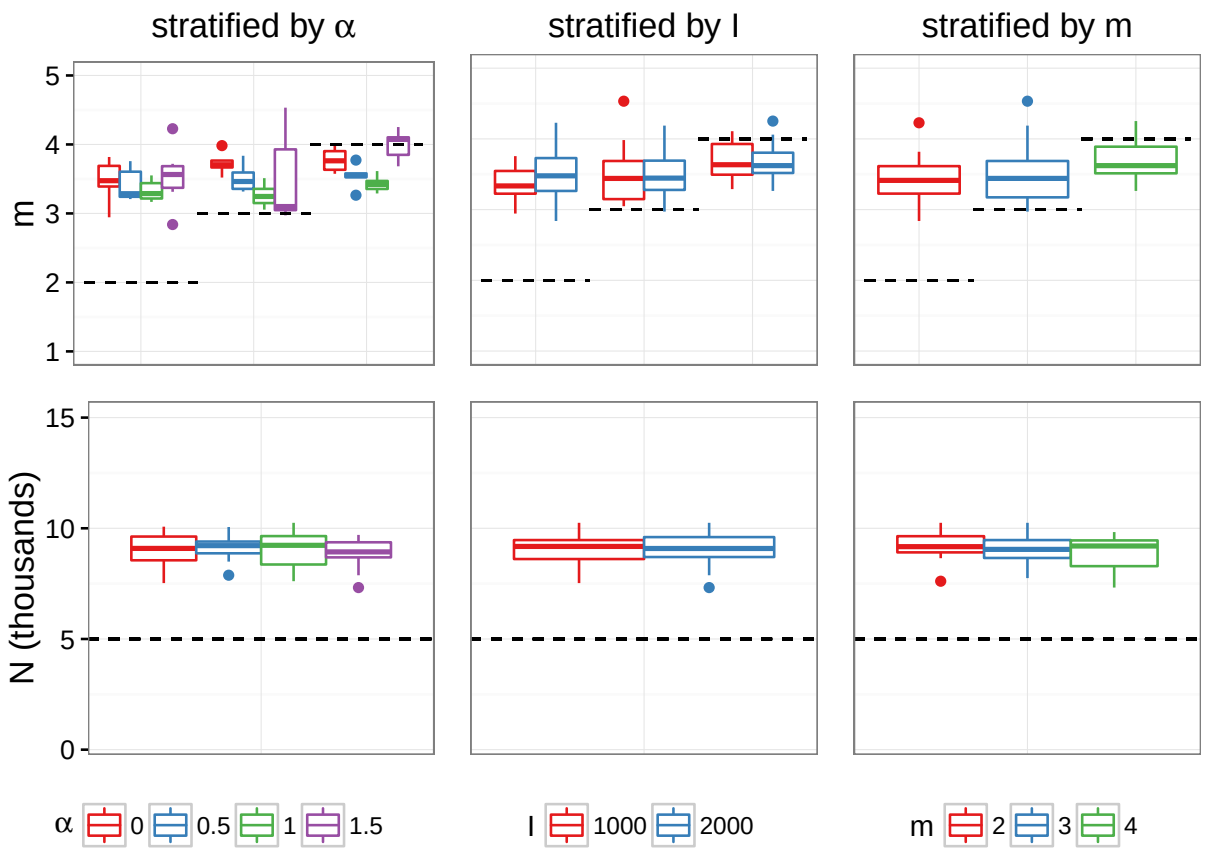


Figure B.23: Posterior mean point estimates for BA model parameters $m$ and $N$ obtained by running *netabc* on simulated data, stratified by true parameter values. First row of plots contains true versus estimated values of $m$; second row contains true versus estimated values of $N$. Columns are stratified by $\alpha$, $I$, and $m$ respectively. Dashed lines indicate true values.

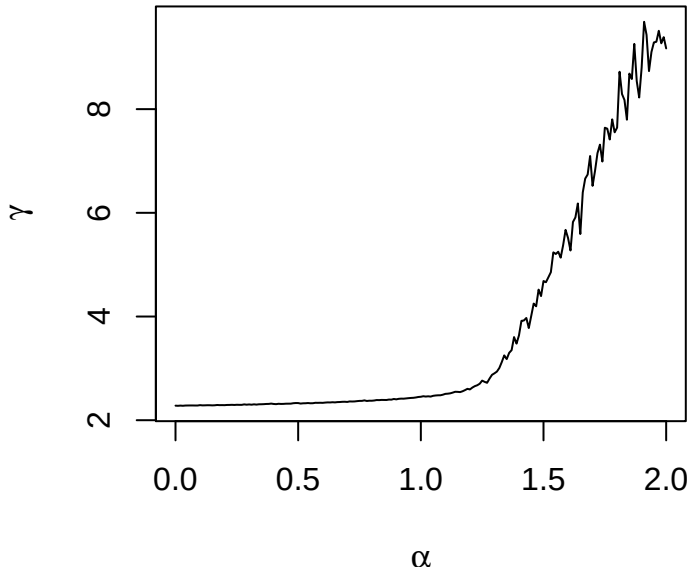


Figure B.24: Relationship between preferential attachment power parameter $\alpha$ and fitted power law exponent $\gamma$ for networks simulated under the BA network model with $N = 5000$ and $m = 2$.

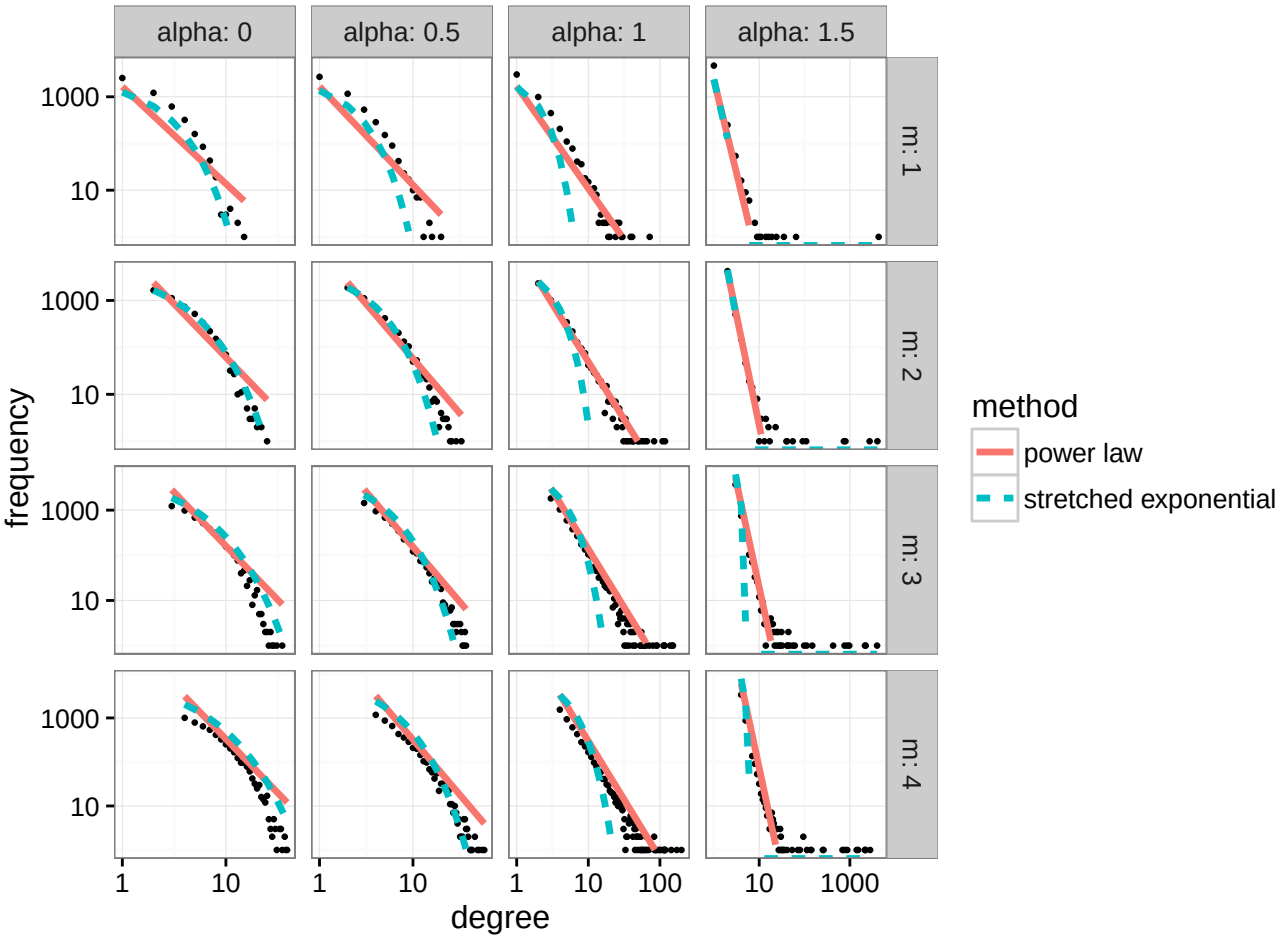


Figure B.25: Best fit power law and stretched exponential curves for degree distributions of simulated Barabási-Albert networks for several values of $\alpha$ and $m$.

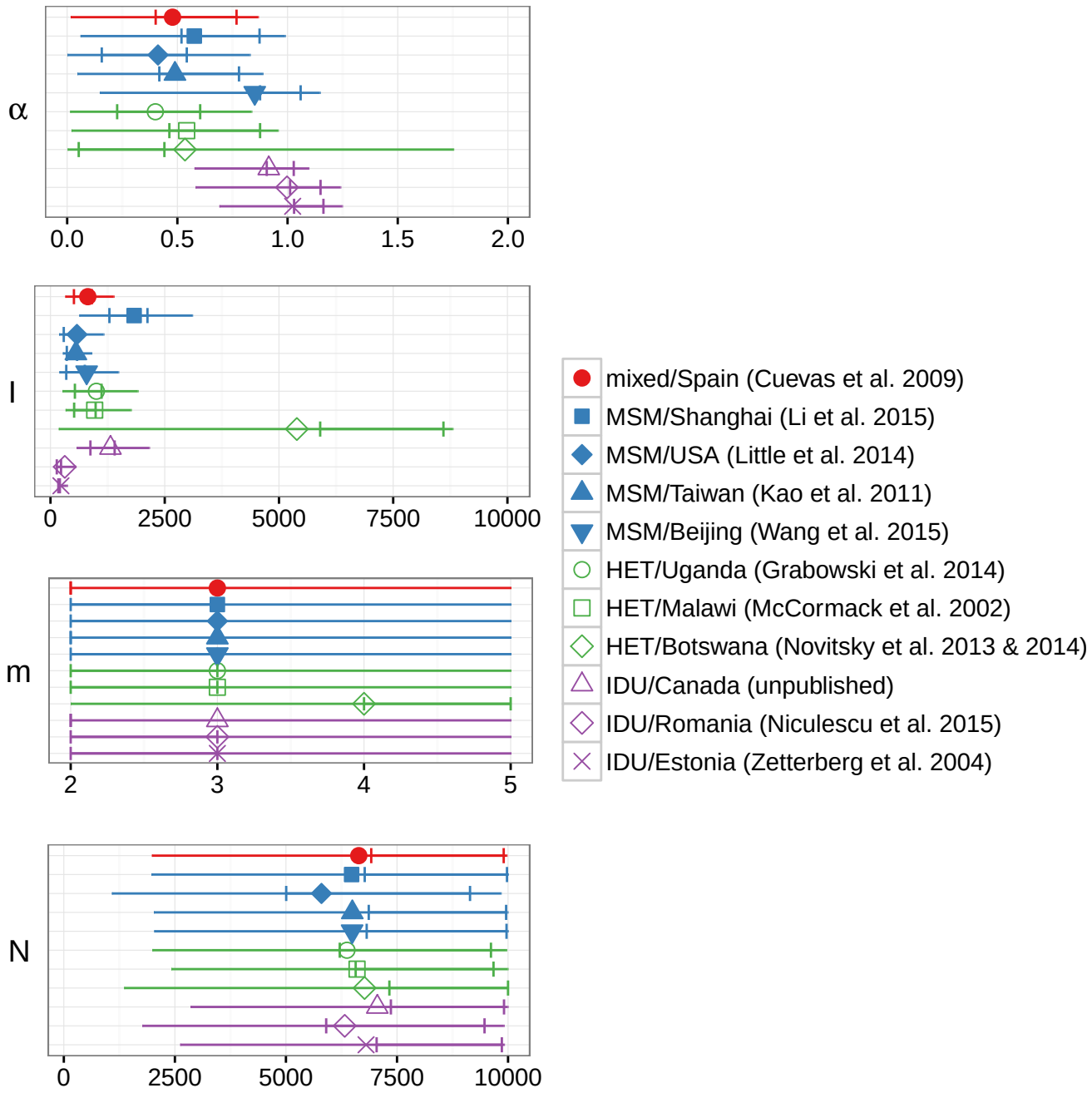


Figure B.26: Posterior mean point estimates and 95% HPD intervals for parameters of the BA network model, fitted to five published HIV datasets with *netabc* using the prior $m \sim \text{DiscreteUniform}(2,5)$. $x$-axes indicate regions of nonzero prior density.