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 Graduate Program)

THE UNIVERSITY OF BRITISH COLUMBIA
(Vancouver)

August 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 ~~network models~~ 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 ~~fit~~ investigate 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"). ~~to simulated transmission trees and applied it to viral phylogenies estimated from six real-world HIV sequence datasets.~~ 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 six real world HIV datasets, and substantial heterogeneity in the parameter estimates was observed. ~~Point estimates~~Depending on the choice of prior, posterior means for the preferential attachment power ranged from ~~0.06 to 1.05~~ 0.26, slightly less than logarithmic, to 1.00, exactly linear. Point estimates of the strength of preferential attachment were higher in injection drug user populations, potentially indicating that high-degree "hub" nodes may play a role in epidemics among this risk group. Our results underscore the importance of considering contact structures when ~~performing phylodynamic inference~~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 α 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).

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

Abstract . . . ii

Preface . . . iii

Table of Contents . . . v

List of Tables . . . vi

List of Figures . . . x

List of Symbols . . . xi

List of Abbreviations . . . xii

Acknowledgements . . . xiii

1 Introduction . . . **1**

1.1 Objective . . . 1

1.2 Phylogenetics and phylodynamics . . . 4

1.2.1 Phylogenetic trees . . . 4

1.2.2 Transmission trees . . . 4

1.2.3 Phylodynamics: linking evolution and epidemiology . . . 6

1.2.4 Tree shapes . . . 7

1.3 Contact networks . . . 9

1.3.1 Overview . . . 9

1.3.2 Scale-free networks and preferential attachment . . . 12

1.3.3 Relationship between network structure and transmission trees . . . 13

1.4 Sequential Monte Carlo . . . 14

1.4.1 Overview and notation . . . 14

1.4.2 Sequential importance sampling . . . 16

1.4.3 Sequential Monte Carlo . . . 17

1.4.4 The sequential Monte Carlo sampler . . . 18

1.5 Approximate Bayesian computation . . . 21

1.5.1 Overview and motivation . . . 21

1.5.2 Algorithms for ABC 23
1.6 Summary 25

2 Reconstructing contact network parameters from viral phylogenies 27
2.1 Discussion 28
2.1.1 *Netabc*: uses, limitations, and possible extensions 28
2.1.2 Analysis of Barabási-Albert model with synthetic data 29
2.1.3 Application to real world HIV data 32

3 Conclusion 36

Bibliography 37

Appendix A Mathematical models, likelihood, and Bayesian inference 48

Appendix B Additional plots 51

List of Tables

List of Figures

1.1 Illustration of a rooted, ultrametric, time-scaled phylogeny. 5

1.2 Illustration of a contact network and transmission tree. 6

1.3 Examples of Barabási-Albert networks with preferential attachment power $\alpha = 0$, 1, and 2. 13

2.1 First and second derivatives of epidemic growth curves at time of sampling for various values of I and N. 31

A.1 Reproduction of Figure 1A from Leventhal *et al.* (2012) used to check the accuracy of our implementation of Gillespie simulation. 51

A.2 Approximation of mixture of Gaussians used by Del Moral *et al.* (2012) and Sisson *et al.* (2009) to test adaptive approximate Bayesian computation (ABC)-sequential Monte Carlo (SMC). 52

A.3 Approximation of mixture of two Gaussians used to test convergence of adaptive ABC-SMC algorithm to a bimodal distribution. 53

A.4 Simulated transmission trees under three different values of BA parameter I 54

A.5 Simulated transmission trees under three different values of BA parameter m 55

A.6 Simulated transmission trees under three different values of BA parameter N 56

A.7 Cross validation accuracy of classifiers for Barabási-Albert (BA) model parameter α for eight epidemic scenarios. 57

A.8 Cross validation accuracy of classifiers for BA model parameter I for eight epidemic scenarios. 58

A.9 Cross validation accuracy of classifiers for BA model parameter m for eight epidemic scenarios. 59

A.10 Cross validation accuracy of classifiers for BA model parameter N for eight epidemic scenarios. 60

A.11 Kernel principal components projection of trees simulated under three different values of BA parameter α, for eight epidemic scenarios. 61

A.12 Kernel principal components projection of trees simulated under three different values of BA parameter I, for eight epidemic scenarios. 62

A.13 Kernel principal components projection of trees simulated under three different values of BA parameter m, for eight epidemic scenarios. 63

A.14 Kernel principal components projection of trees simulated under three different values of BA parameter N, for eight epidemic scenarios. 64

A.15 Grid search kernel scores for testing trees simulated under various α values. The other BA parameters were fixed at $I = 1000$, $N = 5000$, and $m = 2$. 65

A.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 $m = 2$. 66

A.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 $N = 5000$. 67

A.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 $m = 2$. 68

A.19 Point estimates of preferential attachment power α of Barabási-Albert network model, obtained on simulated trees with kernel-score-based grid search. 69

A.20 Point estimates of prevalence at time of sampling I of Barabási-Albert network model, obtained on simulated trees with kernel-score-based grid search. 70

A.21 Point estimates of number of edges per vertex m of Barabási-Albert network model, obtained on simulated trees with kernel-score-based grid search. 71

A.22 Point estimates of number of edges per vertex N of Barabási-Albert network model, obtained on simulated trees with kernel-score-based grid search. 72

A.23 Maximum *a posteriori* point estimates for BA model parameters obtained by running *netabc* on simulated data, for simulations with $m = 3$. 73

A.24 Maximum *a posteriori* point estimates for BA model parameters obtained by running *netabc* on simulated data, for simulations with $m = 3$. 74

A.25 Relationship between preferential attachment power parameter α and power law exponent γ for networks simulated under the BA network model with $N = 5000$ and $m = 2$. 75

A.26 Best fit power law and stretched exponential curves for degree distributions of simulated Barabási-Albert networks for several values of α and m. 75

A.27 Approximate marginal posterior distributions of BA model parameters obtained using kernel-assisted ABC for a network with heterogeneous node behaviour. 76

A.28 Approximate marginal posterior distributions of Barabási-Albert model parameters obtained using kernel-assisted ABC for a network with peer-driven sampling. 77

A.29 Maximum *a posteriori* point estimates and 95% HPD intervals for parameters of the BA network model, fitted to five published HIV datasets with *netabc* using the prior $m \sim DiscreteUniform(2,5)$. 78

A.30 Approximate marginal posterior distributions of BA model parameters for BC data. . . 79

A.31 Approximate marginal posterior distributions of BA model parameters for Cuevas et al. [146] data. 80

A.32 Approximate marginal posterior distributions of BA model parameters for Li et al. [148] data. 81

A.33 Approximate marginal posterior distributions of BA model parameters for Niculescu et al. [145] data. 82

A.34 Approximate marginal posterior distributions of BA model parameters for Novitsky et al. [147] and Novitsky et al. [149] data. 83

A.35 Approximate marginal posterior distributions of BA model parameters for Wang et al. [30] data. 84

A.36 Approximate marginal posterior distributions of BA model parameters for a simulated transmission tree with $\alpha = 0.0$, $I = 1000$, $m = 2$, and $N = 5000$. 85

A.37 Approximate marginal posterior distributions of BA model parameters for a simulated transmission tree with $\alpha = 0.5$, $I = 1000$, $m = 2$, and $N = 5000$. 86

A.38 Approximate marginal posterior distributions of BA model parameters for a simulated transmission tree with $\alpha = 1.0$, $I = 1000$, $m = 2$, and $N = 5000$. 87

A.39 Approximate marginal posterior distributions of BA model parameters for a simulated transmission tree with $\alpha = 1.5$, $I = 1000$, $m = 2$, and $N = 5000$. 88

A.40 Approximate marginal posterior distributions of BA model parameters for a simulated transmission tree with $\alpha = 0.0$, $I = 2000$, $m = 2$, and $N = 5000$. 89

A.41 Approximate marginal posterior distributions of BA model parameters for a simulated transmission tree with $\alpha = 0.5$, $I = 2000$, $m = 2$, and $N = 5000$. 90

A.42 Approximate marginal posterior distributions of BA model parameters for a simulated transmission tree with $\alpha = 1.0$, $I = 2000$, $m = 2$, and $N = 5000$. 91

A.43 Approximate marginal posterior distributions of BA model parameters for a simulated transmission tree with $\alpha = 1.5$, $I = 2000$, $m = 2$, and $N = 5000$. 92

A.44 Approximate marginal posterior distributions of BA model parameters for a simulated transmission tree with $\alpha = 0.0$, $I = 1000$, $m = 3$, and $N = 5000$. 93

A.45 Approximate marginal posterior distributions of BA model parameters for a simulated transmission tree with $\alpha = 0.5$, $I = 1000$, $m = 3$, and $N = 5000$. 94

A.46 Approximate marginal posterior distributions of BA model parameters for a simulated transmission tree with $\alpha = 1.0$, $I = 1000$, $m = 3$, and $N = 5000$. 95

A.47 Approximate marginal posterior distributions of BA model parameters for a simulated transmission tree with $\alpha = 1.5$, $I = 1000$, $m = 3$, and $N = 5000$. 96

A.48 Approximate marginal posterior distributions of BA model parameters for a simulated transmission tree with $\alpha = 0.0$, $I = 2000$, $m = 3$, and $N = 5000$. 97

A.49 Approximate marginal posterior distributions of BA model parameters for a simulated transmission tree with $\alpha = 0.5$, $I = 2000$, $m = 3$, and $N = 5000$. 98

A.50 Approximate marginal posterior distributions of BA model parameters for a simulated transmission tree with $\alpha = 1.0$, $I = 2000$, $m = 3$, and $N = 5000$. 99

A.51 Approximate marginal posterior distributions of BA model parameters for a simulated transmission tree with $\alpha = 1.5$, $I = 2000$, $m = 3$, and $N = 5000$. 100

A.52 Approximate marginal posterior distributions of BA model parameters for a simulated transmission tree with $\alpha = 0.0$, $I = 1000$, $m = 4$, and $N = 5000$. 101
A.53 Approximate marginal posterior distributions of BA model parameters for a simulated transmission tree with $\alpha = 0.5$, $I = 1000$, $m = 4$, and $N = 5000$. 102
A.54 Approximate marginal posterior distributions of BA model parameters for a simulated transmission tree with $\alpha = 1.0$, $I = 1000$, $m = 4$, and $N = 5000$. 103
A.55 Approximate marginal posterior distributions of BA model parameters for a simulated transmission tree with $\alpha = 1.5$, $I = 1000$, $m = 4$, and $N = 5000$. 104
A.56 Approximate marginal posterior distributions of BA model parameters for a simulated transmission tree with $\alpha = 0.0$, $I = 2000$, $m = 4$, and $N = 5000$. 105
A.57 Approximate marginal posterior distributions of BA model parameters for a simulated transmission tree with $\alpha = 0.5$, $I = 2000$, $m = 4$, and $N = 5000$. 106
A.58 Approximate marginal posterior distributions of BA model parameters for a simulated transmission tree with $\alpha = 1.0$, $I = 2000$, $m = 4$, and $N = 5000$. 107
A.59 Approximate marginal posterior distributions of BA model parameters for a simulated transmission tree with $\alpha = 1.5$, $I = 2000$, $m = 4$, and $N = 5000$. 108

List of Symbols

γ exponent of power-law degree distribution in scale-free networks.

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

N total number of nodes in a contact network.

α preferential attachment power parameter in Barabási-Albert networks.

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

List of Abbreviations

ABC approximate Bayesian computation.

BA Barabási-Albert.

ER Erdős-Rényi.

ERGM exponential random graph model.

ESS expected sample size.

HIV human immunodeficiency virus.

HMM hidden Markov model.

HPD highest posterior density.

IDU injection drug users.

IS importance sampling.

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.

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

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,

$$\Pr(\theta \mid T) = \frac{\Pr(T \mid \theta)\Pr(\theta)}{\Pr(T)},$$

where $\Pr(T \mid \theta)$ is the likelihood, $\Pr(\theta)$ is the prior, and $\Pr(T)$ is the marginal probability of T which acts as a normalizing constant on the posterior (see appendix A for a review of mathematical modeling and Bayesian inference, including definitions of these concepts). As we shall show (see **??**), estimating this distribution presents computational challenges beyond those usually encountered in Bayesian inference. Both the likelihood $\Pr(T \mid \theta)$ and the normalizing constant $\Pr(T)$ seem to be intractable, which rules out the use of most common maximum likelihood and Bayesian methods.

~~Calculating the likelihood of the parameters of a contact network model seems likely to be an intractable problem. We have not proven this is the case, but some intuition can be provided by examining the process involved in the likelihood calculation. Consider a contact network model with parameters θ and an estimated transmission tree T with n tips. In general, we do not know the labels of the internal nodes of T, only the labels of its tips. To fit this model using likelihood-based methods, we must calculate the likelihood of θ, that is, $\Pr(T \mid \theta)$. Let $\mathscr{G}$ be the set of all possible contact networks, and $\mathscr{N}$ be the set of all possible labellings of the internal nodes of T. We can write the likelihood as~~

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

~~the last equality following from the fact that T and v depend only on G, not on θ. Although $\Pr(T, v \mid G)$ and $\Pr(G \mid \theta)$ may individually be straightforward to calculate, the number of possible directed graphs on N nodes is $2^{N(N-1)}$ [13], larger if the nodes and edges in the graph may have different labels or attributes. Hence, the number of terms in the sum is at least exponential in n, as there must be at least n nodes in the network. In addition,~~ eq. (1.1) ~~assumes that T is complete, meaning that all infected individuals were sampled. This is rarely the case in practice - most often, we only have access to a subset of the infected individuals. In this case, the likelihood calculation becomes even more complex, because we must also sum over all possible complete trees.~~

~~Depending on the network model studied, it is possible that~~ eq. (1.1) ~~could be simplified into a tractable expression.~~ An alternative to likelihood-based methods, which could be applied to any network model, is provided by approximate Bayesian computation (ABC) [14–17]. 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" [18]; similar kernel functions have been demonstrated to work well as distance functions in ABC [19]. 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 [20]. A recently-developed adaptive algorithm requiring minimal tuning on the part of the user makes SMC an even more attractive approach [21]. 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 [22], 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 [23–26], meaning that their degree distributions follow a power law (although there has been some disagreement, see [6, 27]). Preferential attachment has been postulated as a mechanism by which scale-free networks could be generated [28]. The Barabási-Albert (BA) model [28] 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, human immunodeficiency virus (HIV) is one of the most commonly-studied viruses in a phylodynamic context. Consequently, a large volume of HIV sequence data is publicly available, more than for any other pathogen, and including sequences sampled from diverse geographic and demographic 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 [26]. 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 [29, 30]. 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*) [31]. 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, eg. [32]). The edges or *branches* of the phylogeny connect ancestors to their descendants. Phylogenies may have a *root*, which is a node with no descendants distinguished as the most recent common ancestor of all the extant taxa [33]. 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* [34].

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 [35]. 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 [36]. These concepts are illustrated in fig. 1.1.

1.2.2 Transmission trees

In epidemiology, a *transmission tree* is a graphical representation of an epidemic's progress through a population [37]. 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 [38]. 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 [39], 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

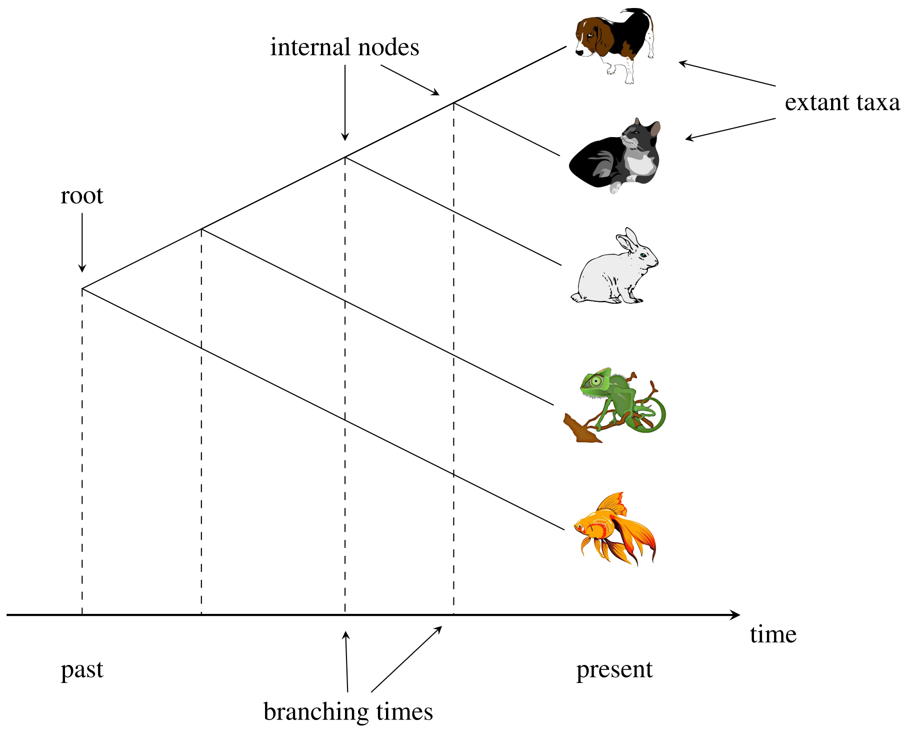

Figure 1.1: Illustration of a rooted, ultrametric, time-scaled phylogeny. The tips of the tree, which represent extant taxa, are placed at the present day on the time axis. Internal nodes, representing extinct common ancestors to the extant taxa, fall in the past. The topology of the tree indicates that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

routes of disease transmission. Contact networks, and their relationships to transmission trees, will be 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, 40]. 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.* 41]. In this work, we reserve the term “transmission tree” for the objects depicted on the right side of fig. 1.2, following *e.g.* [38]. The term “transmission network” is taken to mean the subgraph of the contact network along which transmissions occurred, following *e.g.* [8, 40].

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 [35], 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 [42]. However, in all but the most well-studied of epidemics, transmission trees are not possible to assemble through traditional epidemiological methods [40]. The time and effort to conduct detailed interviews and contact tracing of a sufficient number

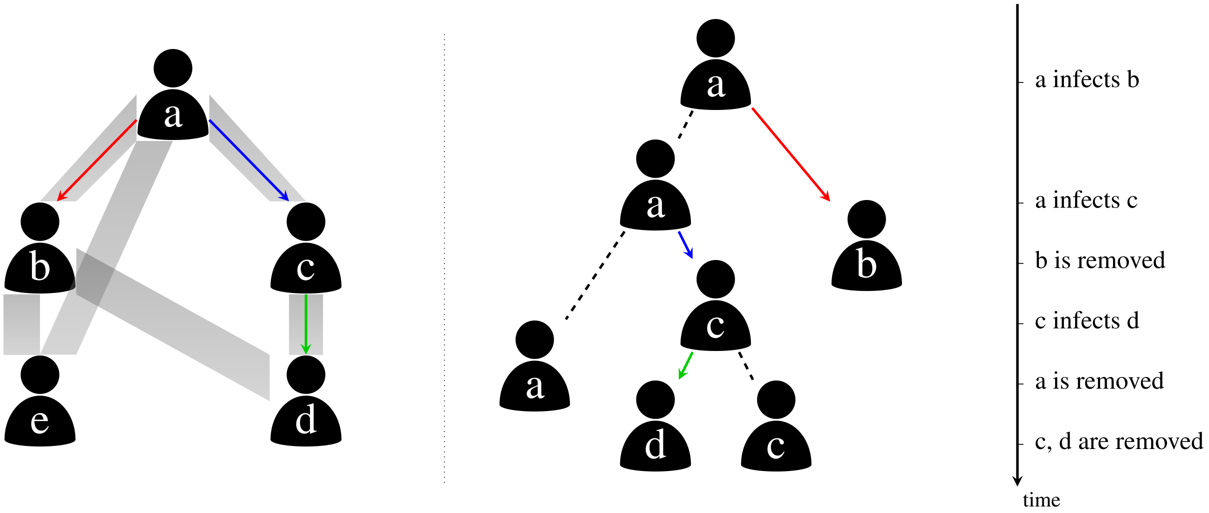

Figure 1.2: Illustration of epidemic spread over a contact network, and the corresponding transmission tree. (Left) A contact network with five hosts, labelled *a* through *e*. Thick shaded edges indicate symmetric contacts among the hosts. The transmission network is indicated by coloured arrows. The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

of infected individuals is usually prohibitive, and may additionally be confounded by misreporting and other challenges [43]. 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* [44] 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 [39]. 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 [45, 46] or coalescent [47, 48] 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.* [49] for a full review.

Due to the relationship between the aforementioned processes, there is a degree of correspondence between viral phylogenies and transmission trees [37, 41, 50, 51]. In particular, the transmission process is quite similar to *allopatric speciation* [52], 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 [53]. 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 [54]. Modern likelihood-based methods of phylogenetic reconstruction [*e.g.* 55, 56] 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 [57–59], which we apply in this work, and least-square dating [60]. Alternatively, the tree may be rooted separately with an outgroup [61] 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 [37, 62]. 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. [63] and Paraskevis et al. [64] were able to accurately recover known transmission trees using viral phylogenies. The problem of accurately estimating transmission trees is an ongoing area of research [32, 54, 65–68]. For example, Hall, Woolhouse, and Rambaut [54] 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 [69]. 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 [70], which measures the imbalance or asymmetry in a rooted tree. For the ith tip of the tree, we define N_i to be the number of branches between that tip and the root. The unnormalized Sackin's index is defined as the sum of all N_i. It is called unnormalized because it does not account for the number of tips in the tree. Among two trees having the same number of tips, the least-balanced tree will have the highest Sackin's index. However, among two equally balanced trees, the larger tree will have a higher Sackin's index. This makes it challenging to compare balances among trees of different sizes. To correct this, Kirkpatrick and Slatkin [71] derive the expected value of Sackin' index under the Yule model [72]. 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) [73], which compares the LTT [35] 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 [42], so large values may indicate that the trees being compared were produced by different epidemic trajectories [73].

Poon et al. [18] developed an alternative pairwise measure which applies the concept of a *kernel function* to phylogenies. Kernel functions, originally developed for support vector machines (SVMs) [74], compare objects in a space $\mathscr{X}$ by mapping them into a feature space $\mathscr{F}$ of high or infinite dimension via a function φ. The similarity between the objects is defined as

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

that is, the inner product of the objects' representations in the feature space. Computing $\varphi(x)$ may be computationally prohibitive due to the dimension of $\mathscr{F}$. The utility of a kernel function K is that it is constructed in such a way that it can compute the inner product without explicitly computing $\varphi(x)$. The kernel function developed in [18] will henceforth be referred to as the *tree kernel*. This kernel maps trees into the space of all possible 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 [75] and did not incorporate branch length information. The version developed by Poon et al. [18] 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. [18], let $N(T)$ denote the set of all nodes in T, $\text{nc}(n)$ be the number of children of node n, c_n^j be the jth child of node n, and l_n be the vector of branch lengths connecting node n to its $\text{nc}(n)$ children. Furthermore, let $\text{nl}(n)$ be the number of children of n which are leaves (we always have $\text{nl}(n) \leq \text{nc}(n)$). The *production rule* of n is the pair $(\text{nc}(n), \text{nl}(n))$. 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 σ is a variance parameter. The tree kernel is defined as

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

where

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

The parameter λ in the above expression, is called the *decay factor* [76], 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. λ penalizes Δ 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 λ and σ 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 [22]. 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, 77, 78]. 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* [79, 80], 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 81].

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 [82]. 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 [81]. 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 [83]. However, these models still assume that every possible contact between a member of group i and a member of group j is equally likely. This assumption is clearly unrealistic for the majority of human communities and can lead to errors in predicted epidemic trajectories when there is substantial heterogeneity present [6, 84, 85]. 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 reproductive number of various outbreaks, which have agreed with values obtained by on-the-ground data collection [86]. 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 [81].

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, 87]. 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 [88]. 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, 89]. 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, 90]. 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 [29, 30].

Contact networks are a challenging type of data to collect, requiring extensive epidemiological investigation in the form of contact tracing [8, 40, 43, 78]. 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 [81]. 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 [23–25], so dense or respondent-driven sampling [91] 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 [92]. However, this network may show little to no agreement with contact data obtained through epidemiological methods [93–95] and therefore may be a poor proxy for the contact network. Other studies [96] 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 [40] (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 [96]. 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 [97] 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 [98] 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 [87] 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 [84] later extended the method to dynamic contact networks and applied it to a sexual network relating 99 individuals investigated during a syphilis outbreak.

1.3.2 Scale-free networks and preferential attachment

A *scale-free* network is one whose degree distribution follows a power law, meaning that the number of nodes in the network with degree k is proportional to $k^{-\gamma}$ for some constant γ [28]. 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 [23–26]. Interestingly, many other types of network, including computer networks [89], biological metabolic networks [99], and academic co-author networks [100], 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* [101], which have been postulated to contribute in varying degree to the spread of diseases such as HIV [38] and severe acute respiratory syndrome (SARS) [102]. Scale-free networks have no epidemic threshold [89], meaning that diseases with arbitrarily low transmissibility ~~can persist~~ 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* [28, 103]. The simplest preferential attachment model is known as the Barabási-Albert (BA) model after its inventors [28]. 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 Jones [27] 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. **drombowski2013** 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 α. 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 β. 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

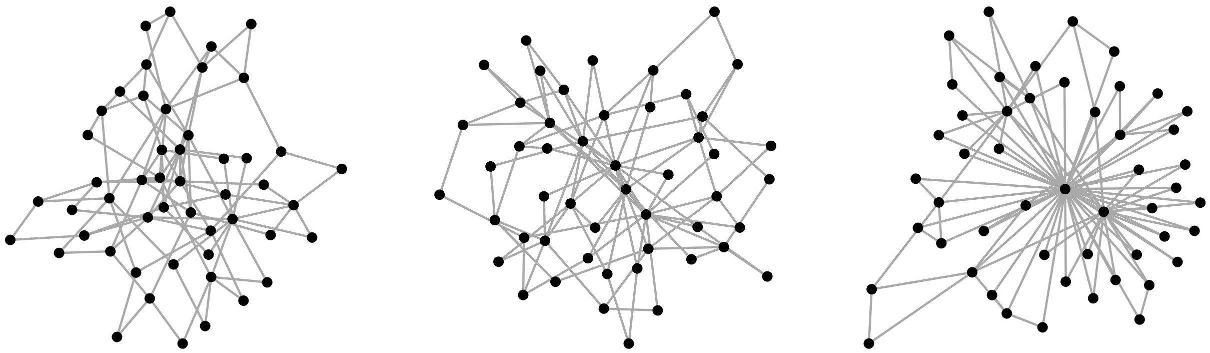

Figure 1.3: Examples of Barabási-Albert networks with preferential attachment power $\alpha = 0$ (left), 1 (centre), and 2 (right). All networks have $N = 50$ nodes and were constructed with $m = 2$ edges per vertex. When $\alpha = 0$, attachments are formed at random and most nodes have low degree. 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.

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 α 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 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. [53]. 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. In particular, for a network model with parameters θ and an input transmission tree T, we are interested in the posterior distribution

$$\Pr(\theta \mid T) = \frac{\Pr(T \mid \theta)\Pr(\theta)}{\Pr(T)}, \tag{1.3}$$

where we have adopted the common abuse of the symbol Pr to refer to a probability density. As alluded to in section 1.1, both the likelihood $\Pr(T \mid \theta)$ and the normalizing constant $\Pr(T)$ are likely computationally intractable (this will be explained further in **??**). 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 π, regardless of its complexity, can always be *represented* by a discrete sample from it. By "represented", we mean that any computation of expectations using π 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 is also known as the *particle filter*. Rather than sampling points one at a time from the target distribution, SMC considers a population of points or “particles”, here denoted $\{x^{(k)}\}$ and indexed by an integer k. The particles are associated with weights, $\{w(x^{(k)})\}$. These weighted particles are a representation for the target distribution. For example, the expected value of the distribution is approximated by

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

where n is the number of particles. 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 sample from the target.

In this work, the distribution of interest is the posterior distribution 1.3. The particles are particular values of the parameters θ 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 $\Pr(\theta \mid T)$. 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.

The goal of this section of the introduction is to describe an algorithm called the SMC sampler [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.

~~Sequential Monte Carlo (SMC) is the name for a family of statistical inference methods that rely on approximating probability distributions of interest with large collections of *particles*, here denoted $\{x^{(k)}\}$~~[115–117]. ~~These collections or *populations* are constructed to form a *Monte Carlo approximation* to some distribution of interest π, meaning that the empirical distribution of the particles converges in distribution to π as the population size gets large~~[113]. ~~The word *sequential* is used because the particle populations are modified in an iterative fashion over time, for example, to incorporate new evidence.~~

To fully describe SMC, we will introduce some notation and terminology. The definitions of these terms will become clearer as they are used. For a sequence $x_1, \ldots, x_d$, 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 kth particle in a population. To ease the notational burden we will omit the superscripts and subscripts on the weight functions w.

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

$$\int_Y K(x,y)\mathrm{d}y = 1 \tag{1.4}$$

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).

1.4.2 Sequential importance sampling

Sequential importance sampling (SIS) [120] is a particle-based method whose aim is to sample from a distribution π on an high-dimensional space, say $\pi(\mathbf{x}) = \pi(x_1, \ldots, x_d)$. The basis of SIS is importance sampling (IS), which is a method of estimating summary statistics of distributions which are known only up to a normalizing constant, and therefore cannot be sampled from directly. That is, if π is such a distribution and f is any real-valued function, IS is concerned with estimating

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

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

$$\int f(x)\pi(x)\mathrm{d}x = \frac{1}{Z}\int w(x)f(x)\mathrm{d}x. \tag{1.5}$$

Since η can be sampled from exactly, and γ and f can both be evaluated pointwise, the integral $\int w(x)f(x)\mathrm{d}x$ can be approximated by a Monte Carlo estimate. 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. (1.5). Although this is a simple and elegant approach, the drawback is that the variance of the estimate is proportional to the variance of the importance weights [117], which may be quite large if η and γ are very different. Therefore, the practical use of IS on its own is limited, since it depends on finding an importance distribution similar to π, which we usually know very little about *a priori*.

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

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

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

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

The importance weights for η can be written recursively as

$$w(\mathbf{x_i}) = \frac{\pi(\mathbf{x_i})}{\eta(\mathbf{x_i})} = \frac{\pi(x_i \mid \mathbf{x_{i-1}})\pi(\mathbf{x_{i-1}})}{\eta(x_i \mid \mathbf{x_{i-1}})\eta(\mathbf{x_{i-1}})} = \frac{\pi(x_i \mid \mathbf{x_{i-1}})}{\eta(x_i \mid \mathbf{x_{i-1}})} \cdot w(\mathbf{x_{i-1}}). \tag{1.6}$$

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

Algorithm 1 Sequential importance sampling.

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

1.4.3 Sequential Monte Carlo

The importance distribution η constructed with SIS is merely an approximation to π, and may be a fairly poor one in practice depending on the application. Try as we might to keep the variances of the weights low, the cumulative errors at each sequential step tend to push many of the weights to very low values [115]. This results in a poor approximation to π, 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 rejuvination*, 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.

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
    $w^{(k)} \leftarrow \pi(x_1^{(k)}) / \eta(x_1^{(k)})$
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
        $w(\mathbf{x_i}^{(k)}) \leftarrow [\pi(x_i^{(k)} \mid \mathbf{x_{i-1}^{(k)}}) / \eta(x_i^{(k)} \mid \mathbf{x_{i-1}^{(k)}})] \cdot w(\mathbf{x_{i-1}}^{(k)})$
    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
```

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. When the ESS drops below the threshold (conventionally $n/2$ [117]), particles are resampled according to their importance weights. This results in the removal of low-weight particles from the population, and also equalizes all the weights. Various resampling strategies beyond the basic sampling with replacement have been proposed [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 $\pi(x)$, by sequentially sampling from d distributions of lower but increasing dimension. Del Moral, Doucet, and Jasra [118] developed an *SMC sampler* with an alternative objective: to sample sequentially from d distributions $\pi_1, \ldots, \pi_d$, all of the same dimension and defined on the same space. The π_i are assumed to form a related sequence, such as posterior distributions attained by sequentially considering new evidence. As with SIS, we assume that $\pi_i(x) = \gamma_i(x)/Z_i$, where γ_i is known pointwise and the normalizing constant Z_i is unknown.

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

$$\eta_i(x') = \int \eta_{i-1}(x) K_i(x, x') \mathrm{d}x, \tag{1.7}$$

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

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

where $q(x, x')$ 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 η appears straightforward, the drawback is that the importance distribution itself becomes intractable. In particular, evaluating $\eta_i(x)$ involves a i-dimensional integral of the type in eq. (1.7). As it is necessary to evaluate $\eta(x)$ pointwise to perform IS, this construction appears to have defeated the purpose of providing an importance distribution for each π_i. Del Moral, Doucet, and Jasra [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. This brings us back to the setting described above in section 1.4.2, namely of building up an importance distribution of dimension d sequentially through lower-dimensional distributions. We can write $\tilde{\pi}_i$ in terms of $\tilde{\pi}_{i-1}$ by noticing that

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

and hence

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

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

$$w(\mathbf{x_i}) = \frac{\tilde{\pi}_i(\mathbf{x_i})}{\eta_i(\mathbf{x_i})} \tag{1.8}$$

$$= \frac{\tilde{\pi}_{i-1}(\mathbf{x_{i-1}})\pi_i(x_i)L(x_i,x_{i-1})}{\eta_{i-1}(\mathbf{x_{i-1}})\pi_{i-1}(x_{i-1})K_i(x_{i-1},x_i)} \tag{1.9}$$

$$= w(\mathbf{x_{i-1}}) \cdot \frac{\pi_i(x_i)L_{i-1}(x_i,x_{i-1})}{\pi_{i-1}(x_{i-1})K_i(x_{i-1},x_i)} \tag{1.10}$$

$$\propto w(\mathbf{x_{i-1}}) \cdot \frac{\gamma_i(x_i)L_{i-1}(x_i,x_{i-1})}{\gamma_{i-1}(x_{i-1})K_i(x_{i-1},x_i)}. \tag{1.11}$$

The final key piece of information is to notice that, because the L_i are Markov kernels, π_i is simply the marginal in $\mathbf{x_{i-1}}$ of $\tilde{\pi}$. Therefore, a sample from $\tilde{\pi}_i$ automatically gets us a sample from π_i, by considering only the ith component of $\mathbf{x_i}$. In fact, since the weight update eq. (1.11) depends only on the ith 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. These are all the ingredients we need to apply SIS. The sequences of kernels L and K should be chosen based on the problem at hand to minimize the variance in the importance weights as well as possible. For a fixed choice of K_i, the backward kernels L_i which minimize this variance are called the *optimal* backward kernels. The full SMC sampler algorithm is presented as algorithm 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
    $w^{(k)} \leftarrow \gamma_1(x_1^{(k)})/\eta_1(x_1^{(k)})$
    Normalize the weights so that $\sum w = 1$
end for
for $i = 2$ to $d$ do
    for $k = 1$ to $n$ do
        Sample $x_i^{(k)}$ from $K(x_{i-1}^{(k)}, x_i)$                ▷ Extend the $k$th particle
        $w^{(k)} \leftarrow w^{(k)} \cdot \frac{\gamma_i(x_i)L_{i-1}(x_i,x_{i-1})}{\gamma_{i-1}(x_{i-1})K_i(x_{i-1},x_i)}$
    end for
    Normalize the weights so that $\sum w = 1$
    if ESS$(w) < T$ then                                        ▷ $T$ 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 (**??**), that the posterior distribution

$$\Pr(\theta \mid T) = \frac{\Pr(T \mid \theta)\Pr(\theta)}{\Pr(T)}$$

for a contact network model with parameters θ and an input transmission tree T does not fall in this category (note that, as in the previous section, we are using Pr to denote a probability density). 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 π which can be evaluated up to a normalizing constant Z, that is, $\pi(x) = \gamma(x)/Z$. Both algorithms calculate the ratio $\pi(x)/\pi(x') = \gamma(x)/\gamma(x')$ 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, it is required to calculate the importance weights. In the context of Bayesian inference, this ratio is a likelihood ratio, which must be calculated by computing the individual likelihoods and dividing them. If the likeilhood is intractible, this is clearly not a viable approach.

Approximate Bayesian computation (ABC) [14–16] was developed to estimate posterior distributions with intractable likelihoods, which have arisen frequently in the domain of population genetics [17, 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. This conveniently allows us to apply the existing MCMC and SMC 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.

~~Most mathematical models are amenable to fitting via one or both of the approaches, ML or Bayesian inference, discussed above. However, there are some, particularly in the domain of population genetics~~ [17, 125], ~~for which calculation of either the likelihood or the product of the likelihood and the prior may be infeasible. For example, one or both of these quantities may be expressible only as an intractable integral. ABC is designed for such cases, where standard likelihood-based techniques for model fitting cannot be applied.~~

~~Ordinarily, Bayesian inference targets the posterior distribution $\Pr(\theta \mid y)$. That is, in the Bayesian framework,~~ 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, namely the similarity of simulated datasets to the observed data. If datasets simulated under the model closely resemble the real data, it follows that the model is a reasonable approximation to the real-world process generating the observed data. More formally, let y be the observed data to which we are trying to fit a model with parameters θ. 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 ρ defined on the space of all possible data our model could generate. ABC aims to sample from the joint posterior distribution of model parameters and simulated datasets z which are within some small distance ε of the observed data y,

$$\pi_\varepsilon(\theta, z \mid y) = \frac{\Pr(\theta)\Pr(z \mid \theta)\mathbb{I}_{A_{\varepsilon,y}}(z)}{\int_{A_{\varepsilon,y}\times\Theta}\Pr(\theta)\Pr(z \mid \theta)\mathrm{d}\theta}. \tag{1.12}$$

Here, $A_{\varepsilon,y}$ is an ε-ball around y with respect to ρ, Θ is the space of all possible model parameters, and $\mathbb{I}$ is the indicator function [126]. The distribution $\pi_\varepsilon(\theta, z \mid y)$ will be referred to as the *ABC target distribution*. The term $\Pr(z \mid \theta)$ appears to be the bothersome likelihood again, but this will turn out not to be a problem because we are simulating z ourselves. In fact, it is possible to sample from this distribution exactly (see next section).

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 θ. z is a simulated 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 **??**). The distance function ρ must compare two transmission trees - the observed tree y, and the simulated tree z. We will define this distance function using the tree kernel discussed in section 1.2.4. 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 true tree. Otherwise, the probability density is zero.

~~As we shall see in the next section, this distribution can be sampled from exactly. The word "approximate" derives from assumption that, for a suitably chosen distance ρ and a small enough ε, the marginal in z of this distribution approximates the posterior of interest [126]. That is,~~ 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

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

The intuition for why this approximation might be reasonable comes from the fact that, when $\varepsilon = 0$, the ε-ball around y should contain only y itself, hence the integral on the left is exactly equal to the posterior. Thus, by taking ε small, we should attain something close to the posterior if ρ 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. [16, 129]), ρ 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

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

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].

Summary statistics can be useful if the data are high-dimensional or of a complex type, but 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. [19] proposed the use of a kernel function (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 kernel function to compute the weights of the particles. The work by Poon [22], 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 fall into one of 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 (θ, z), so that the ABC target distribution can be written $\pi_\varepsilon(x \mid y)$.

Rejection ABC is the simplest method, and also the one which was first proposed [14, 15]. The algorithm, outlined in algorithm 4, repeats the following steps until a desired number of samples from the target distribution are obtained. Parameter values θ are sampled according to the prior distribution $\pi(\theta)$. Then, a simulated dataset z is generated from the model with the sampled parameter values. By definition, the probability density of obtaining the particular dataset z is $f(z \mid \theta)$. Finally, the parameters are sampled if the distance of z from the observed data y is less than ε, that is, with probability $\mathbb{I}_{A_{\varepsilon,y}}(z)$. Putting this all together, the parameters θ are sampled with probability proportional to

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

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

Rejection ABC is easy to understand and implement, but it is not generally computationally feasible. If the posterior is very different from the prior, a very large number of samples may need to be taken in order to find a simulated dataset which is close to z. The inefficiency is compounded by the curse

Algorithm 4 Rejection ABC.

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

of dimensionality - the measure of the ε-ball around y decreases exponentially with the number of dimensions. ABC-MCMC (algorithm 5) was designed to overcome these hurdles [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

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

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

Some of the same computational inefficiencies arise with ABC-MCMC as with rejection. For example, in regions of low posterior density, the probability to simulate a dataset proximal to the observed data is low. Various strategies have been developed to mitigate this, including reducing the tolerance level ε as the chain progresses [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 ε_i. The intention is for the algorithm to progress smoothly through a sequence of target distributions which ends at the ABC approximation to the posterior. The initial value ε_0 is set to ∞, which makes the first distribution in the sequence

$$\pi_0(\theta,z) = \frac{\Pr(\theta)\Pr(z \mid \theta)}{\int_{\mathbb{R}\times\Theta}\Pr(\theta)\Pr(z \mid \theta)\mathrm{d}\theta\mathrm{d}z}.$$

For brevity, we have ommitted the dependence on the observed data y. In the terminology of algorithm 2, the numerator is the first of the γ's, that is, $\gamma_0 = \Pr(\theta)\Pr(z \mid \theta)$. Sampling in proportion to γ_0 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 is problem-specific, and so appropriate kernels must be chosen for ABC. Several options have been proposed [21, 131, 134].

Algorithm 6 ABC-SMC.

```
for $k = 1$ to $n$ do
    Sample $x_1^{(k)}$ from $\Pr(\theta)\Pr(z \mid \theta)$        ▷ Draw $\theta$ from the prior and simulate $z$ from $\theta$
    $w^{(k)} \leftarrow 1/n$
end for
for $i = 2$ to $i_{\max}$ do        ▷ $i_{\max}$ is a user-specified number of iterations
    for $k = 1$ to $n$ do
        Sample $x_i^{(k)}$ from $K(x_{i-1}^{(k)}, x_i)$        ▷ Extend the $k$th particle
        $w^{(k)} \leftarrow w^{(k)} \cdot \dfrac{\gamma_i(x_i)L_{i-1}(x_i, x_{i-1})}{\gamma_{i-1}(x_{i-1})K_i(x_{i-1}, x_i)}$
    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
```

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

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

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

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

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

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

1.6 Summary

Our method integrates the four distinct research areas just described: phylogenetics, contact networks, sequential Monte Carlo, and approximate Bayesian computation. The first two topics together form the problem domain. Phylogenetic data is 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

2.1 Discussion

2.1.1 *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 [29, 30]. Despite this, few methods exist for investigating contact network parameters in a phylodynamic framework [although see 87, 96, 98, 106, 135, for related work]. Kernel-ABC is a model-agnostic method which can be used to investigate any quantity that affects tree shape [22]. In this work, we developed *netabc*, a method based on kernel-ABC to infer the parameters of a contact network model. The method is general, meaning that it can be used to infer parameters of any network model, as long as it allows simulated networks can be easily generated. We have included generators for the BA model discussed here, as well as the ER and Watts-Strogatz (WS) network models. Instructions for adding additional models are available in the project's online documentation. We have made *netabc* publicly available at github.com/rmcclosk/netabc under a permissive open source license, to encourage other researchers to apply and extend our method.

Several alternative network models and modelling frameworks have been developed which may provide useful future targets for kernel-assisted ABC. Waring models [27, 136] 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. [96] to characterize the transmission network in the United Kingdom. Exponential random graph models (ERGMs) [137] 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. The author 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 kernel-assisted 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 [138] and the log-linear adjacency matrix parameterization applied by Morris [78].

The two-step process of simulating a contact network and subsequently allowing an epidemic to spread over that network carries with it the assumption that the contact network is static over the duration of the epidemic. Clearly this assumption is invalid, as people make and break partnerships on a regular basis. 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, 40]. In principle, kernel-assisted ABC could be adapted to dynamic contact networks by using a method such as that developed by Robinson, Cohen, and Colijn [139] 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. 63], and is frequently used in practice [e.g. 38], the topologies of a viral phylogeny and transmission tree can differ significantly [37, 54] 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, 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 [32, 54, 65, 66, 68], 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 Portable Operating System Interface (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.

2.1.2 Analysis of Barabási-Albert model with synthetic data

The preferential attachment power α had a very strong influence on tree shape in the range of values we considered (**????**). Although the tree kernel was the most effective classifier for α, a Sackin's index of tree imbalance performed nearly as well (**??**). This result was intuitive: high α values produce networks with few well-connected "superspreader" nodes which are involved in a large number of transmissions, resulting in a highly unbalanced ladder-like tree structure (**??**). There was no observable bias in the estimates of α obtained with *netabc*, however ~~the variation in these estimates was higher for $\alpha < 1$ than for $\alpha \geq 1$~~ there appeared to be weaker identifiability for $\alpha < 1$ than for $\alpha \geq 1$ (**????**). The relationship between α and the power law exponent γ may explain this result (fig. A.25). The γ values associated with $\alpha = 0$ 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 kernel-assisted ABC. The dynamics of the SI model, and the coalescent process [47], offer a potential explanation for the identifiability of I. 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 distribution of coalescence times in the tree should therefore be informative about I [53]. This information is captured by the tree kernel, and also by the nLTT statistic, which both performed quite well in classifying I (**??**).

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. There is an intuitive explanation for this result, namely that adding additional nodes does not change the edge density or overall shape of a BA network. This can be illustrated by imagining that we add a small number of nodes to a network after the epidemic simulation has already been completed. It is possible that none of these new nodes attains a connection to any infected node. Thus, running the simulation again on the new, larger network could produce the exact same transmission tree as before. On the other hand, when I is large relative to N, the coalescent dynamics discussed above also apply. That is, 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 (**??** and fig. A.10) corroborates this hypothesis, as the nLTT uses only information about the coalescence times. When all BA parameters were simultaneously estimated with kernel-assisted ABC, N was nearly always over-estimated by approximately a factor of two (**????**). 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.

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 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. 2.1, 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 kernel-assisted 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 (fig. A.20).

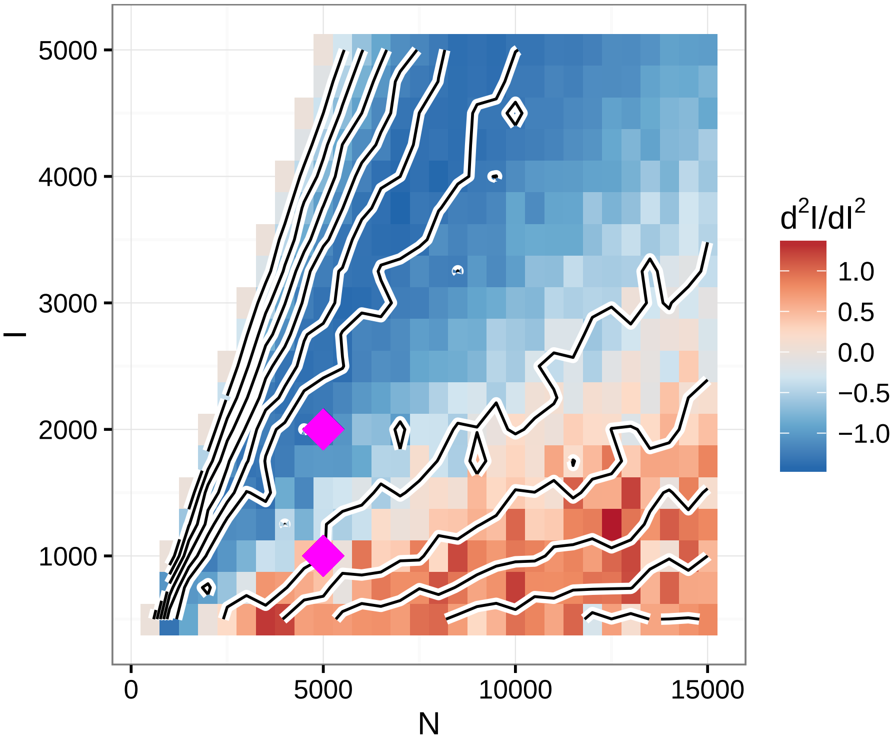

Figure 2.1: First and second derivatives of epidemic growth curves at time of sampling for various values of *I* and *N*. Networks were simulated under the BA model with $\alpha = 1.0$, $m = 2$, and *N* varied along the values shown on the *x*-axis. Transmission trees were sampled at the time when *I* 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 kernel-assisted 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 kernel-ABC. The exception to this was the value $m = 1$, which produces networks without cycles whose associated trees were more easily distinguished. However, all the analyses presented here did not take the absolute size of the transmission trees into account, as the branch lengths were rescaled by their mean. Because higher *m* values imply higher edge density, an epidemic should spread more quickly for higher *m* than lower *m* with the same per-edge transmission probability. Hence, considering the absolute height of the trees may improve our method's ability to reconstruct *m*.

In addition to the tree height, many summary statistics have been developed to capture particular details of tree shape. Two of these, Sackin's index and the ratio of internal to terminal branch lengths, were correlated with every BA parameter. Classifiers based on Sackin's index and the nLTT similarity measure performed well in some cases, though poorly in others. ABC is often applied using a vector of summary statistics [126, 141], 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 [142]. Hence, an avenue for future improvement of our method may be the inclusion of additional summary statistics to supplement the tree kernel. In addition, all four parameters were more accurately classified when the number of tips in the transmission trees was larger, underscoring the importance of adequate sampling for accurate phylodynamic inference.

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

As noted by Lintusaari et al. [143], uniform priors on model parameters may translate to highly informative priors on quantities of interest. We observed a non-linear relationship between the preferential attachment power α and the power law exponent γ (fig. A.25). Therefore, placing a uniform prior on α between 0 and 2 is equivalent to placing an informative prior that γ is close to 2. Therefore, if we were primarily interested in γ rather than α, a more sensible choice of prior might have a shape informed by fig. A.25 and be bounded above by approximately $\alpha = 1.5$. This would uniformly bound γ in the region $2 \leq \gamma \leq 4$ commonly reported in the network literature [23–25, 96]. We note however that Jones and Handcock [144] estimated γ values greater than four for some datasets, in one case as high as 17, indicating that a wider range of permitted γ 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 α were the most accurate, although the variance of the estimates was high and the confidence intervals were wide for $\alpha < 1$ (**????** and figs. A.23 and A.24). The estimates of N and I were both biased, and the estimates of m were largely uninformative. Despite these issues, a major result of our 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.

2.1.3 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 α differed substantially for three 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 α 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.

~~With respect to the preferential attachment power α, the six datasets analysed fell into two categories~~ (**??**). ~~First, we estimated a preferential attachment power close to 1, indicating linear preferential attachment, for the BC data and the outbreaks studied by~~ Niculescu et al. [145] and Wang et al. [30]. ~~These values were robust to specifying different priors for m. All three datasets were sampled from populations in which we would expect a high degree of epidemiological relatedness:~~ Niculescu et al. [145] ~~studied a recent outbreak among Romanian IDU,~~ Wang et al. ~~sampled acutely infected MSM in Beijing, China, and the BC data constituted a phylogenetic IDU cluster. These are all contexts in which we would expect some of the assumptions of the BA model, such as a connected network, relatively high mean degree, and preferential attachment dynamics, to hold.~~

~~The remaining three datasets~~ (Cuevas et al. [146], Novitsky et al. [147], and Li et al. [148]) ~~had estimated values of α below 0.5 when $m = 1$ was included in the prior, but these were not robust to changing the prior to exclude $m = 1$. For the~~ Cuevas et al. ~~data, model misspecification is likely partially responsible. While the authors found that a large proportion of the samples were epidemiologically linked, these were mainly in small local clusters rather than the single large component postulated by the BA model. In addition, the mixed risk groups in the dataset would be unlikely to significantly interact, further weakening any global preferential attachment dynamics. The dataset studied by~~ Novitsky et al. [147] ~~originated from a densely sampled population where the predominant risk factor was believed to be heterosexual exposure. Although the MAP estimate of α was almost unchanged when the value $m = 1$ was excluded from the prior, the confidence interval shrank substantially.~~

For all datasets we examined, the posterior mean estimates for α were sub-linear, ranging from 0.27 to 0.73. 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. They found α values ranging from 0.26 to 0.62, depending on the gender and time period considered. Our estimates of α for the Niculescu et al. [145] was above this range under both priors, as were the estimates for the Wang et al. [30] 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 Novitsky et al. [147] 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 [149], the authors noted that their sample was gender-biased, being composed of approximately 75% women. Our estimate of α for these data was 0.55 or 0.53, depending on the prior on m; de Blasio, Svensson, and Liljeros 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 (**??** and fig. A.29). This finding is in line with Dombrowski et al. [150], who reanalyzed a network of IDUs in Brooklyn, USA, collected between 1991 and 1993 [151]. They found that the 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.

One somewhat surprising result was the difference between parameter estimates for the Li et al. [148] and Wang et al. [30] datasets. Both groups studied cohorts of acutely infected MSM in major Chinese cities (Shanghai and Beijing respectively); yet, the Li et al. data was estimated to have a lower preferential attachment power and larger infected population than the Wang et al. [30] data…

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 γ of the real data sets we evaluated. However, the posterior means for α 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. A.26, 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 γ 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 γ 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.

Rothenberg and Muth [152] estimated the power law exponents for 15 major network studies of sexual and injection drug use networks carried out between 1981 and 2000. The estimated γ values for "all contacts" (that is, including both interviewed and non-interviewed individuals) were between 2.08 and 2.87 for all but one of the networks…

The estimates of the prevalence I were largely robust to the choice of prior on m (**??** and fig. A.29). Niculescu et al. [145] reported that 494 new HIV infections were diagnosed among the studied population (IDUs in Bucharest) during the study period. Our estimate of the prevalence was higher (X), although the

95% highest posterior density (HPD) interval did contain the value 494. In addition, it is not unreasonable that a substantial fraction of the HIV-positive individuals in an IDU population could be undiagnosed, given that these populations are often marginalized.

The individuals sampled by Wang et al. [30] were recruited by following a prospective cohort of 2000 MSM in Beijing, China. Of the 2000, 179 new infections were identified during the study period. This number is much lower than the estimated prevalence of X obtained with *netabc*, however the authors did not claim to have sampled the entire sexual network. In a nationwide survey, Wu et al. [153] estimated that there were 24,198 MSM living in Beijing, of whom 5.7%, or 1379, were HIV-infected, which is within the 95% confidence interval of our estimate of I.

Li et al. [148] studied an MSM population in Shanghai, China. Despite the apparent similarity to the context of the Wang et al. [30] study, the estimated HIV prevalence of the Li et al. data was higher (X). The aforementioned survey [153] estimated that there were 14,511 MSM living in Shanghai, with an HIV prevalence of 6.8% or approximately 987 individuals. This was within the 95% confidence interval of the estimate obtained with *netabc*. It is worth noting that Li et al. claimed that there were 80,000 MSM living in Shanghai, which is a significantly different estimate than that obtained by Wu et al. [153].

Cuevas et al. [146] 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 X we obtained with ABC, even without considering the likely possibility that there were more infected than sampled individuals. Again, this error may in part be due to model misspecification, which highlights the importance of selecting a model appropriate to the dataset at hand.

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 γ were within the range of values reported in the literature (**??**).

Chapter 3

Conclusion

Due to the rapid advancement of nucleotide sequencing technology, viral sequence data data have become increasingly feasible to collect on a population level. Through phylodynamic methods, these data offer a window into epidemiological processes which 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 [40]. 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 [48, 154], 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 which 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, 90]. 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 [23–25]. Of the four parameters we considered, the preferential attachment power α was the most readily estimable. Estimating α 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 [91] 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] Frank Harary and Edgar M Palmer. *Graphical Enumeration*. Elsevier, 2014.

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

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

[16] 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.

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

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

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

[20] 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.

[21] 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.

[22] 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.

[23] 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.

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

[25] 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.

[26] 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.

[27] 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.

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

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

[30] 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.

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

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

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

[34] 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.

[35] 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.

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

[37] 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.

[38] 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).

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

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

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

[42] 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.

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

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

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

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

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

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

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

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

[51] 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.

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

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

[54] 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.

[55] 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.

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

[57] 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.

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

[59] 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.

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

[61] 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.

[62] 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.

[63] 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.

[64] 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.

[65] 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.

[66] 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.

[67] 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.

[68] 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.

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

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

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

[72] 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.

[73] 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.

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

[75] 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.

[76] 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.

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

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

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

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

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

[82] 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.

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

[84] 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.

[85] 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.

[86] 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).

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

[88] Duncan J Watts and Steven H Strogatz. "Collective dynamics of 'small-world' networks". In: *Nature* 393.6684 (1998), pp. 440–442.

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

[90] 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.

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

[92] 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.

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

[94] 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.

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

[96] 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.

[97] 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.

[98] 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.

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

[100] 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.

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

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

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

[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] 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.

[136] 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.

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

[138] 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.

[139] 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.

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

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

[142] 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.

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

[144] 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.

[145] 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.

[146] 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.

[147] 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.

[148] 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.

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

[150] 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.

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

[152] 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.

[153] 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.

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

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

[156] 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

A *mathematical model* is a formal description of a hypothesized relationship between some observed data, x and outcomes y. A *parametric* model defines a family of possible relationships between data and outcomes, parameterized by one or more numeric parameters θ. 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, θ. 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 $x = \{x_1, \dots, x_n\}$ and $y = \{y_1, \dots, y_n\}$ are vectors of real numbers. The linear model postulates that the outcomes are linearly related to the data, modulo some noise introduced by measurement error, environmental fluctuations, and other external factors. Formally, $y_i = \beta x_i + \varepsilon_i$, where β is the slope of the linear relationship, and ε_i is the error associated with measurement i. We can make this model a statistical one by hypothesizing a distribution for the error terms ε_i; most commonly, it is assumed that they are normally distributed with variance σ. In mathematical terms, $y_i \sim \beta x_i + \mathscr{N}(0, \sigma^2)$, where "$\sim$" means "is distributed as". We can see from this formulation that the model is parametric, with parameters θ = (β, σ). Moreover, we can write down the probability density π of observing outcome y_i given the parameters,

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

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

For a general model, the probability density of y given the parameters θ is also known as the *likelihood*, written $\mathscr{L}$, of θ. That is, $\mathscr{L}(\theta \mid y) = f(y \mid \theta)$ for the model's probability density function (pdf)

f. The higher the value of the likelihood, the more likely the observations y are under the model. Thus, the likelihood provides a natural criterion for fitting the model parameters: we want to pick θ such that the probability density of our observed outcomes y is as high as possible. The parameters that optimize the likelihood are known as the *ML* estimates, denoted $\hat{\theta}$. That is,

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

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

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

In principle, $f(y \mid \theta)\pi(\theta)$ can be optimized numerically just like $\mathscr{L}(\theta \mid y)$, which would also optimize the posterior distribution. The resulting optimal parameters are called the maximum *a posteriori* (MAP) estimates. However, from a Bayesian perspective, θ is not a fixed quantity to be estimated, but rather a random variable with an associated distribution (the posterior). Therefore, the MAP estimate by itself is of limited value without associated statistics about the posterior distribution, such as the mean or credible intervals. Unfortunately, to calculate such statistics, it is necessary to evaluate the normalizing constant in the denominator of eq. (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 ~~whose states are indexed by possible model parameters. The transition probability from state θ_1 to state θ_2 is taken to be~~ on the space of possible model parameters. The transition density from parameters θ_1 to θ_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 θ. Therefore, if a long enough random walk is performed on the chain, the distribution of states visited will be a Monte Carlo approximation of $f(\theta \mid y)$, from which we can calculate statistics of interest. Actually performing this random walk is straightforward and can be accomplished via the Metropolis-Hastings algorithm [155, 156] (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

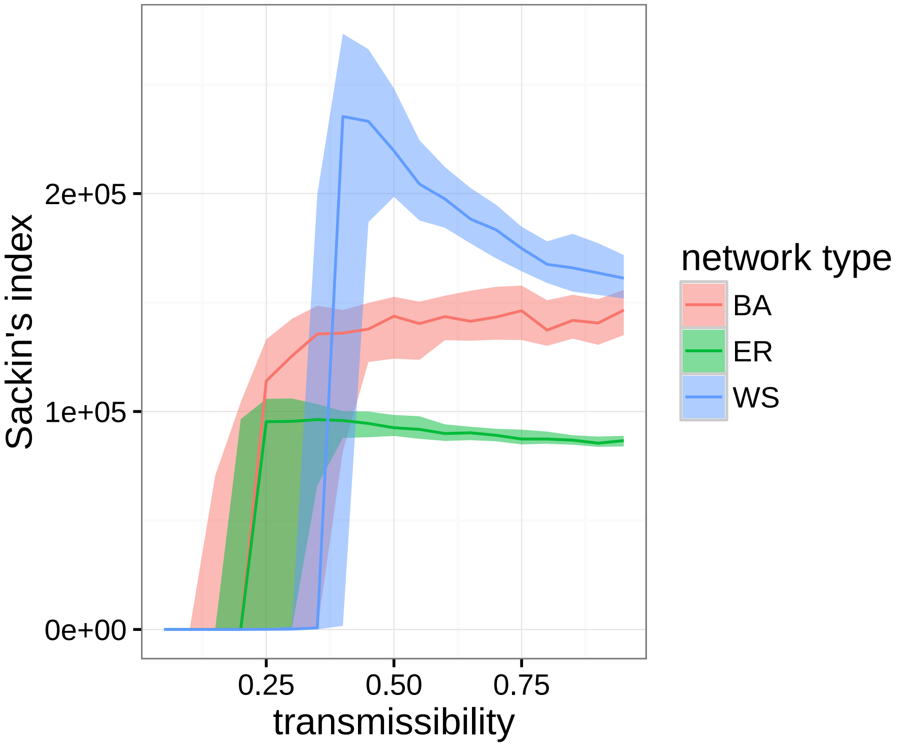

Figure A.1: Reproduction of Figure 1A from Leventhal *et al.* (2012) used to check the accuracy of our implementation of Gillespie simulation. Transmission trees were simulated over three types of network, with pathogen transmissibility varying from 0 to 1. Sackin's index was calculated for each simulated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

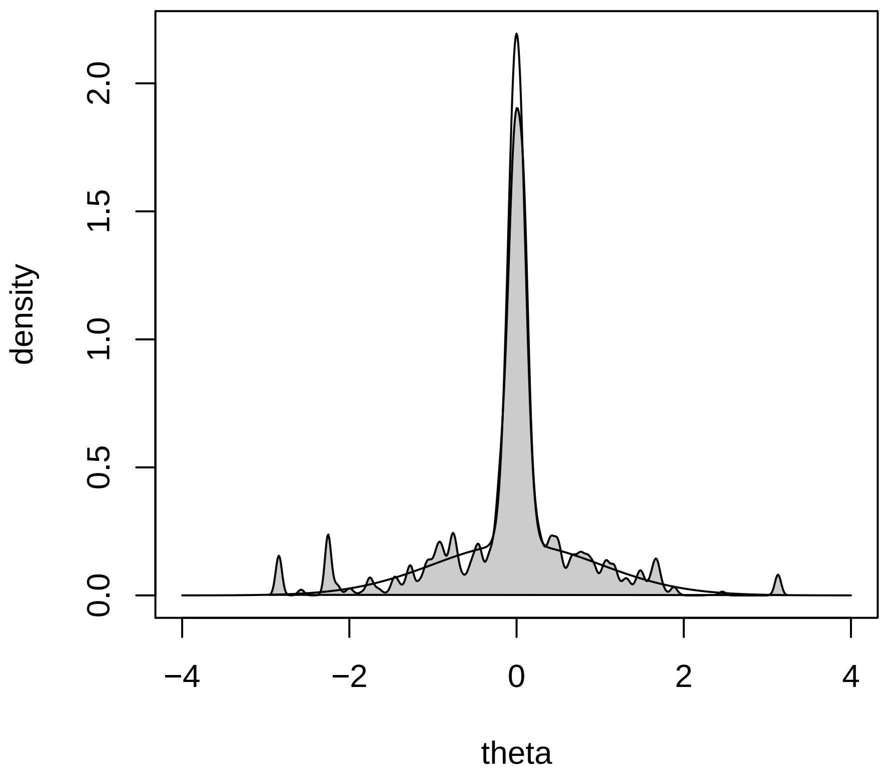

Figure A.2: Approximation of mixture of Gaussians used by Del Moral *et al.* (2012) and Sisson *et al.* (2009) to test SMC. Solid black line indicates true distribution. Grey shaded area shows ABC approximation obtained with our implementation of adaptive ABC-SMC, using 10000 particles with one simulated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

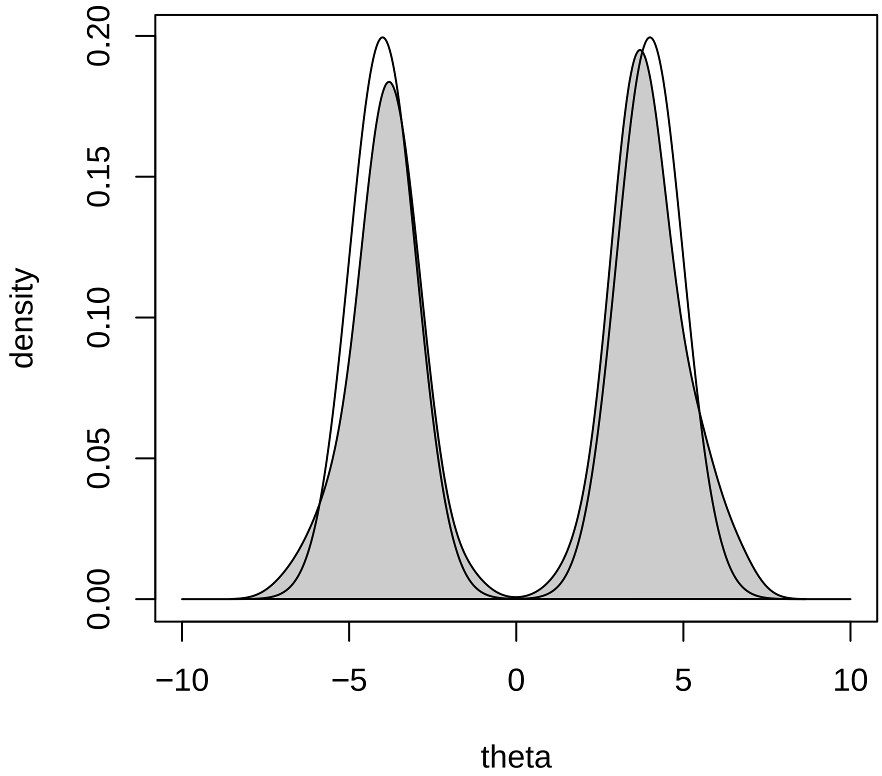

Figure A.3: Approximation of mixture of two Gaussians used to test convergence of SMC algorithm to a bimodal distribution. Solid black line indicates true distribution. Grey shaded area shows ABC-SMC approximation obtained with our implementation, using 10000 particles with one simulated data point per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

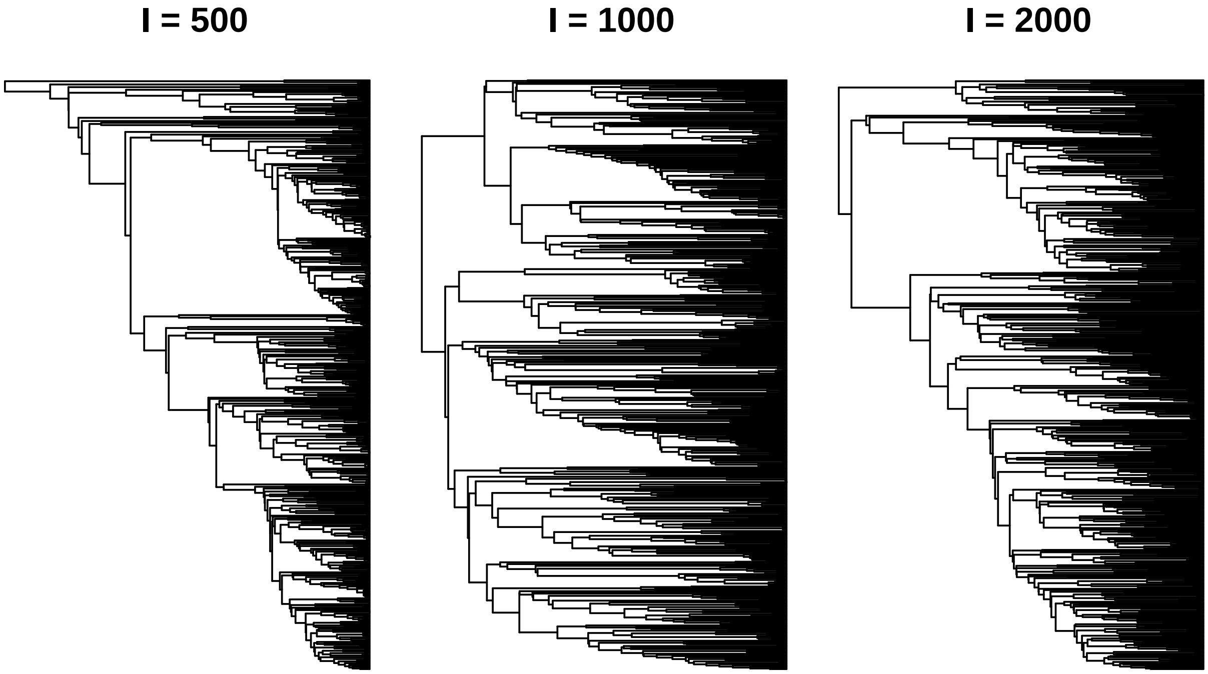

Figure A.4: Simulated transmission trees under three different values of BA parameter I. Epidemics were simulated on BA networks with parameters $\alpha = 1.0$, $m = 2$, and $N = 5000$. Epidemics were simulated until $I = 500$, 1000, or 2000 nodes were infected. Transmission trees were created by sampling 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.

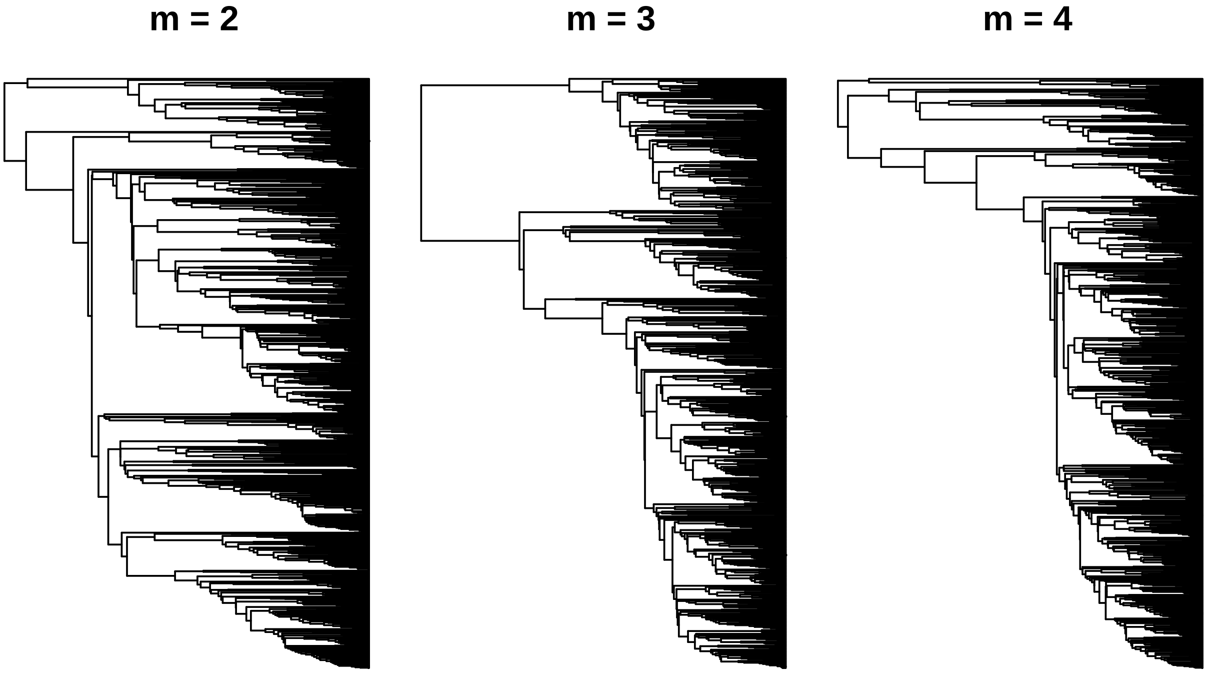

Figure A.5: Simulated transmission trees under three different values of BA parameter m. Epidemics were simulated on BA networks with parameters α = 1.0, $N = 5000$, and m = 2, 3, or 4. Epidemics were simulated until $I = 1000$ nodes were infected. Transmission trees were created by sampling 500 infected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

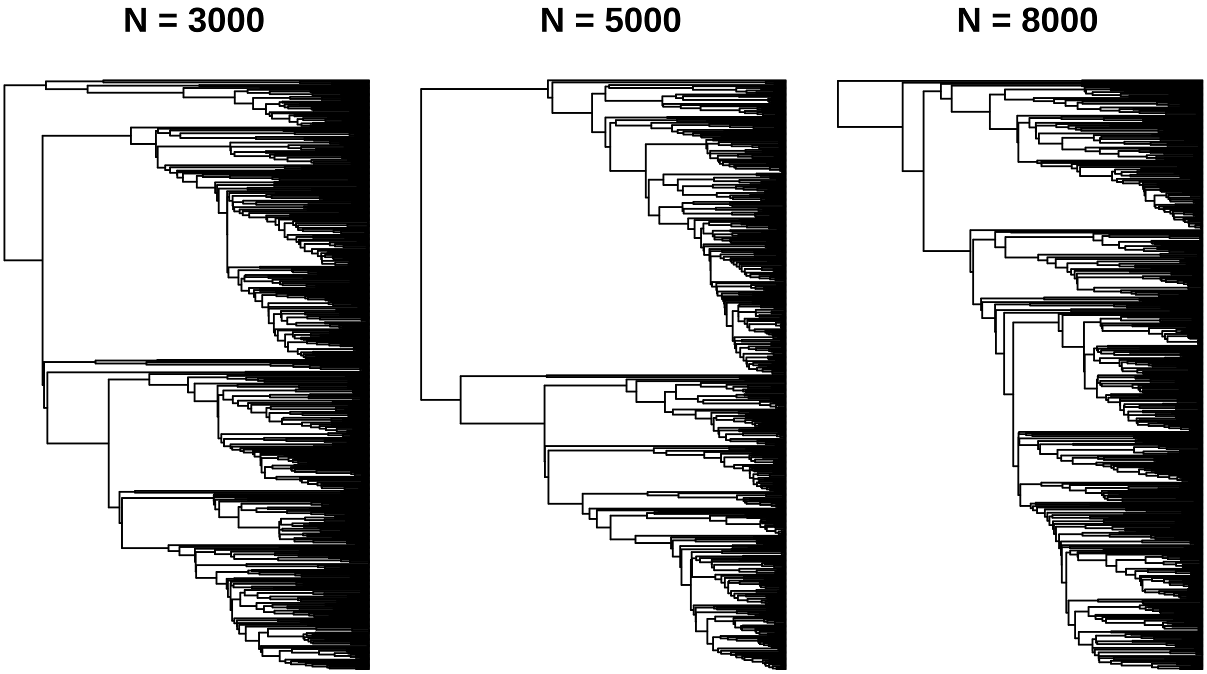

Figure A.6: Simulated transmission trees under three different values of BA parameter N. Epidemics were simulated on BA networks with parameters α = 1.0, $m = 2$, and N = 3000, 5000, or 8000. Epidemics were simulated until $I = 1000$ nodes were infected. Transmission trees were created by sampling 500 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.

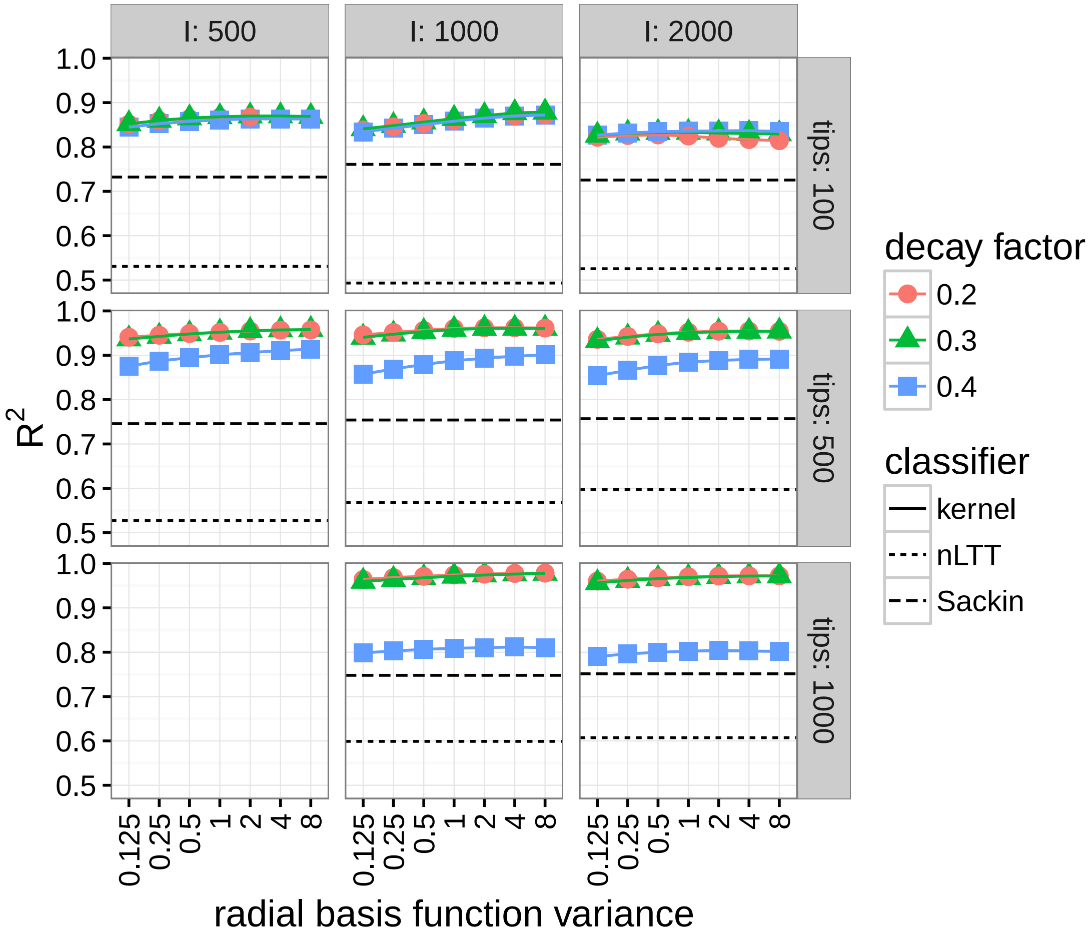

Figure A.7: Cross validation accuracy of classifiers for BA model parameter α for eight epidemic scenarios. Solid lines and points are R^2 of tree kernel kSVR under various kernel meta-parameters. Dashed and dotted lines are R^2 of linear regression against Sackin's index, and SVR using nLTT.

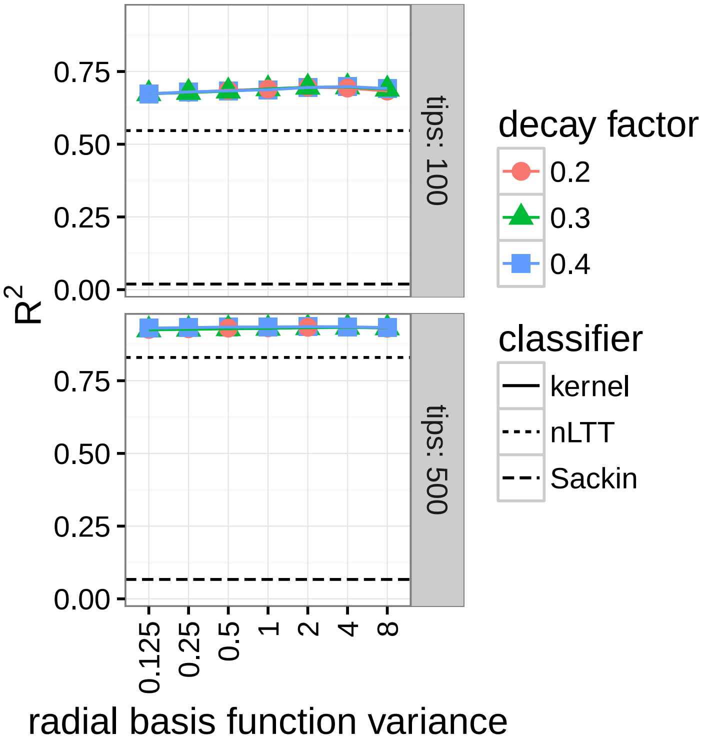

Figure A.8: Cross validation accuracy of classifiers for BA model parameter I for eight epidemic scenarios. Solid lines and points are R^2 of tree kernel kSVR under various kernel meta-parameters. Dashed and dotted lines are R^2 of linear regression against Sackin's index, and SVR using nLTT.

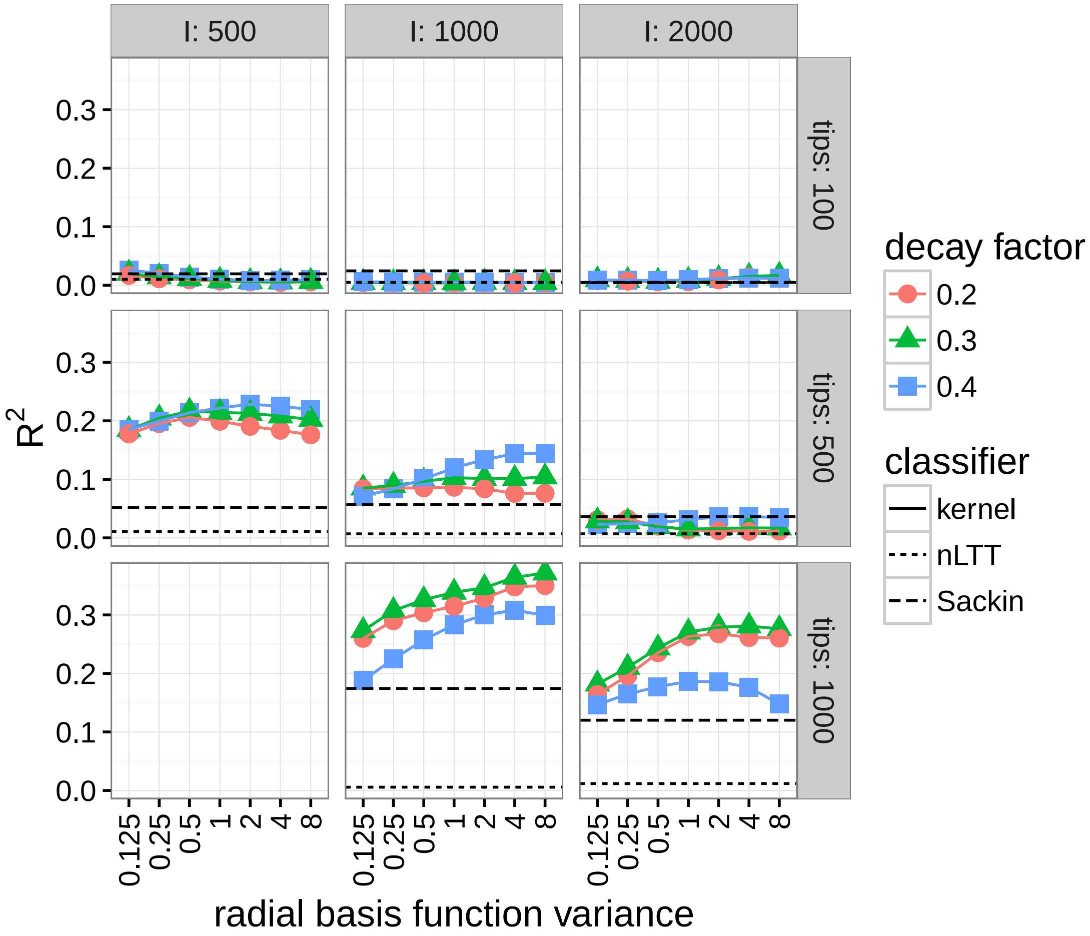

Figure A.9: Cross validation accuracy of classifiers for BA model parameter *m* for eight epidemic scenarios. Solid lines and points are R^2 of tree kernel kSVR under various kernel meta-parameters. Dashed and dotted lines are R^2 of linear regression against Sackin's index, and SVR using nLTT.

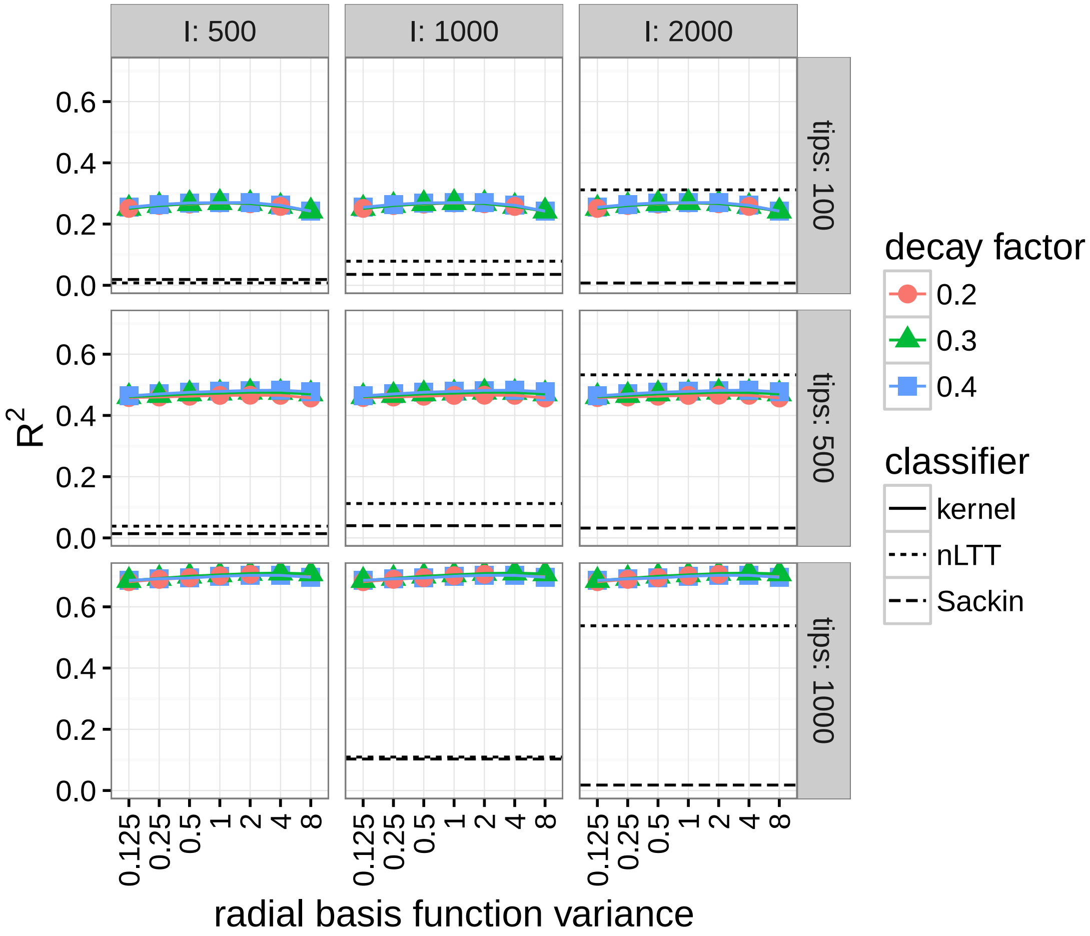

Figure A.10: Cross validation accuracy of classifiers for BA model parameter N for eight epidemic scenarios. Solid lines and points are R^2 of tree kernel kSVR under various kernel meta-parameters. Dashed and dotted lines are R^2 of linear regression against Sackin's index, and SVR using nLTT.

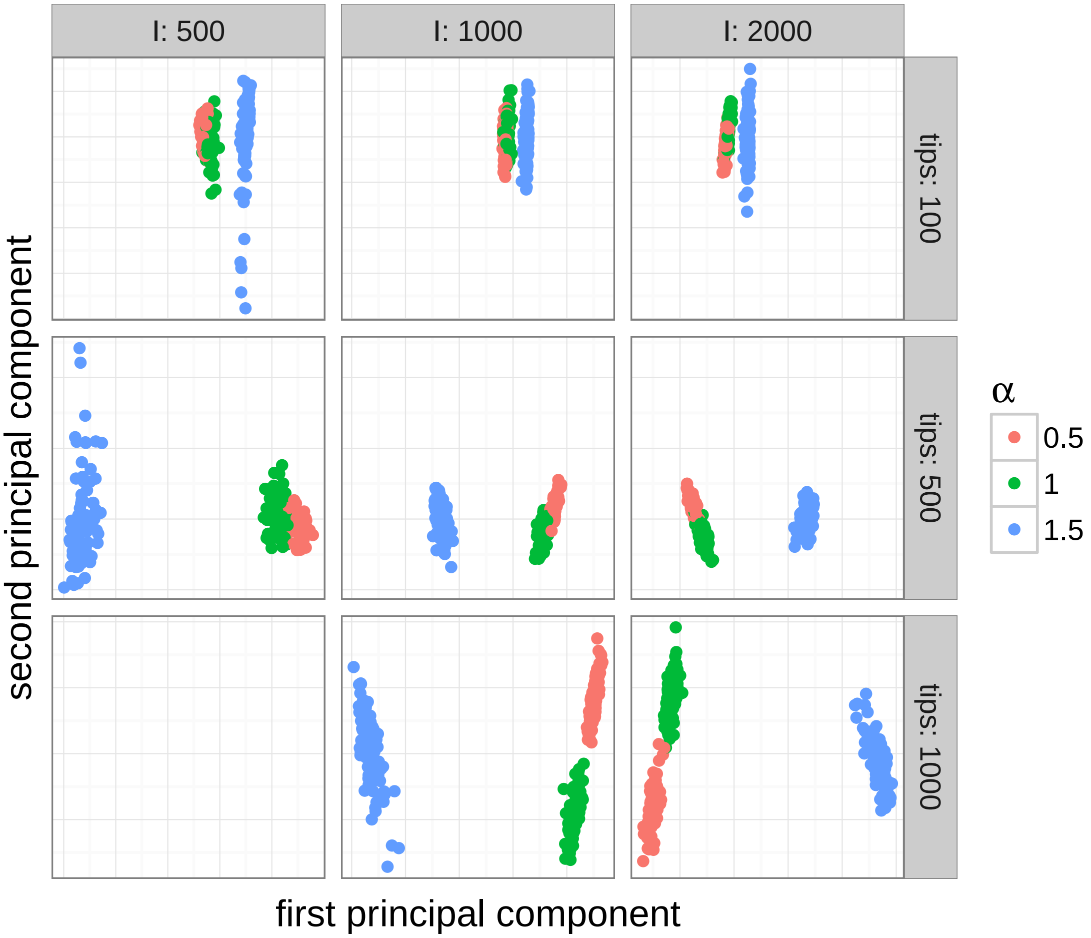

Figure A.11: Kernel principal components projection of trees simulated under three different values of BA parameter α, for eight epidemic scenarios.

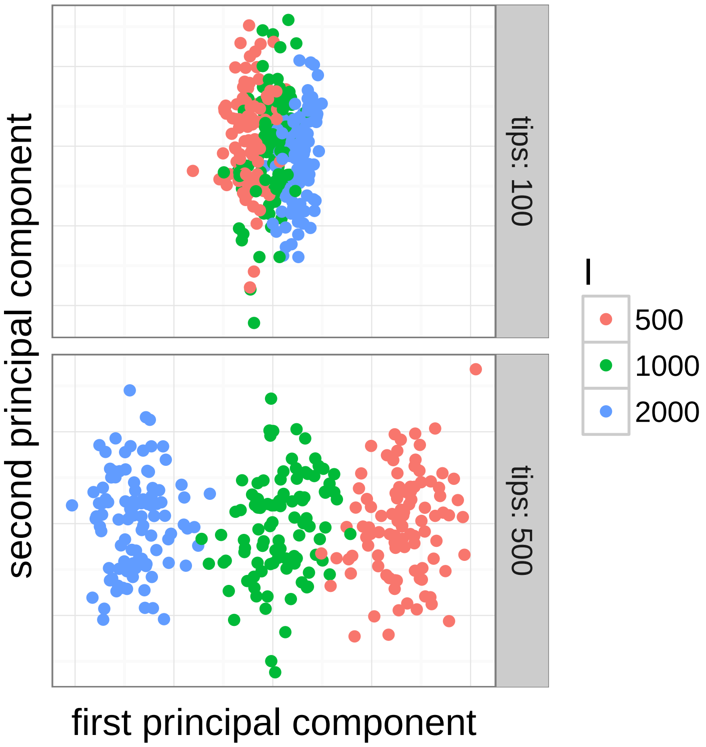

Figure A.12: Kernel principal components projection of trees simulated under three different values of BA parameter *I*, for eight epidemic scenarios.

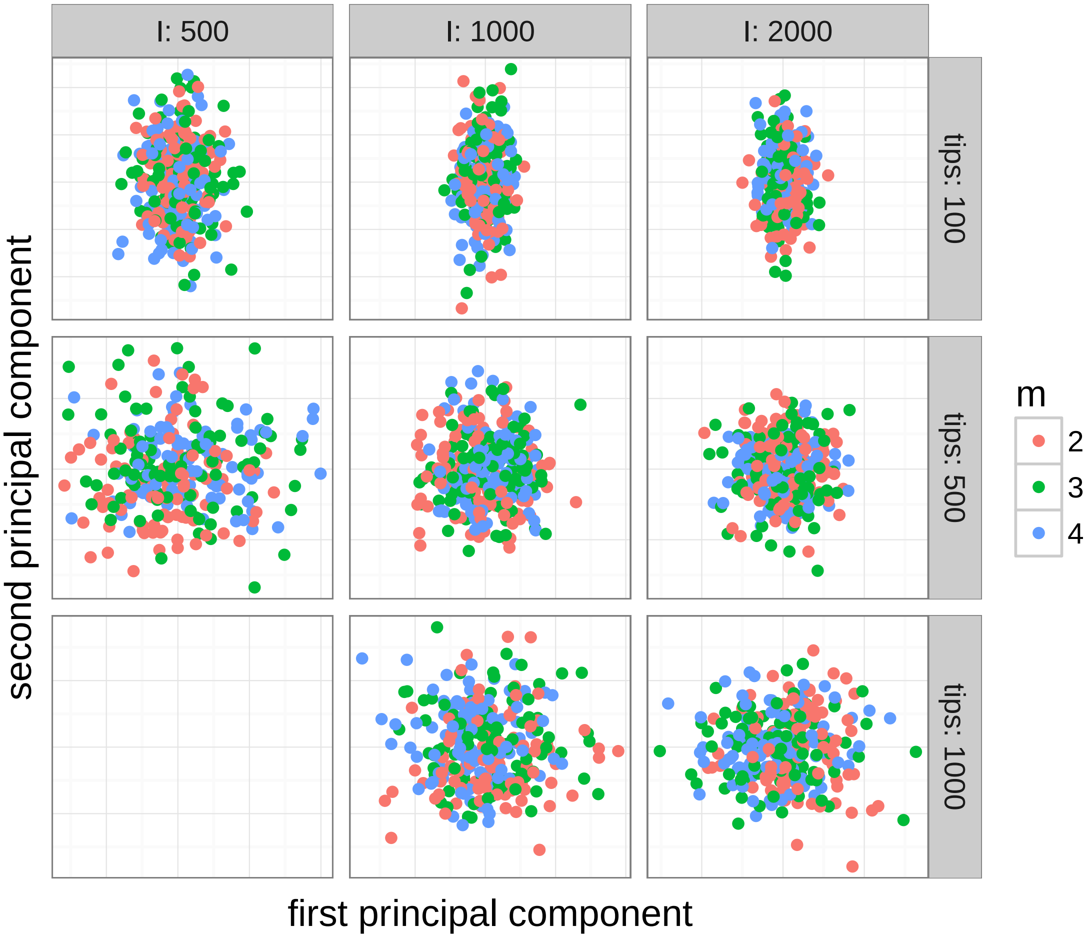

Figure A.13: Kernel principal components projection of trees simulated under three different values of BA parameter *m*, for eight epidemic scenarios.

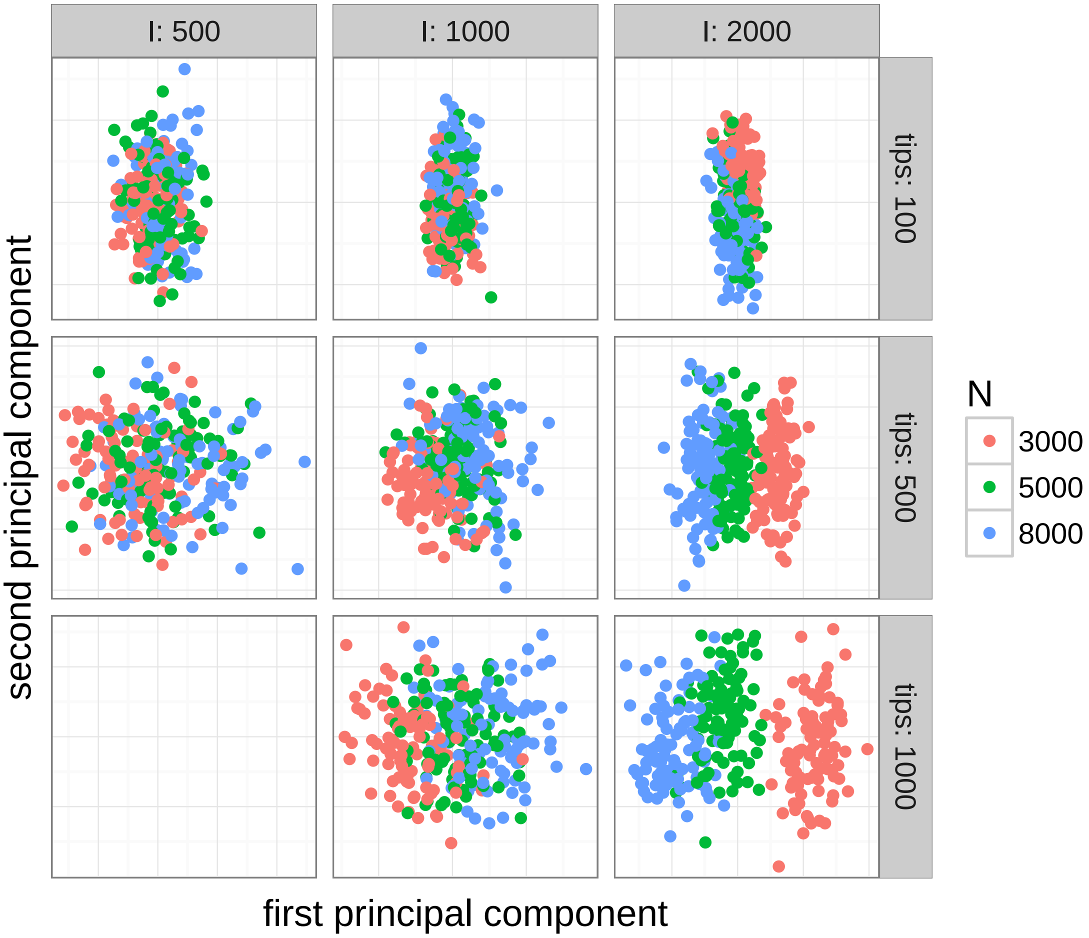

Figure A.14: Kernel principal components projection of trees simulated under three different values of BA parameter N, for eight epidemic scenarios.

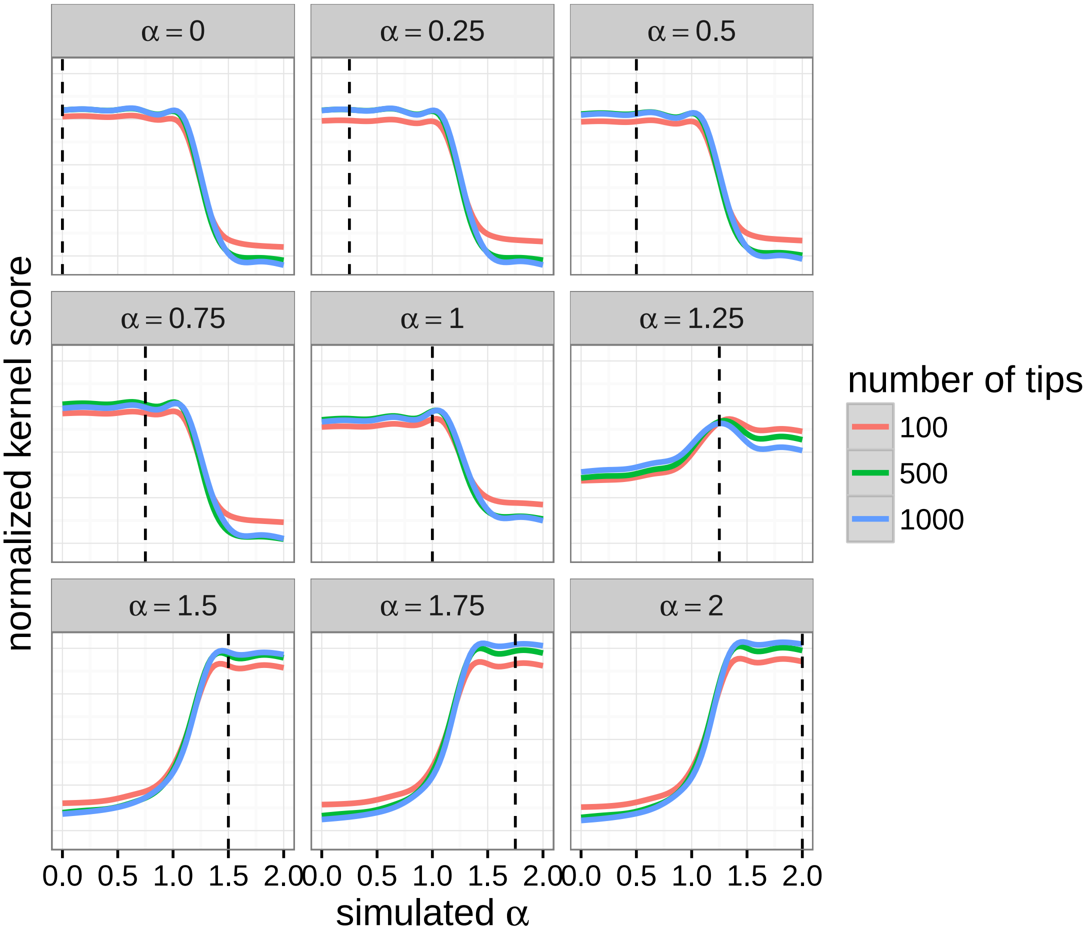

Figure A.15: Grid search kernel scores for testing trees simulated under various α values. The other BA parameters were fixed at $I = 1000$, $N = 5000$, and $m = 2$.

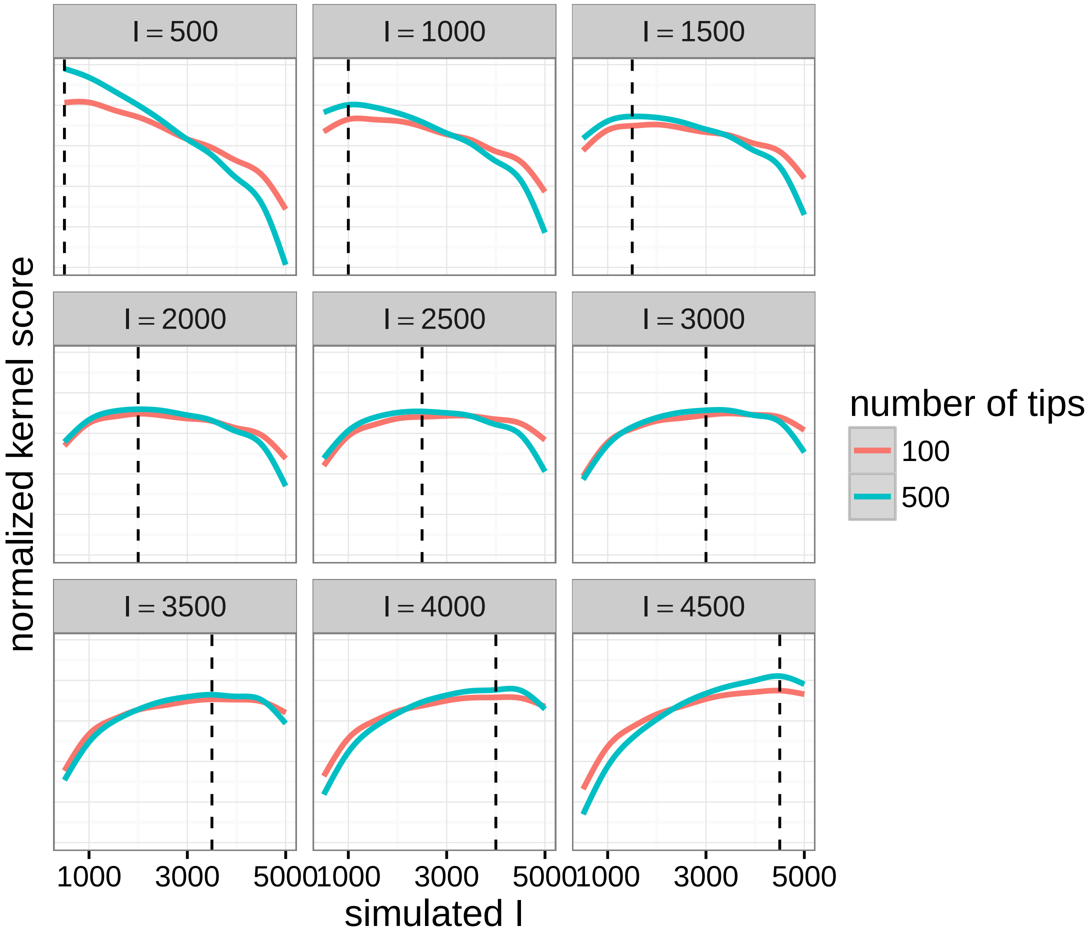

Figure A.16: Grid search kernel scores for testing trees simulated under various I values. The other BA parameters were fixed at $\alpha = 1.0$, $N = 5000$, and $m = 2$.

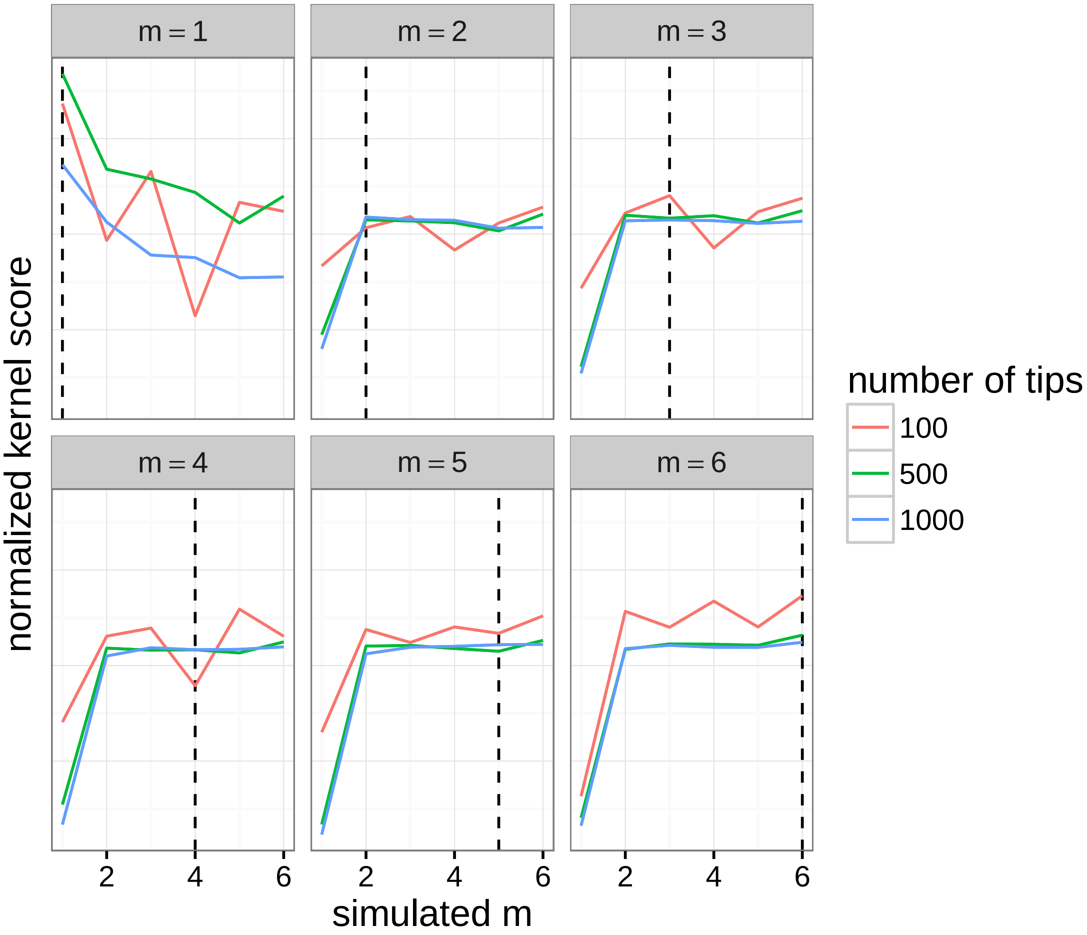

Figure A.17: Grid search kernel scores for testing trees simulated under various m values. The other BA parameters were fixed at $\alpha = 1.0$, $I = 1000$, and $N = 5000$.

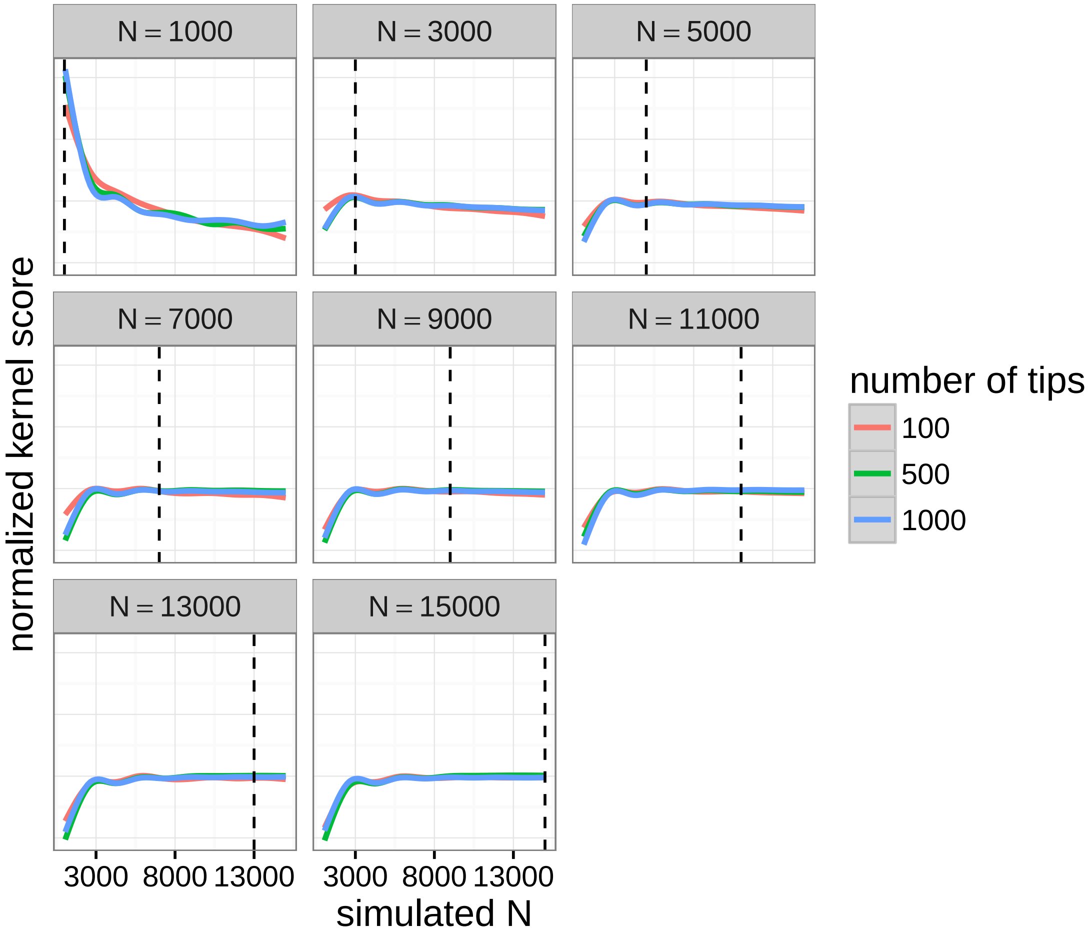

Figure A.18: Grid search kernel scores for testing trees simulated under various N values. The other BA parameters were fixed at $\alpha = 1.0$, $I = 1000$, and $m = 2$.

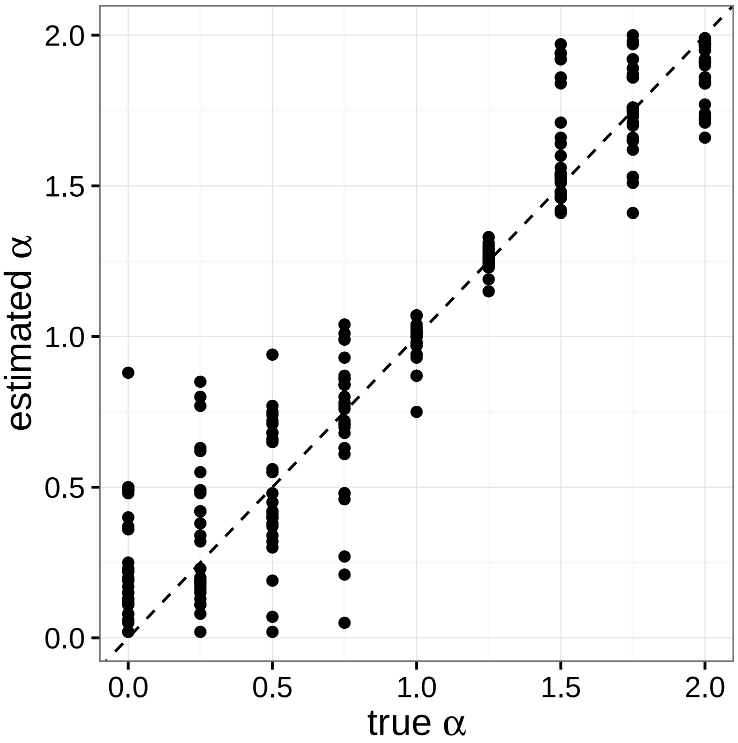

Figure A.19: Point estimates of preferential attachment power α of Barabási-Albert network model, obtained on simulated trees with kernel-score-based grid search. Test trees were simulated according to several values of α (x-axis) with other model parameters fixed at $m = 2$, $N = 5000$, and $I = 1000$. The test trees were compared to trees simulated along a narrowly spaced grid of α 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 α (y-axis).

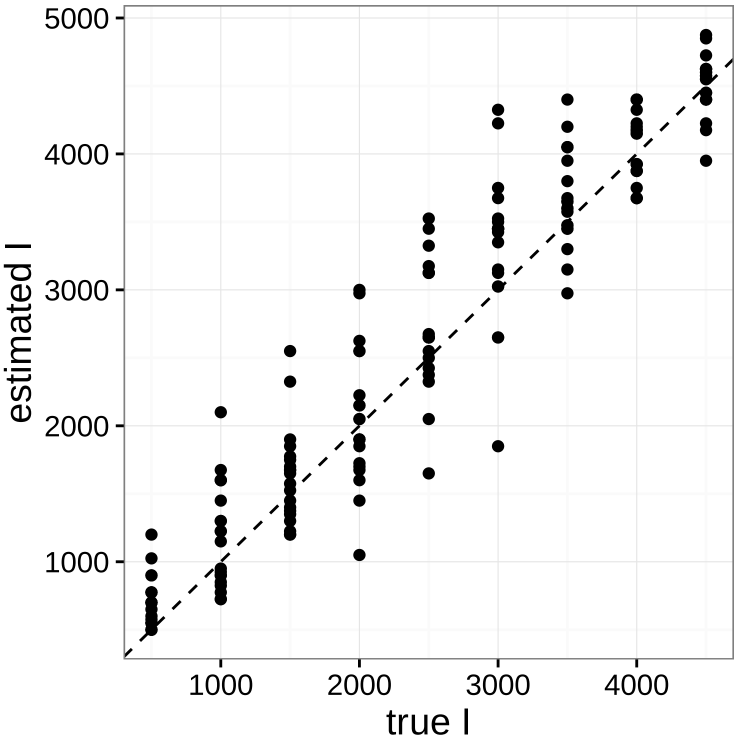

Figure A.20: Point estimates of prevalence at time of sampling I of Barabási-Albert network model, obtained on simulated trees with kernel-score-based grid search. Test trees were simulated according to several values of I (x-axis) with other model parameters fixed at $\alpha = 1$, $m = 2$, and $N = 5000$. 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).

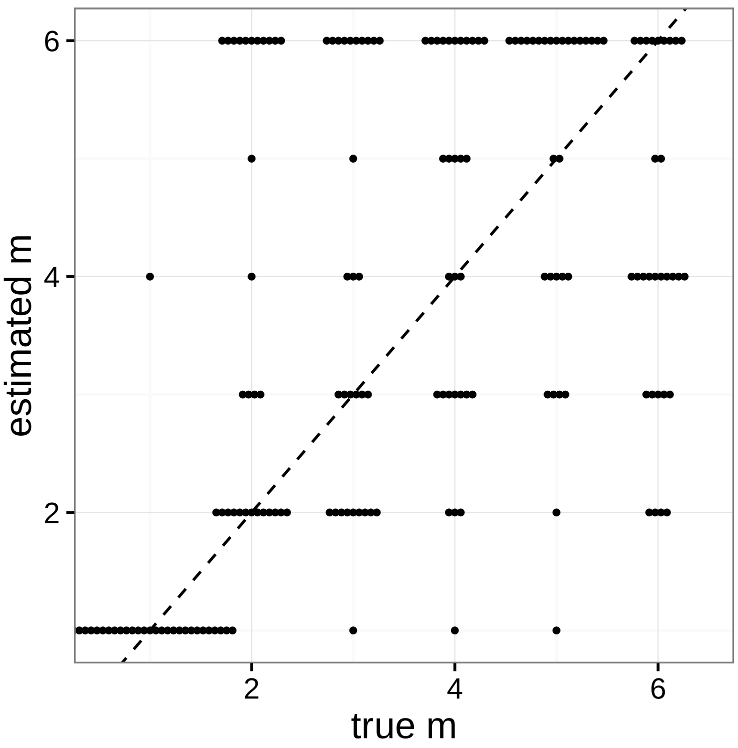

Figure A.21: Point estimates of number of edges per vertex *m* of Barabási-Albert network model, obtained on simulated trees with kernel-score-based grid search. Test trees were simulated according to several values of *m* (*x*-axis) with other model parameters fixed at $\alpha = 1$, $I = 1000$, and $N = 5000$. 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).

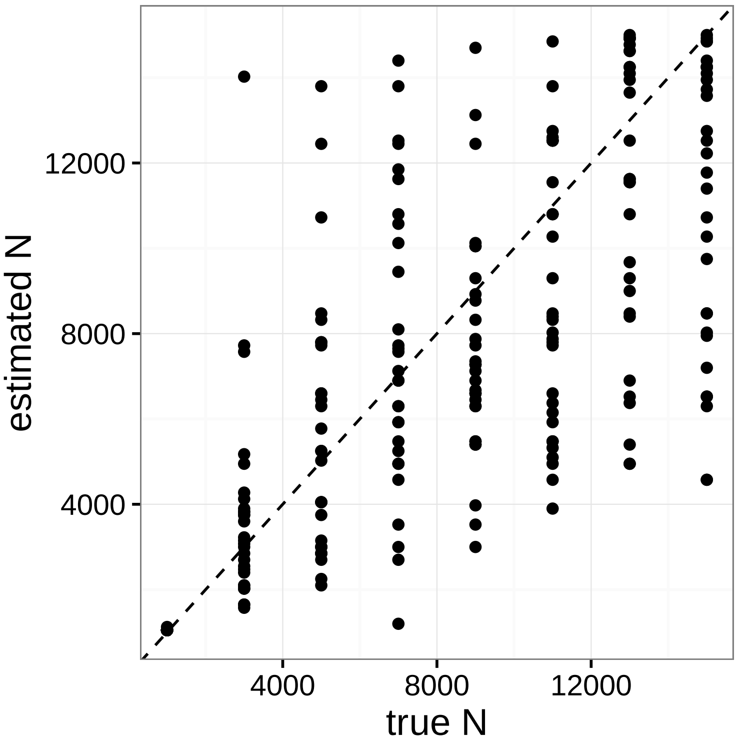

Figure A.22: Point estimates of number of edges per vertex *N* of Barabási-Albert network model, obtained on simulated trees with kernel-score-based grid search. Test trees were simulated according to several values of *N* (*x*-axis) with other model parameters fixed at $\alpha = 1$, $m = 2$, and $I = 1000$. 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).

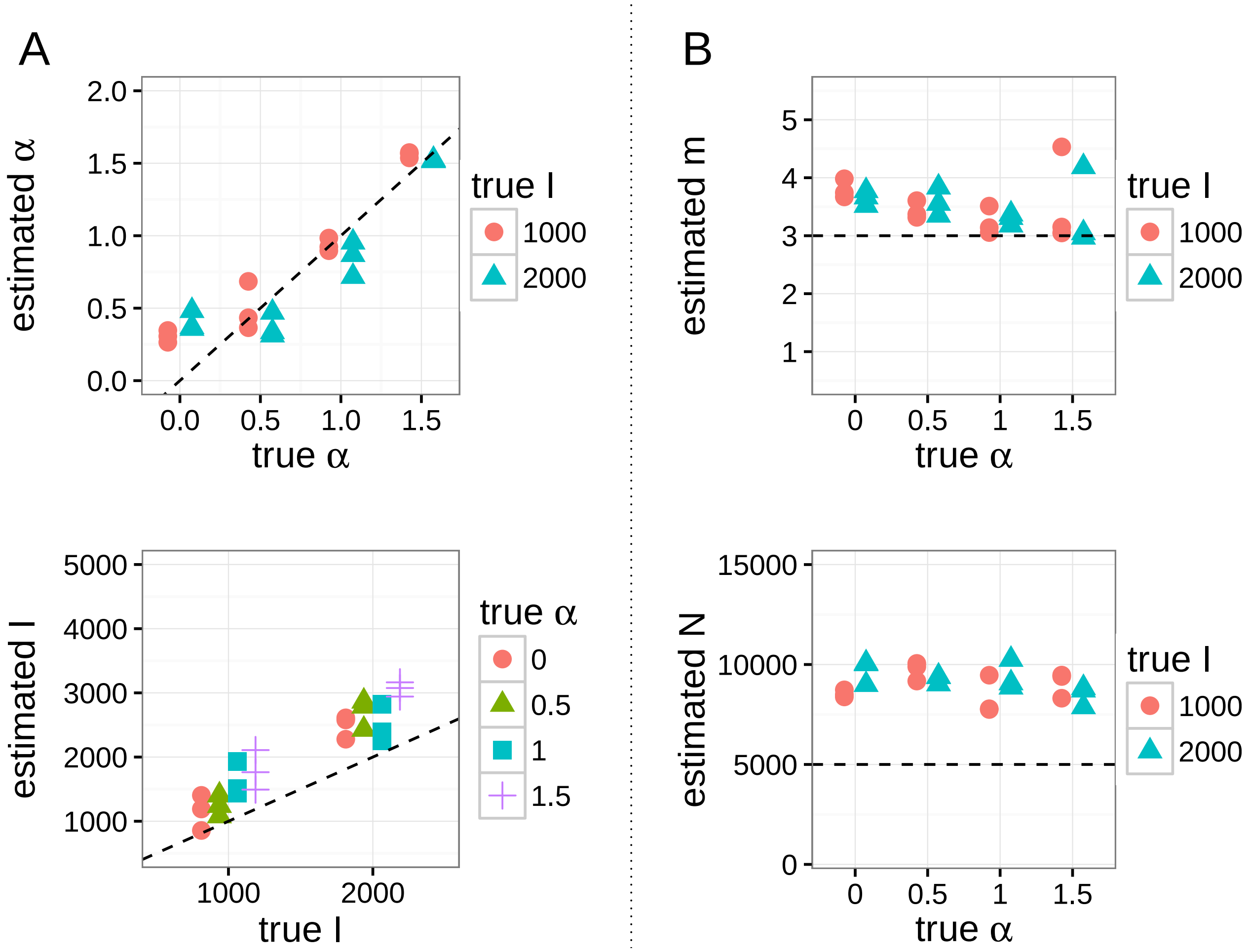

Figure A.23: Maximum *a posteriori* point estimates for BA model parameters obtained by running *netabc* on simulated data, for simulations with $m = 3$. Dashed lines indicate true values. (A) Estimates of α and I which were varied in these simulations against known values. (B) Estimates of m and N which were held fixed in these simulations at the values $m = 3$ and $N = 5000$.

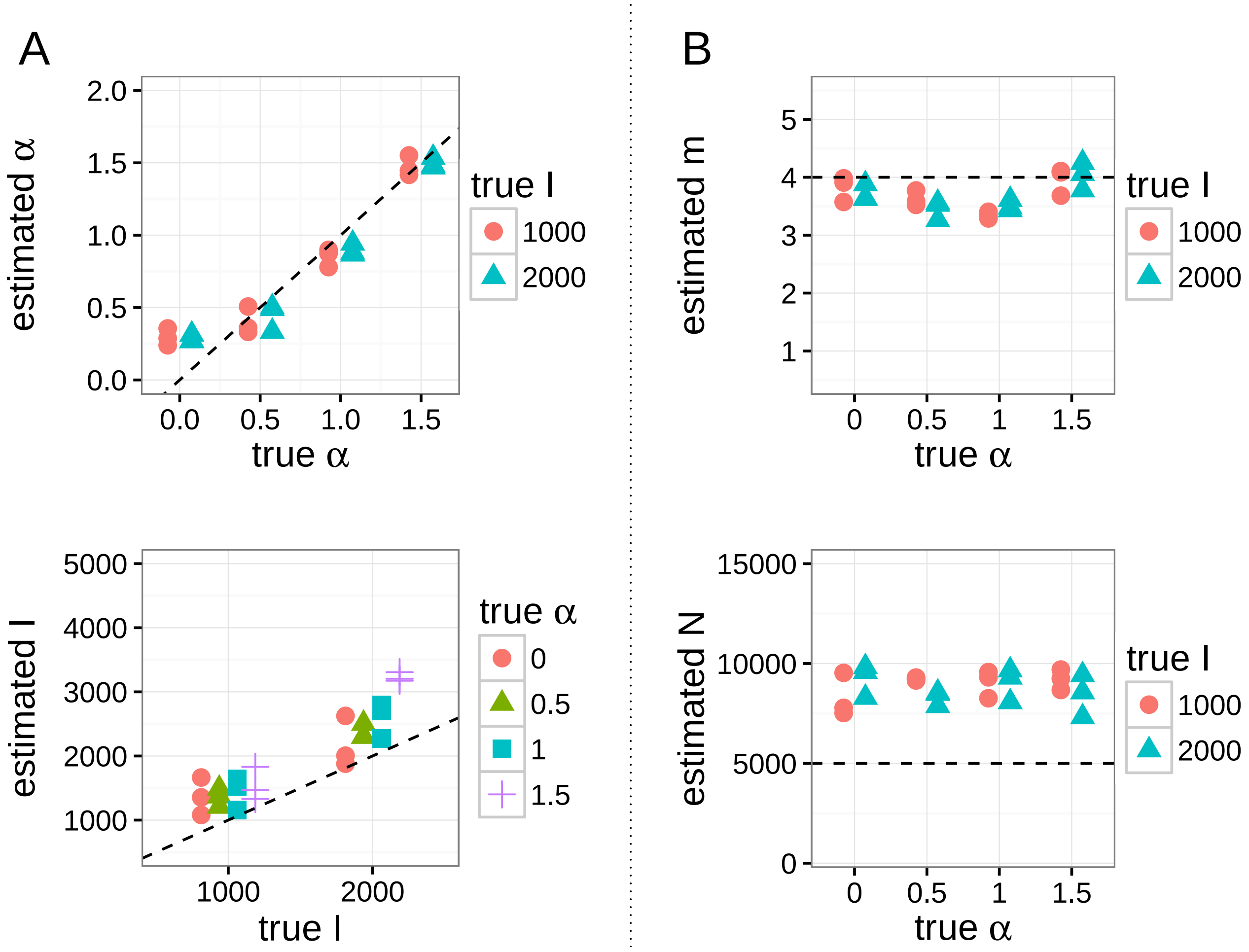

Figure A.24: Maximum *a posteriori* point estimates for BA model parameters obtained by running *netabc* on simulated data, for simulations with $m = 4$. Dashed lines indicate true values. (A) Estimates of α and I which were varied in these simulations against known values. (B) Estimates of m and N which were held fixed in these simulations at the values $m = 4$ and $N = 5000$.

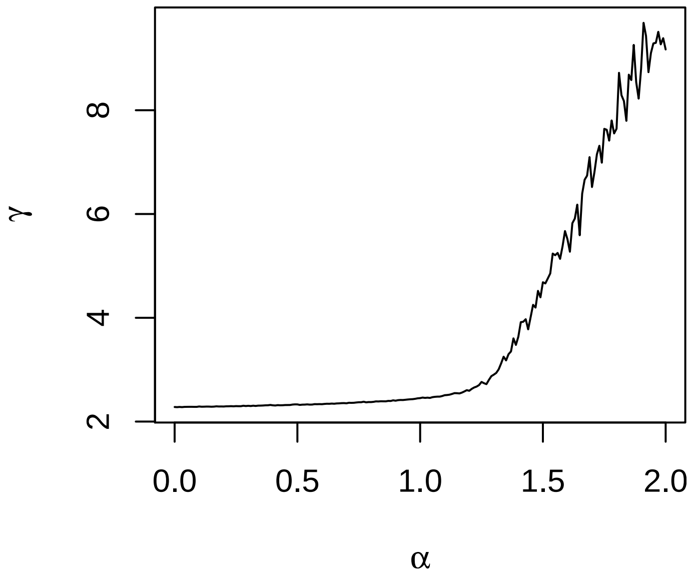

Figure A.25: Relationship between preferential attachment power parameter α and power law exponent γ for networks simulated under the BA network model with $N = 5000$ and $m = 2$.

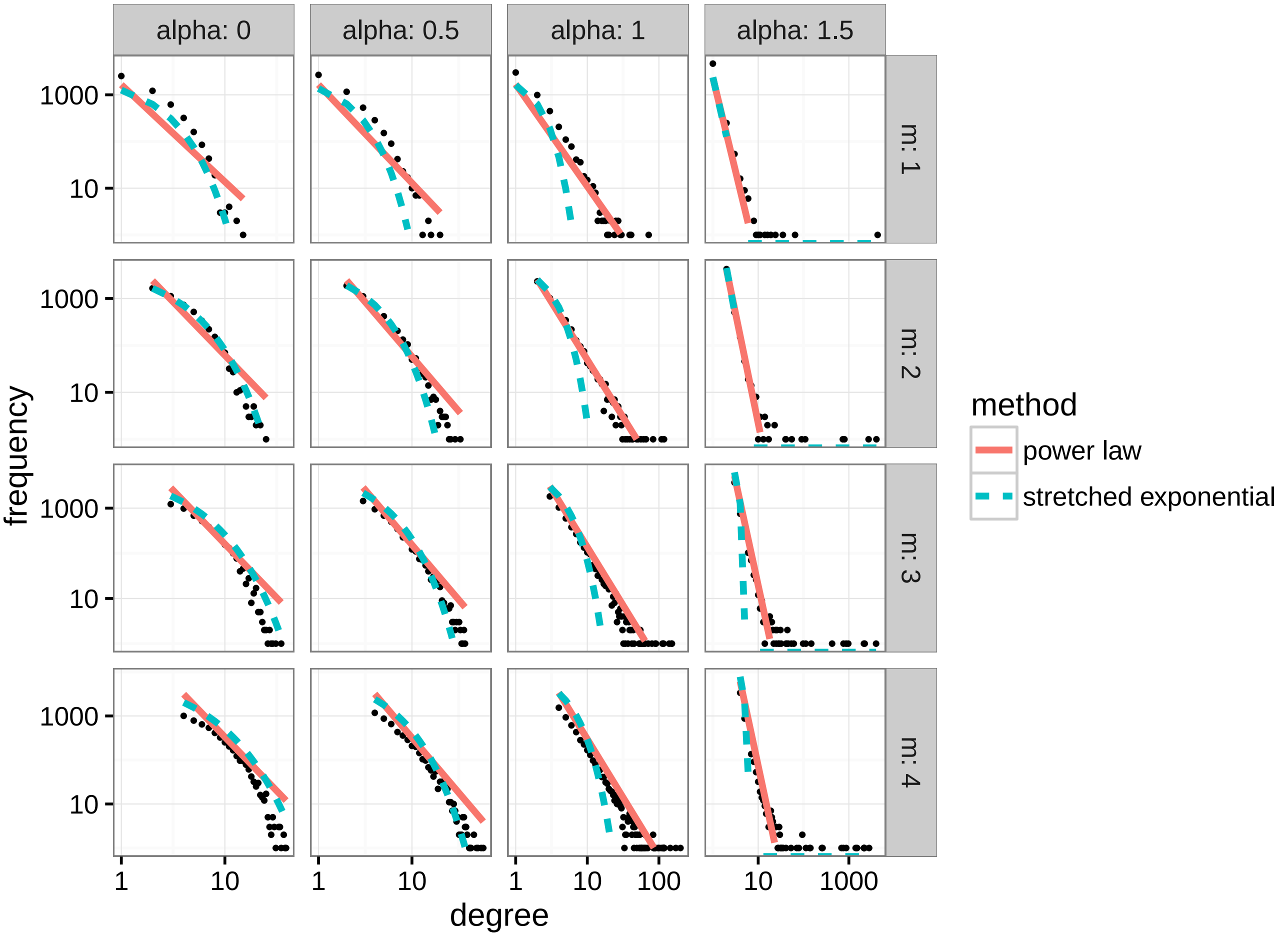

Figure A.26: Best fit power law and stretched exponential curves for degree distributions of simulated Barabási-Albert networks for several values of α and m.

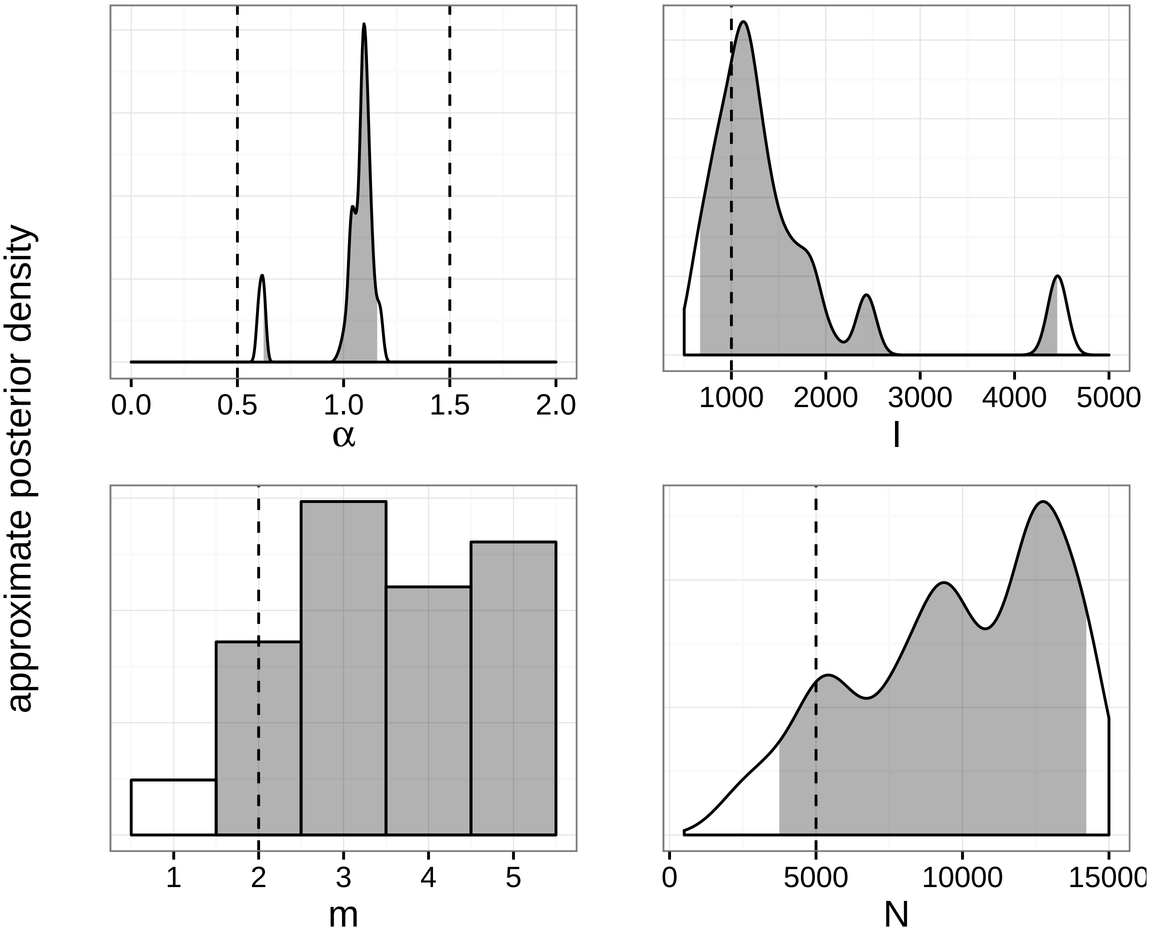

Figure A.27: Approximate marginal posterior distributions of Barabási-Albert model parameters obtained using kernel-assisted ABC for a network with heterogeneous node behaviour. Half of the nodes were attached with $\alpha = 0.5$, and the other half with $\alpha = 1.5$ (vertical dashed lines, top left). Other parameter values were $m = 2$, $I = 1000$, and $N = 5000$ (vertical dashed lines, other than top left). Shaded areas indicate 95% highest posterior density intervals.

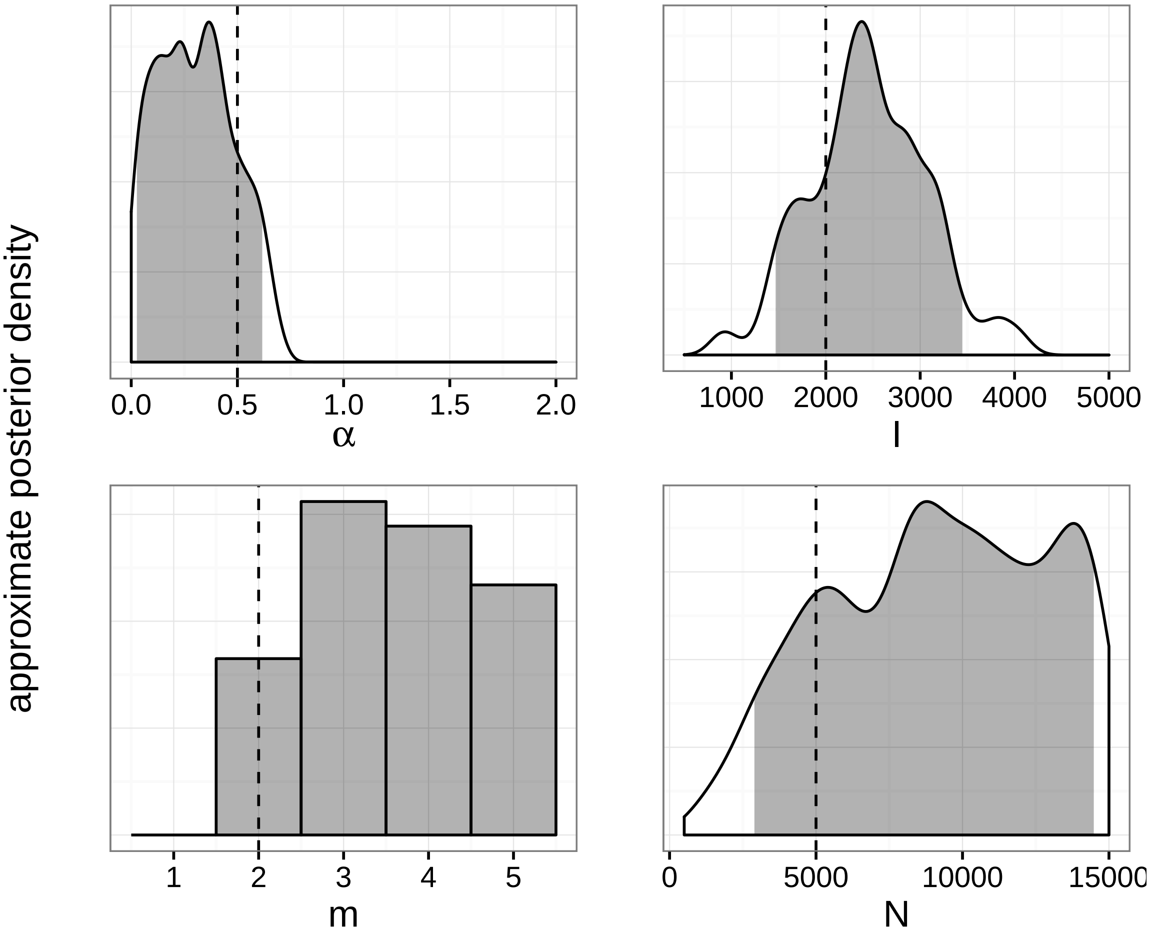

Figure A.28: Approximate marginal posterior distributions of Barabási-Albert model parameters obtained using kernel-assisted ABC for a network with peer-driven sampling. An epidemic was simulated in the usual fashion, but rather than being sampled at random, infected nodes were sampled with a probability two times higher if they had any sampled neighbours in the contact network. Vertical dashed lines indicate true parameter values, and shaded areas indicate 95% highest posterior density intervals.

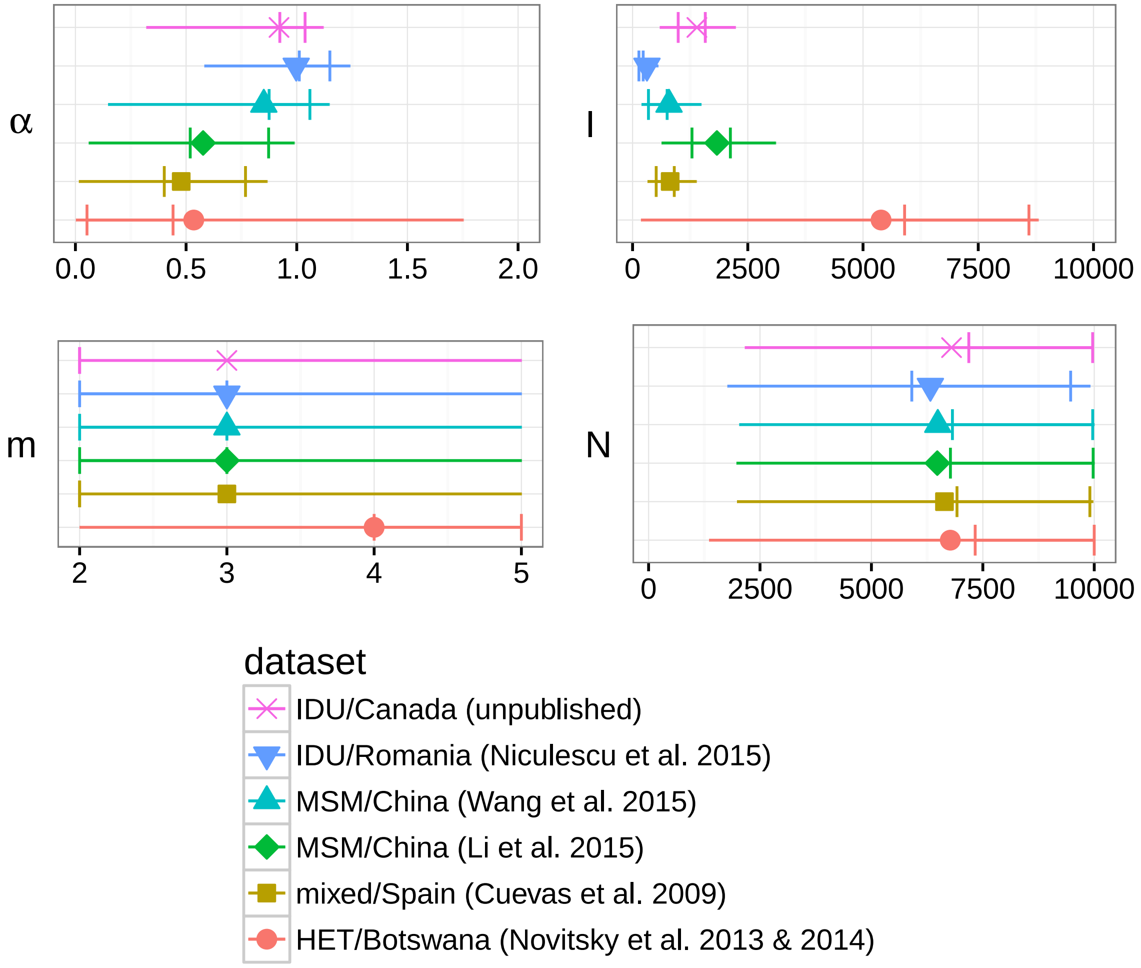

Figure A.29: Maximum *a posteriori* point estimates and 95% HPD intervals for parameters of the BA network model, fitted to five published HIV datasets with *netabc* using the prior $m \sim$ DiscreteUniform(2, 5). x-axes indicate regions of nonzero prior density. In particular, the prior on m was DiscreteUniform(2, 5).

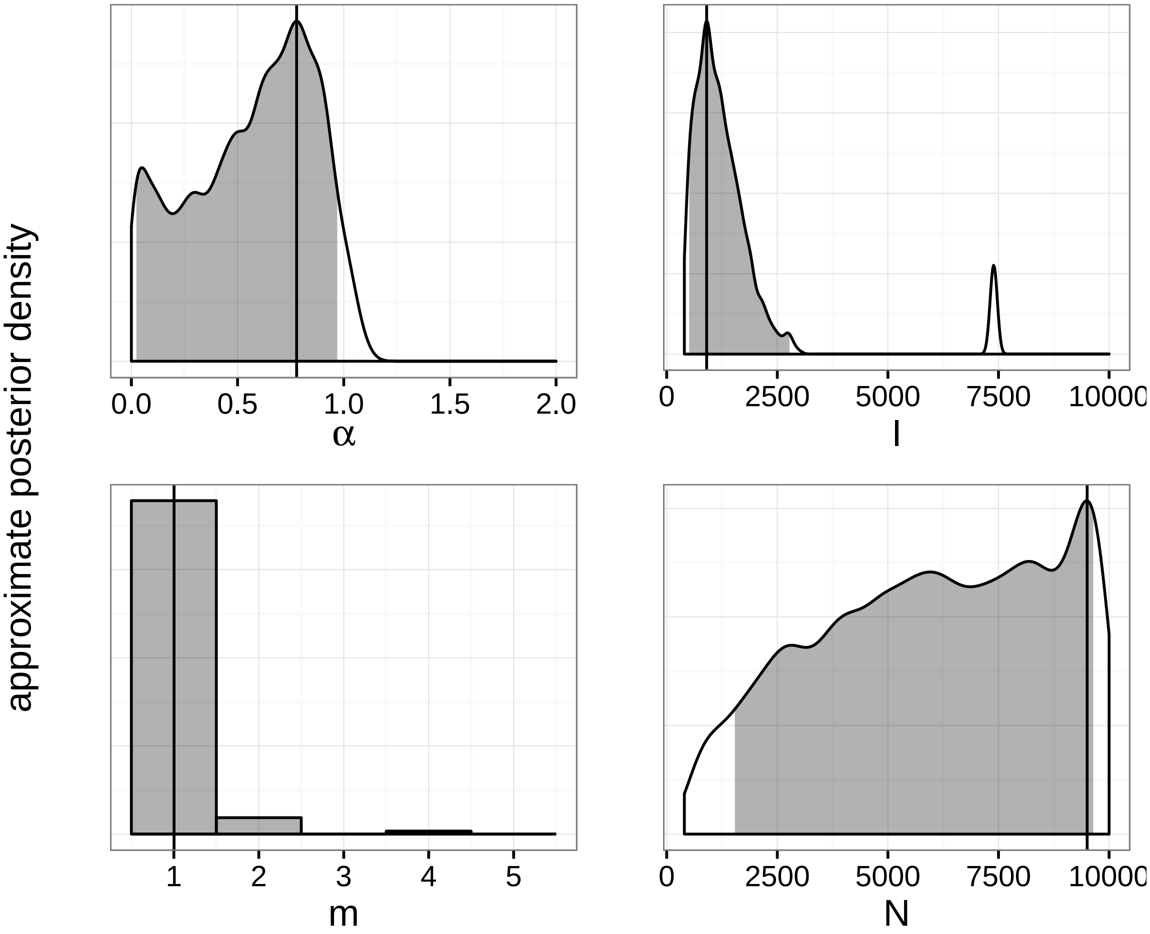

Figure A.30: Approximate marginal posterior distributions of BA model parameters for BC data. Vertical lines indicate maximum *a posteriori* estimates, and shaded areas are 95% highest posterior density intervals. *x*-axis indicates regions of nonzero prior density.

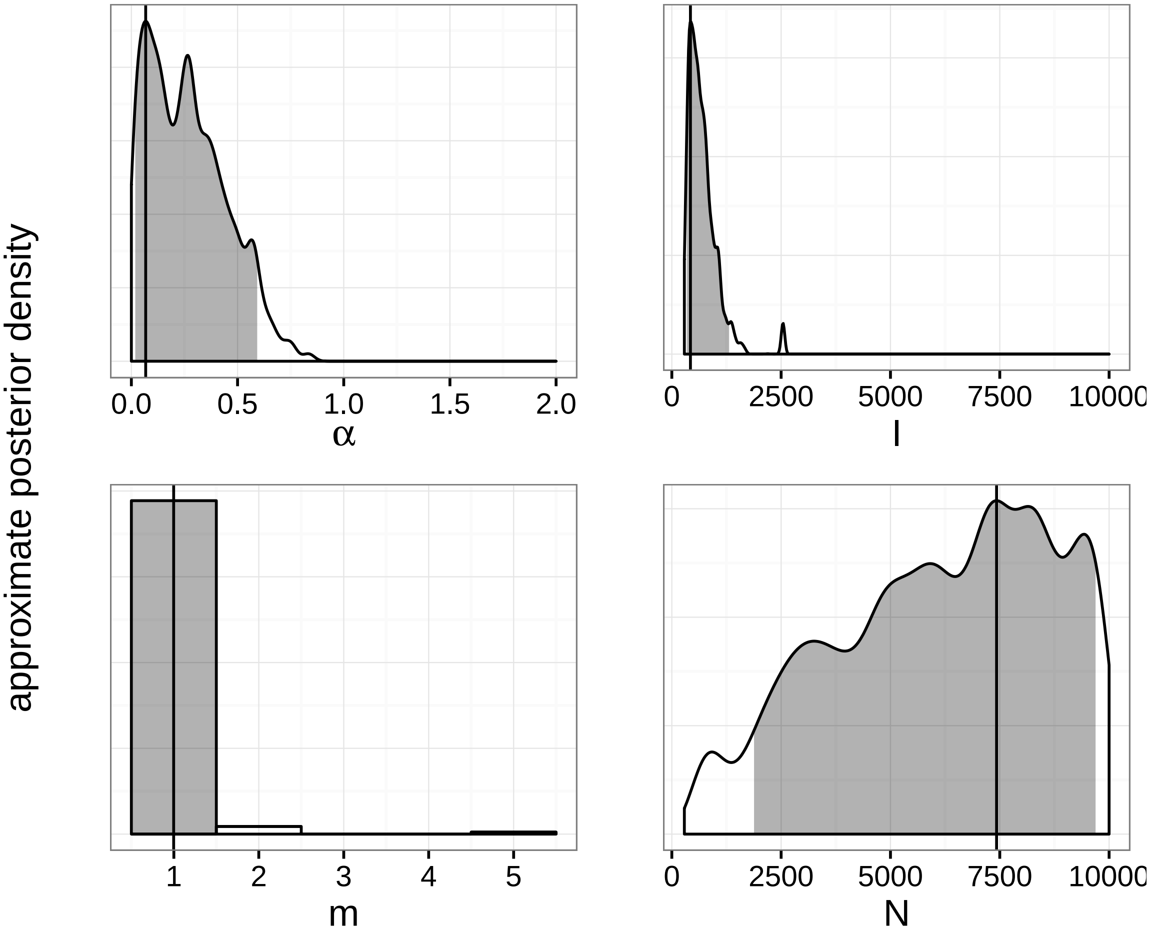

Figure A.31: Approximate marginal posterior distributions of BA model parameters for Cuevas et al. [146] data. Vertical lines indicate maximum *a posteriori* estimates, and shaded areas are 95% highest posterior density intervals. *x*-axis indicates regions of nonzero prior density.

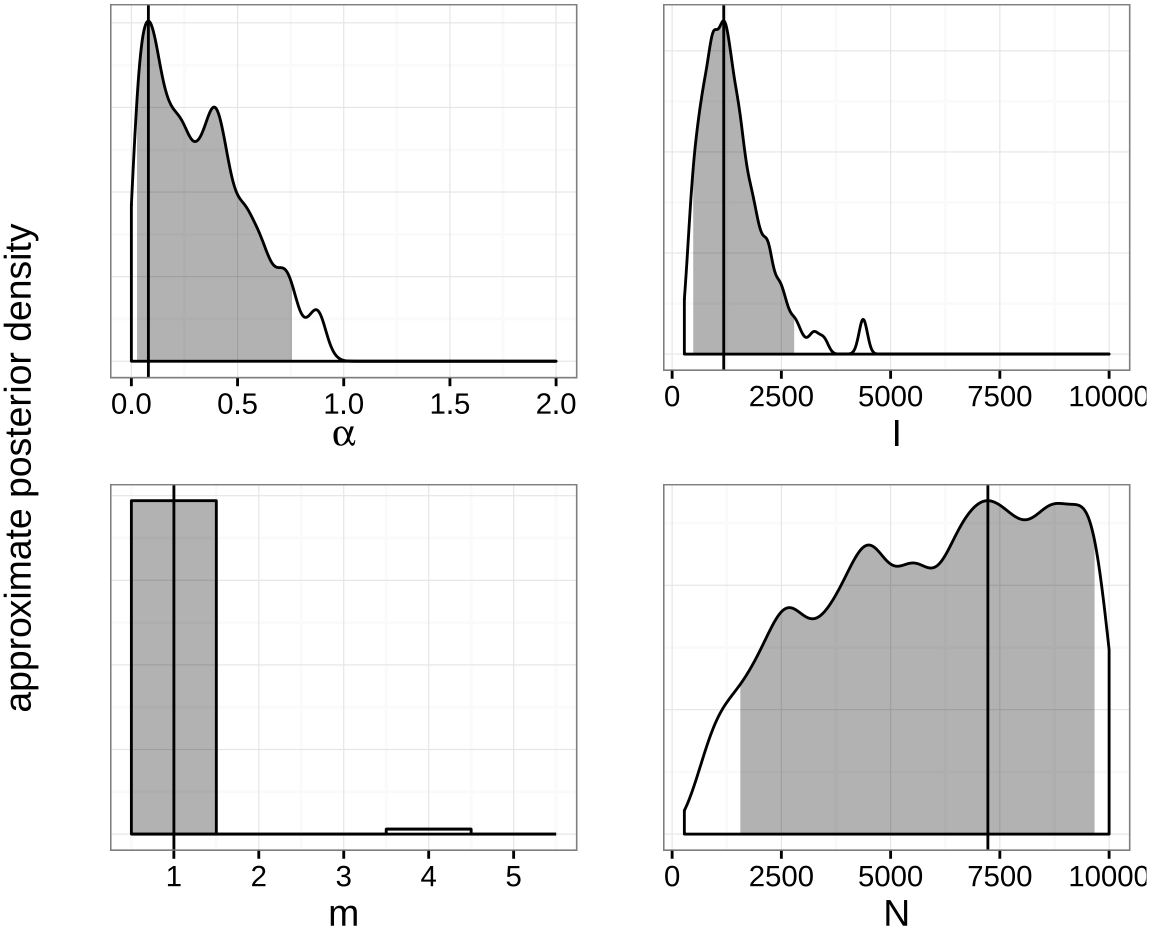

Figure A.32: Approximate marginal posterior distributions of BA model parameters for Li et al. [148] data. Vertical lines indicate maximum *a posteriori* estimates, and shaded areas are 95% highest posterior density intervals. *x*-axis indicates regions of nonzero prior density.

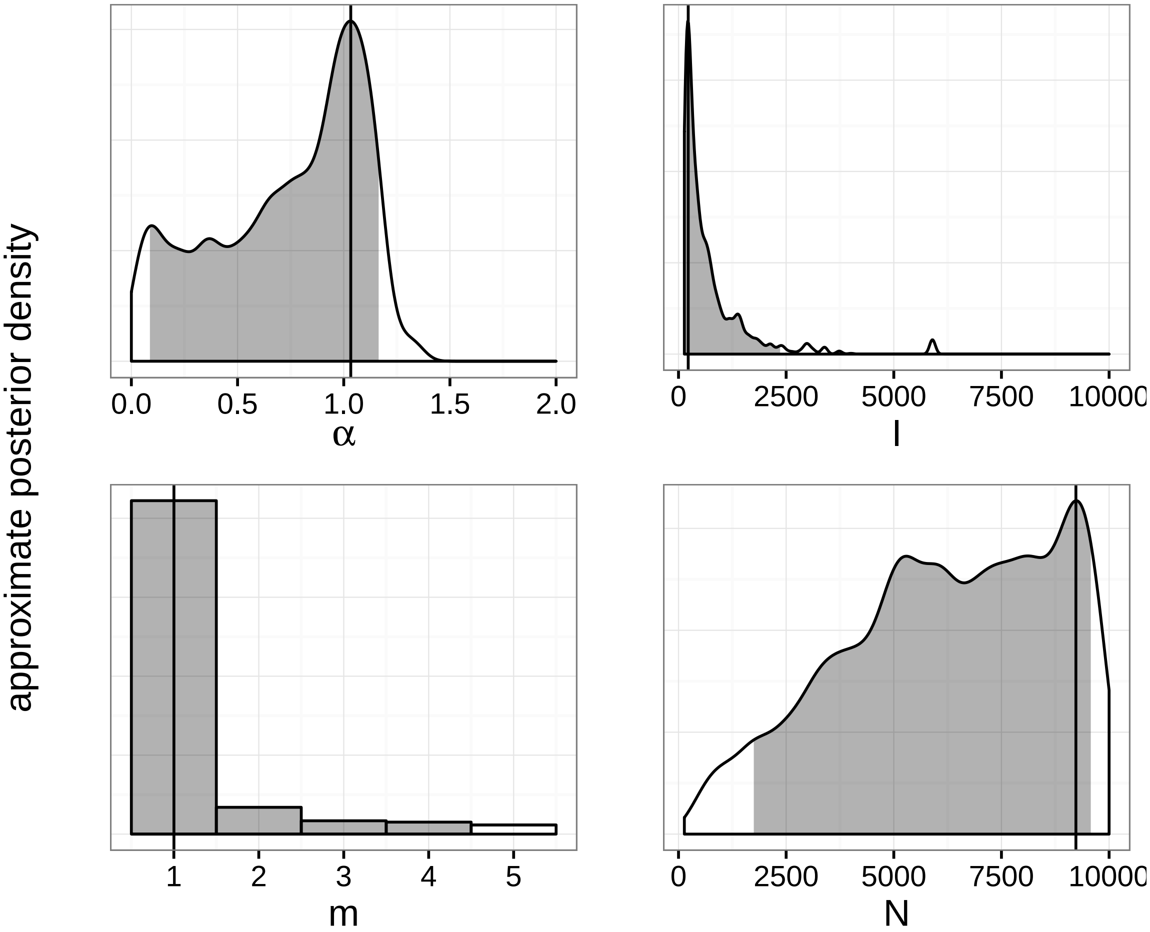

Figure A.33: Approximate marginal posterior distributions of BA model parameters for Niculescu et al. [145] data. Vertical lines indicate maximum *a posteriori* estimates, and shaded areas are 95% highest posterior density intervals. *x*-axis indicates regions of nonzero prior density.

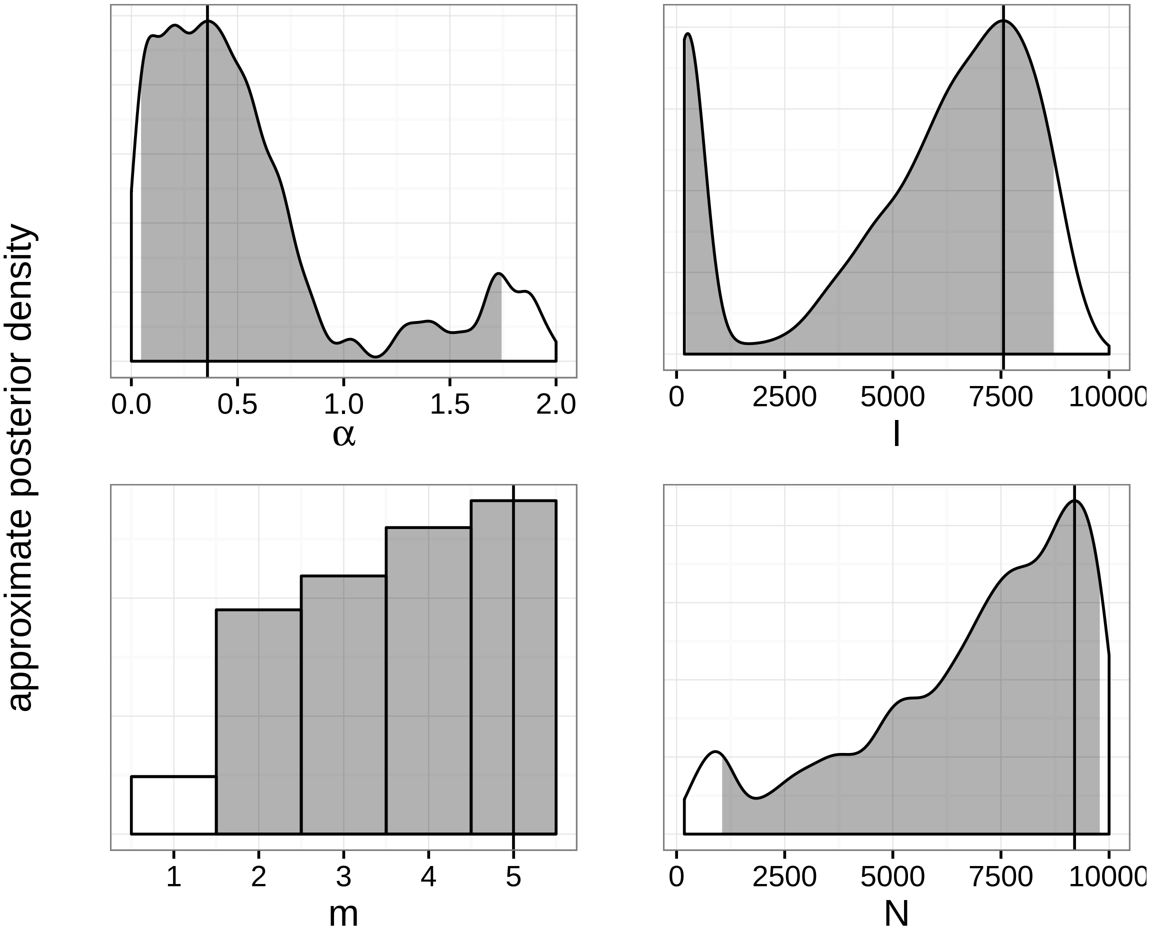

Figure A.34: Approximate marginal posterior distributions of BA model parameters for Novitsky et al. [147] and Novitsky et al. [149] data. Vertical lines indicate maximum *a posteriori* estimates, and shaded areas are 95% highest posterior density intervals. *x*-axis indicates regions of nonzero prior density.

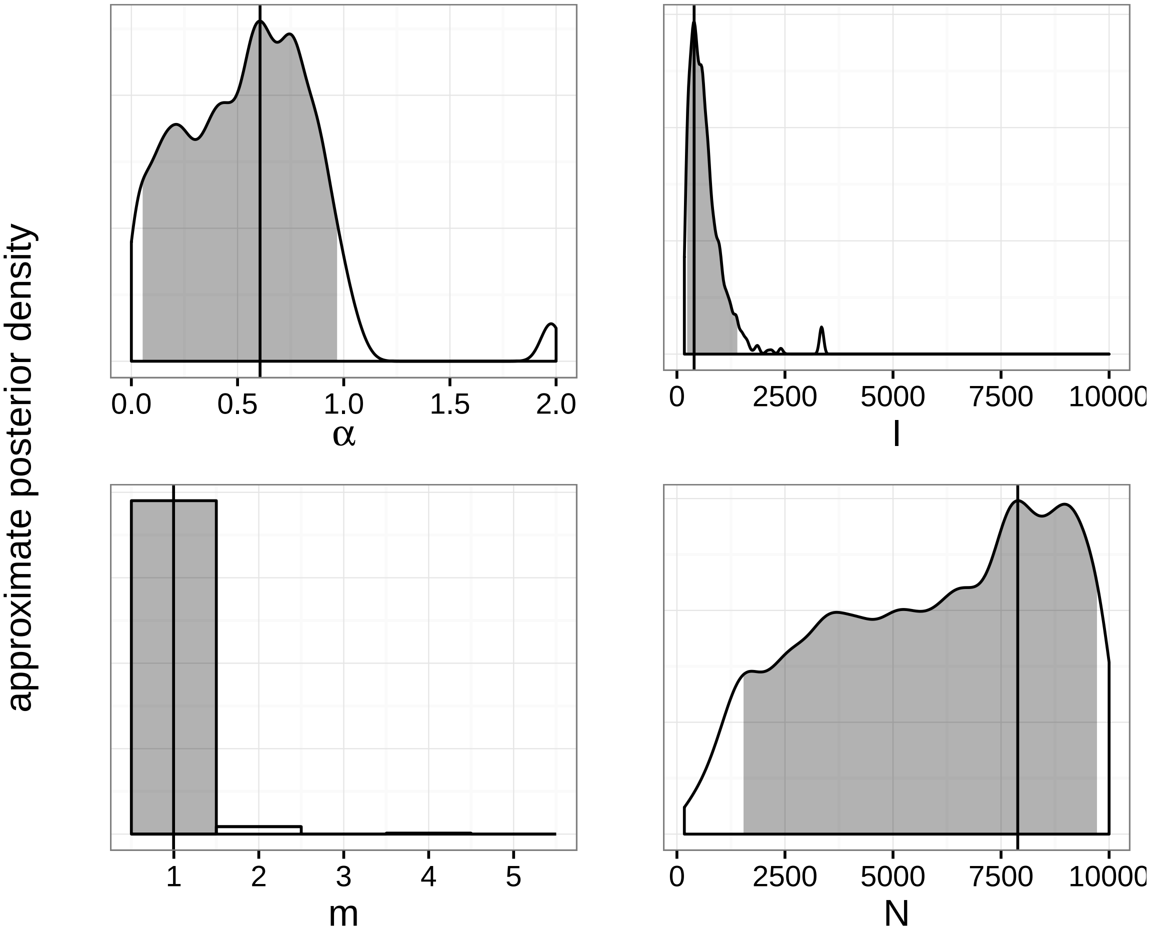

Figure A.35: Approximate marginal posterior distributions of BA model parameters for Wang et al. [30] data. Vertical lines indicate maximum *a posteriori* estimates, and shaded areas are 95% highest posterior density intervals. *x*-axis indicates regions of nonzero prior density.

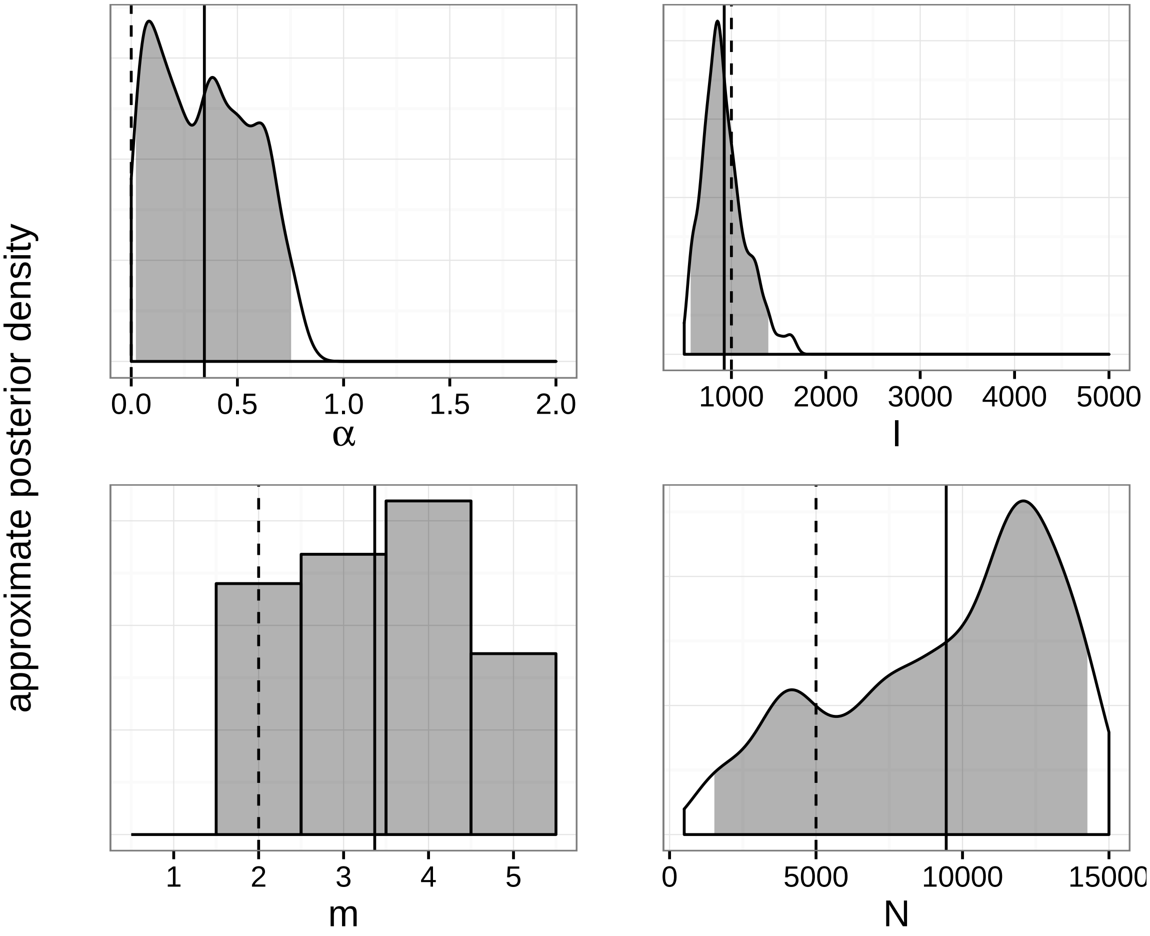

Figure A.36: Approximate marginal posterior distributions of BA model parameters obtained by applying *netabc* to a simulated transmission tree with BA parameter values $\alpha = 0.0$, $I = 1000$, $m = 2$, and $N = 5000$. Vertical dashed lines indicate true values. Shaded areas are 95% highest posterior density intervals. x-axes indicate regions of nonzero prior density.

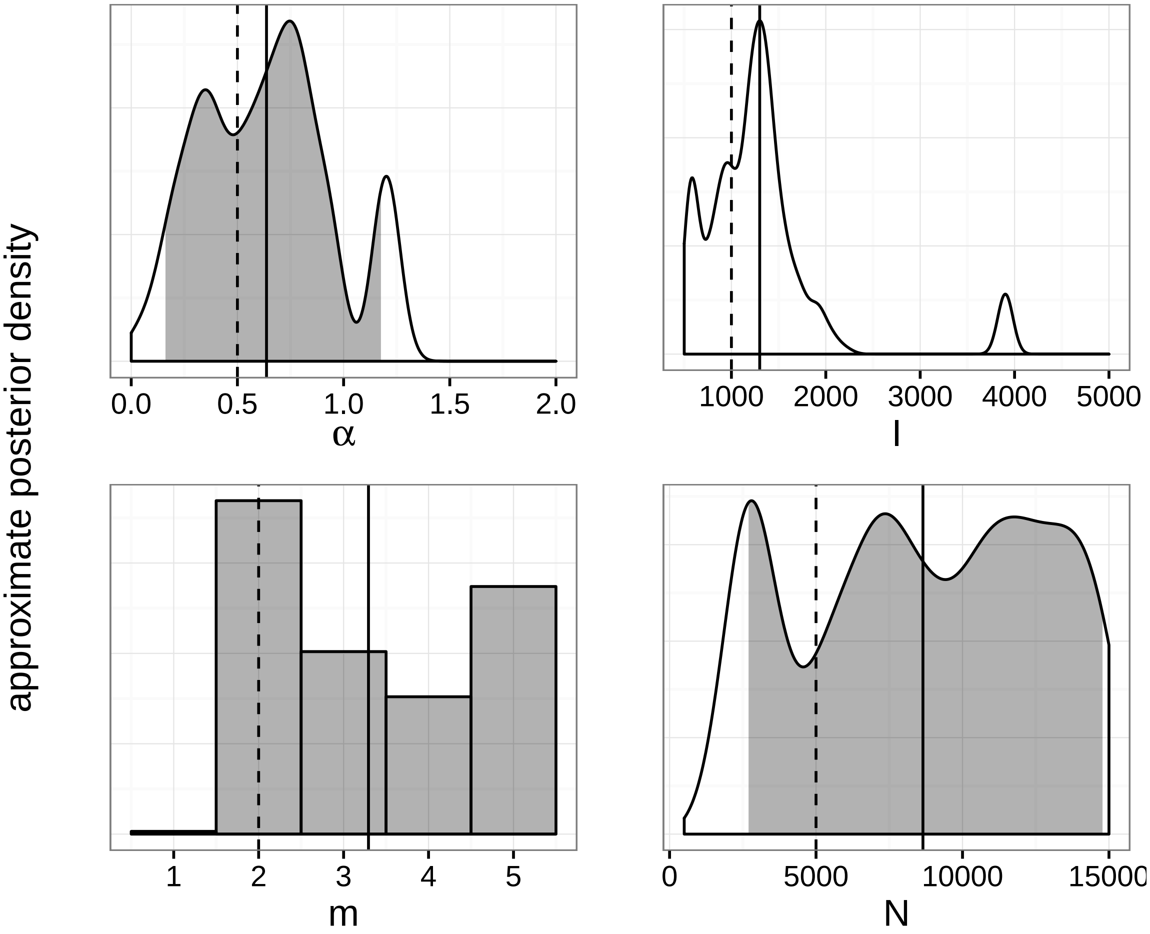

Figure A.37: Approximate marginal posterior distributions of BA model parameters obtained by applying *netabc* to a simulated transmission tree with BA parameter values $\alpha = 0.5$, $I = 1000$, $m = 2$, and $N = 5000$. Vertical dashed lines indicate true values. Shaded areas are 95% highest posterior density intervals. x-axes indicate regions of nonzero prior density.

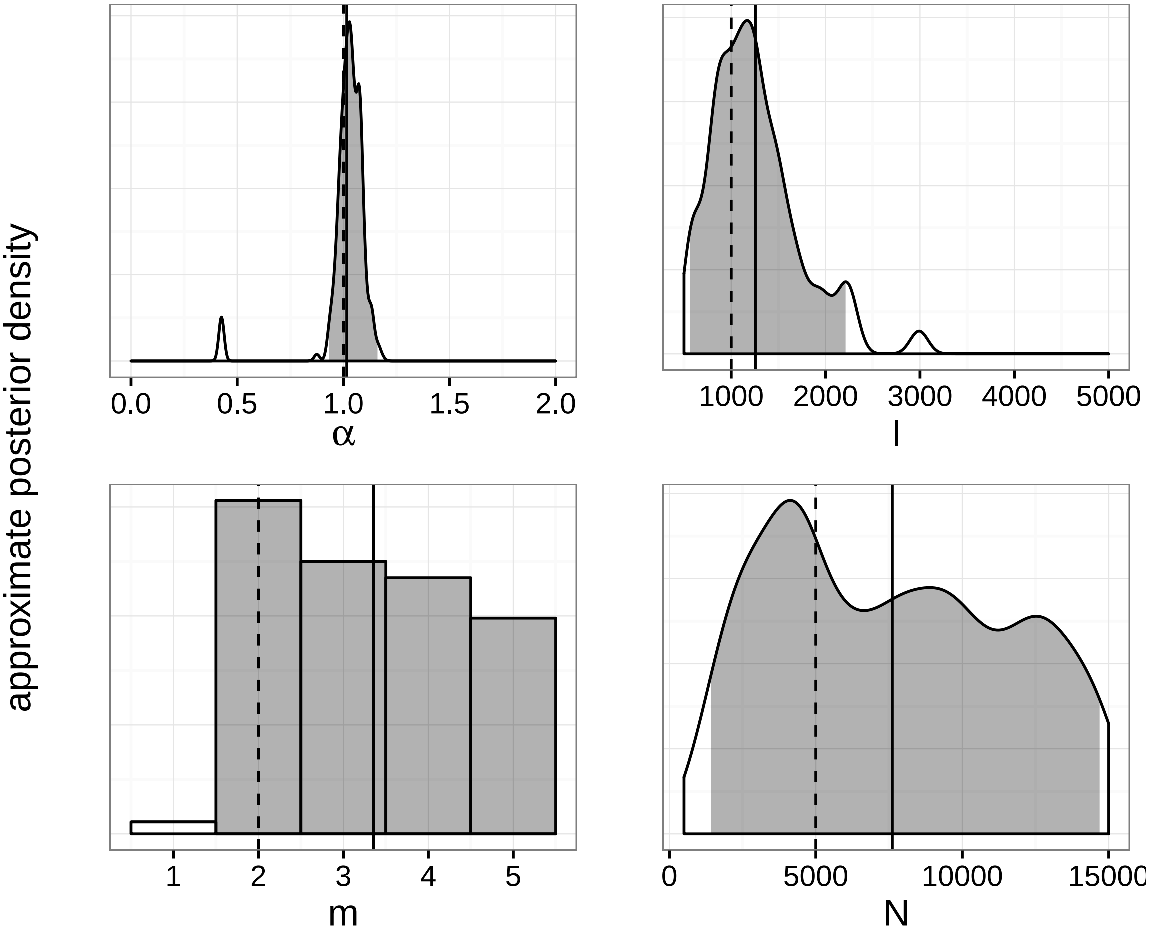

Figure A.38: Approximate marginal posterior distributions of BA model parameters obtained by applying *netabc* to a simulated transmission tree with BA parameter values $\alpha = 1.0$, $I = 1000$, $m = 2$, and $N = 5000$. Vertical dashed lines indicate true values. Shaded areas are 95% highest posterior density intervals. x-axes indicate regions of nonzero prior density.

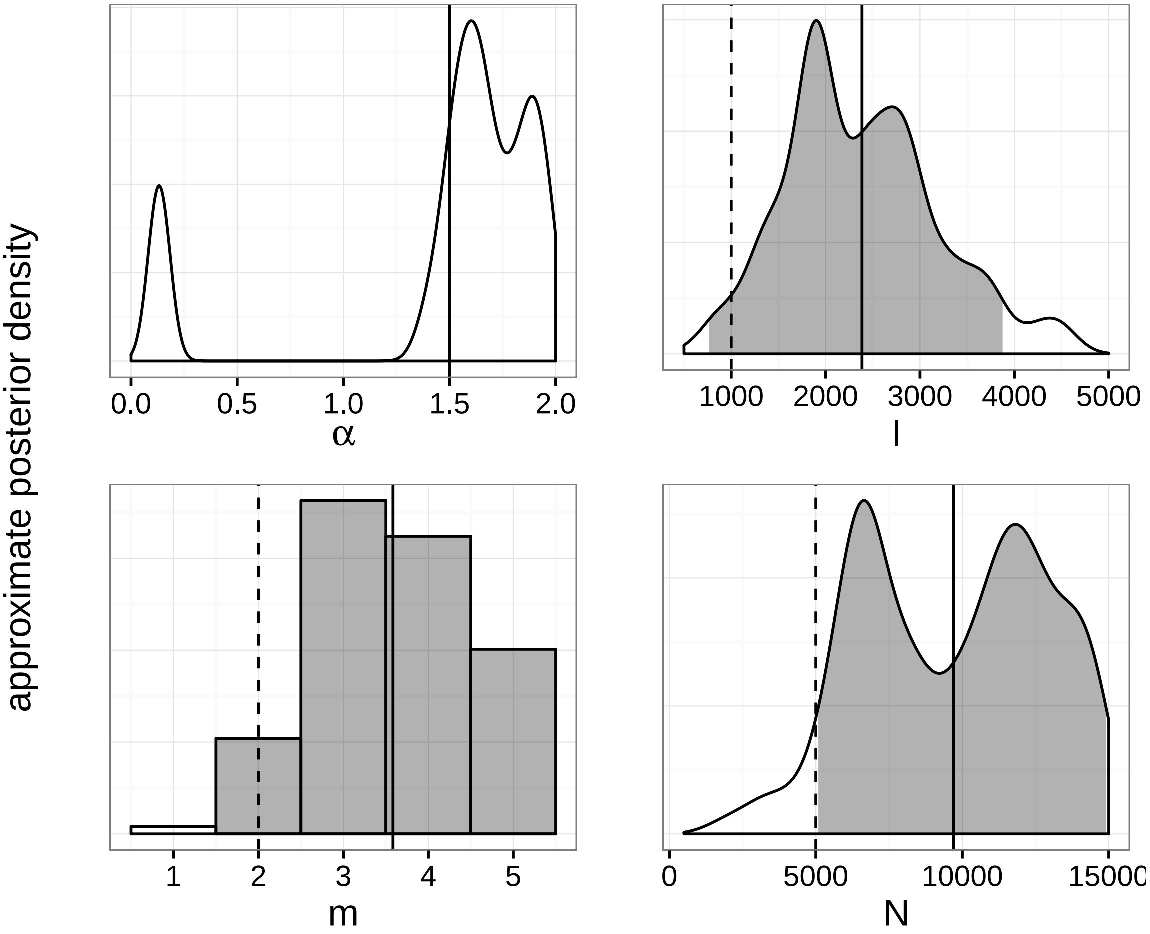

Figure A.39: Approximate marginal posterior distributions of BA model parameters obtained by applying *netabc* to a simulated transmission tree with BA parameter values $\alpha = 1.5$, $I = 1000$, $m = 2$, and $N = 5000$. Vertical dashed lines indicate true values. Shaded areas are 95% highest posterior density intervals. x-axes indicate regions of nonzero prior density.

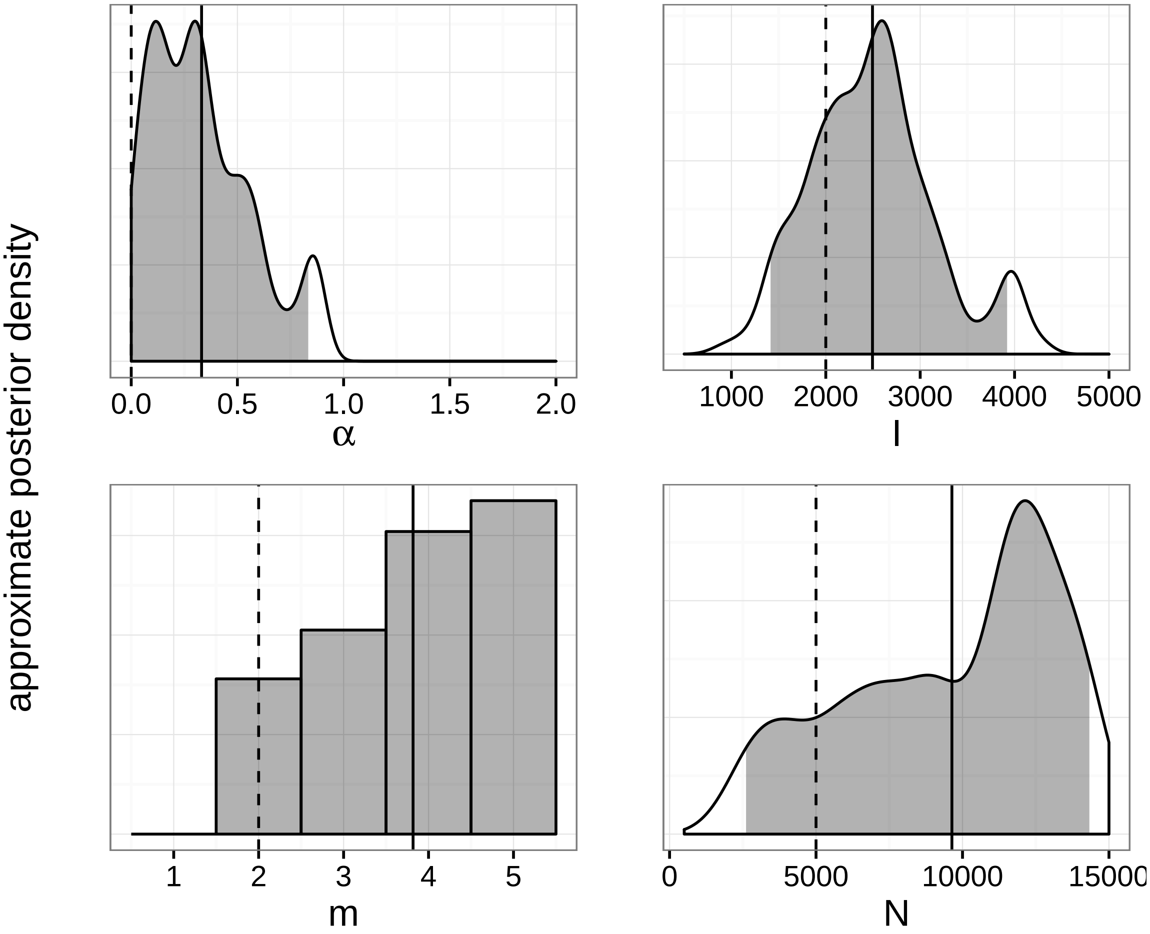

Figure A.40: Approximate marginal posterior distributions of BA model parameters obtained by applying *netabc* to a simulated transmission tree with BA parameter values $\alpha = 0.0$, $I = 2000$, $m = 2$, and $N = 5000$. Vertical dashed lines indicate true values. Shaded areas are 95% highest posterior density intervals. x-axes indicate regions of nonzero prior density.

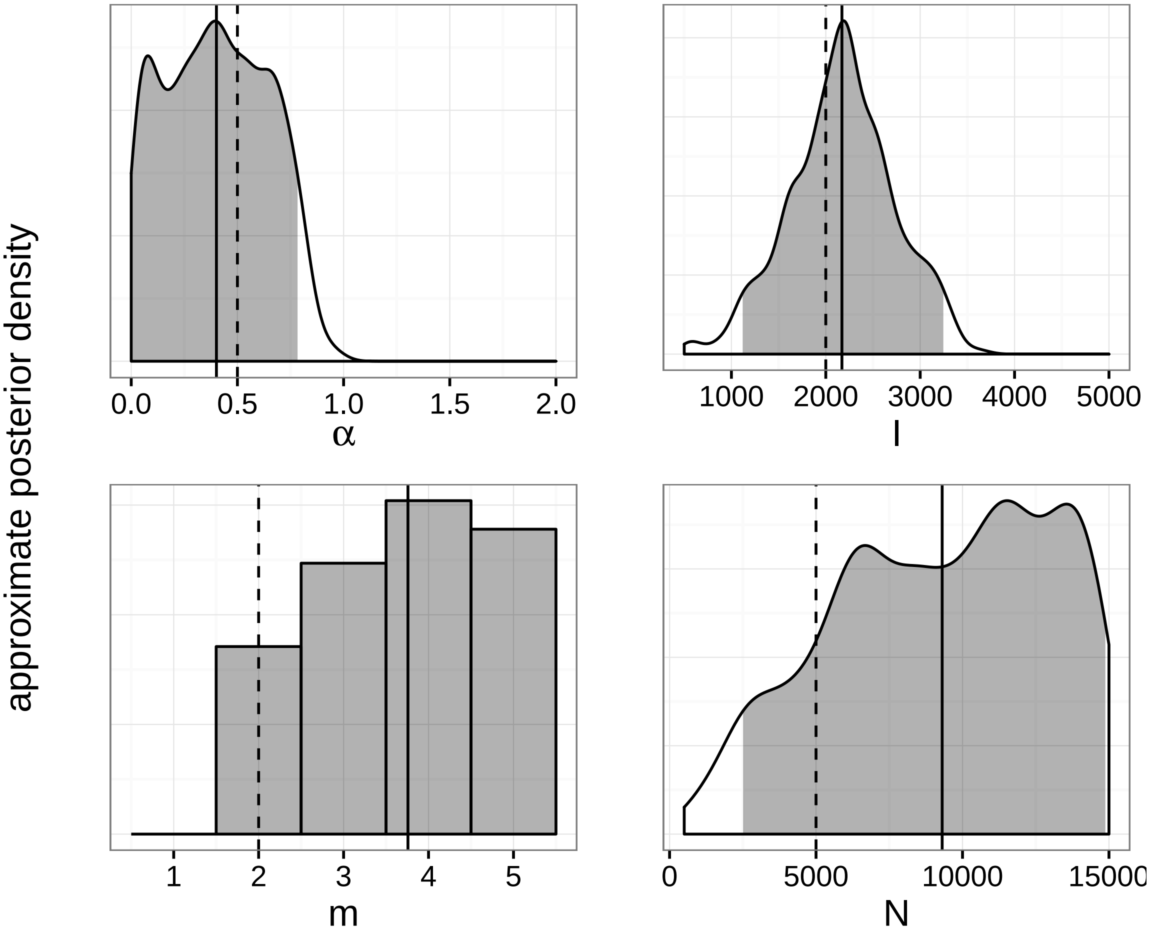

Figure A.41: Approximate marginal posterior distributions of BA model parameters obtained by applying *netabc* to a simulated transmission tree with BA parameter values $\alpha = 0.5$, $I = 2000$, $m = 2$, and $N = 5000$. Vertical dashed lines indicate true values. Shaded areas are 95% highest posterior density intervals. x-axes indicate regions of nonzero prior density.

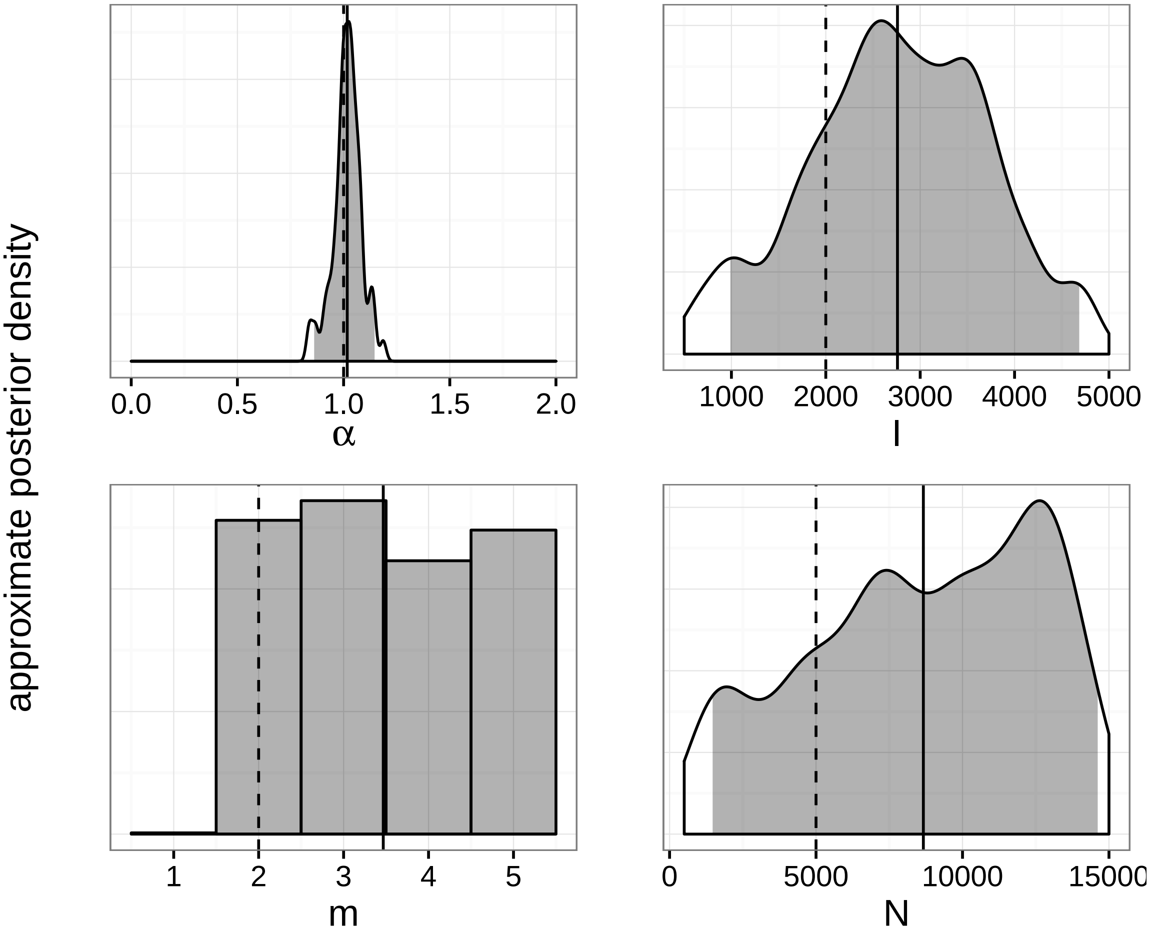

Figure A.42: Approximate marginal posterior distributions of BA model parameters obtained by applying *netabc* to a simulated transmission tree with BA parameter values $\alpha = 1.0$, $I = 2000$, $m = 2$, and $N = 5000$. Vertical dashed lines indicate true values. Shaded areas are 95% highest posterior density intervals. x-axes indicate regions of nonzero prior density.

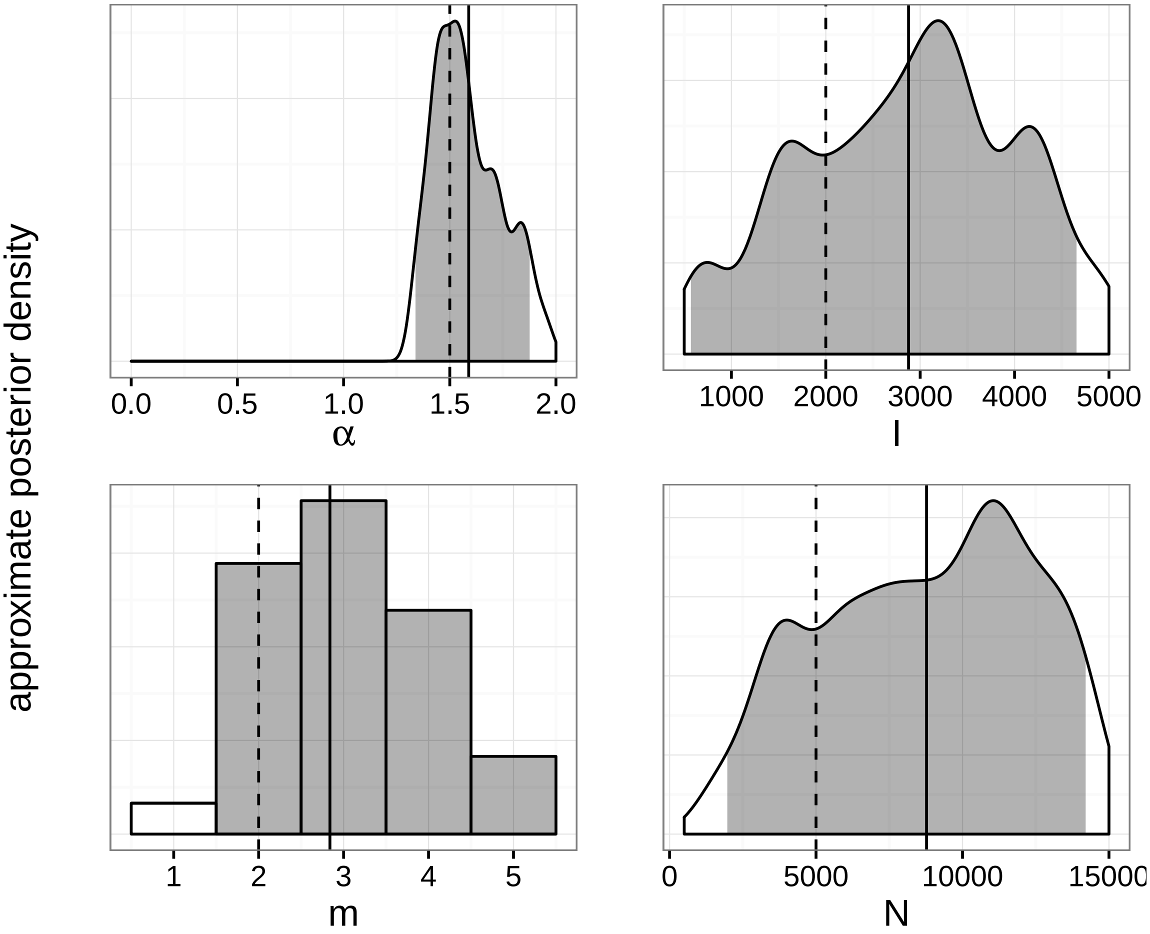

Figure A.43: Approximate marginal posterior distributions of BA model parameters obtained by applying *netabc* to a simulated transmission tree with BA parameter values $\alpha = 1.5$, $I = 2000$, $m = 2$, and $N = 5000$. Vertical dashed lines indicate true values. Shaded areas are 95% highest posterior density intervals. x-axes indicate regions of nonzero prior density.

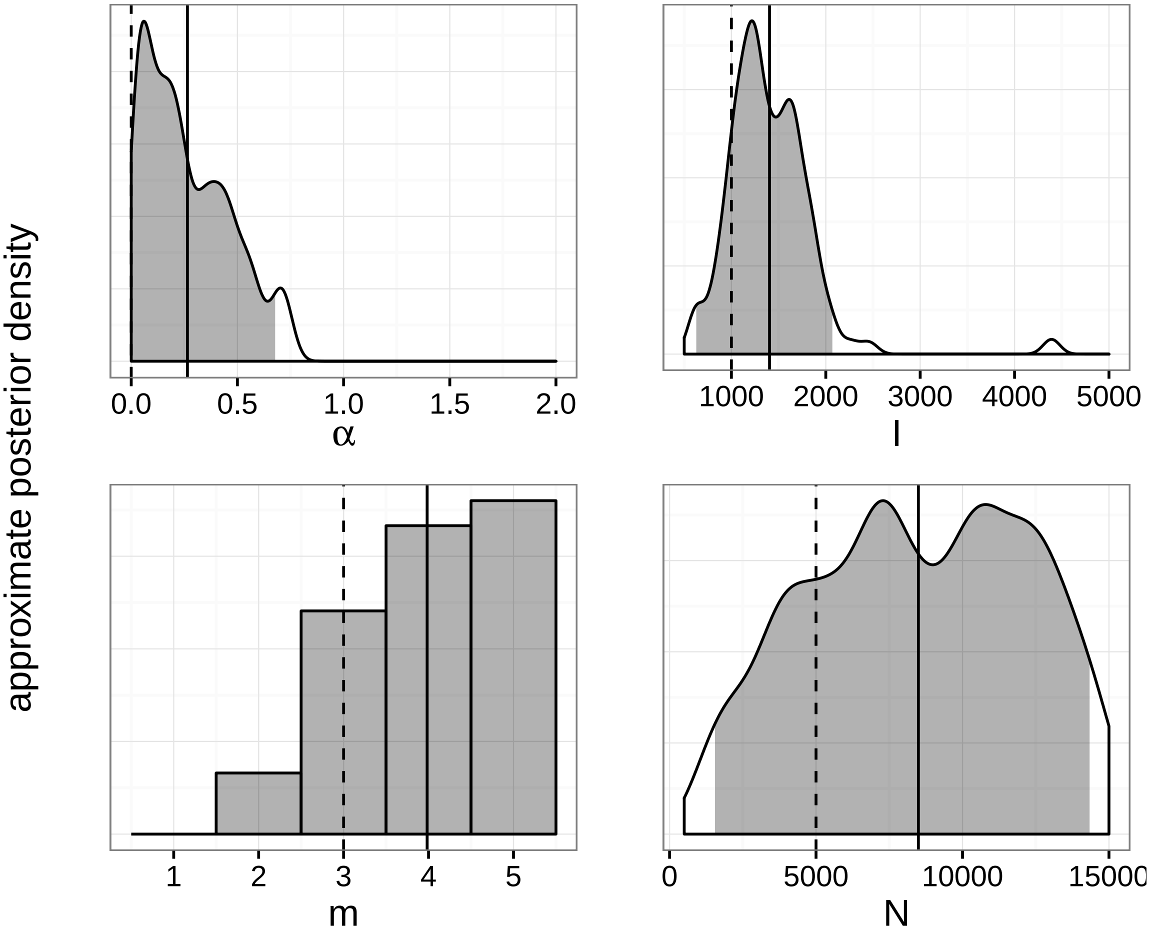

Figure A.44: Approximate marginal posterior distributions of BA model parameters obtained by applying *netabc* to a simulated transmission tree with BA parameter values $\alpha = 0.0$, $I = 1000$, $m = 3$, and $N = 5000$. Vertical dashed lines indicate true values. Shaded areas are 95% highest posterior density intervals. x-axes indicate regions of nonzero prior density.

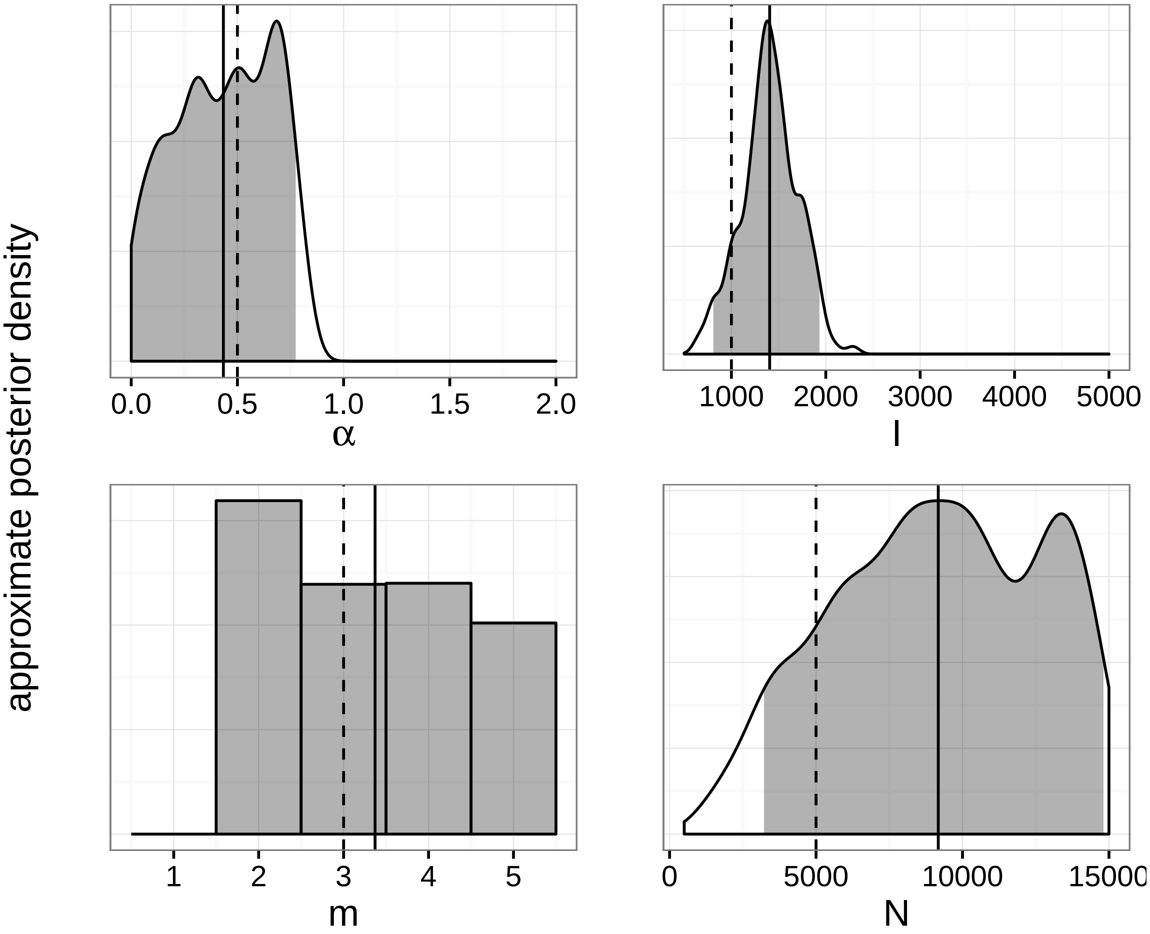

Figure A.45: Approximate marginal posterior distributions of BA model parameters obtained by applying *netabc* to a simulated transmission tree with BA parameter values $\alpha = 0.5$, $I = 1000$, $m = 3$, and $N = 5000$. Vertical dashed lines indicate true values. Shaded areas are 95% highest posterior density intervals. *x*-axes indicate regions of nonzero prior density.

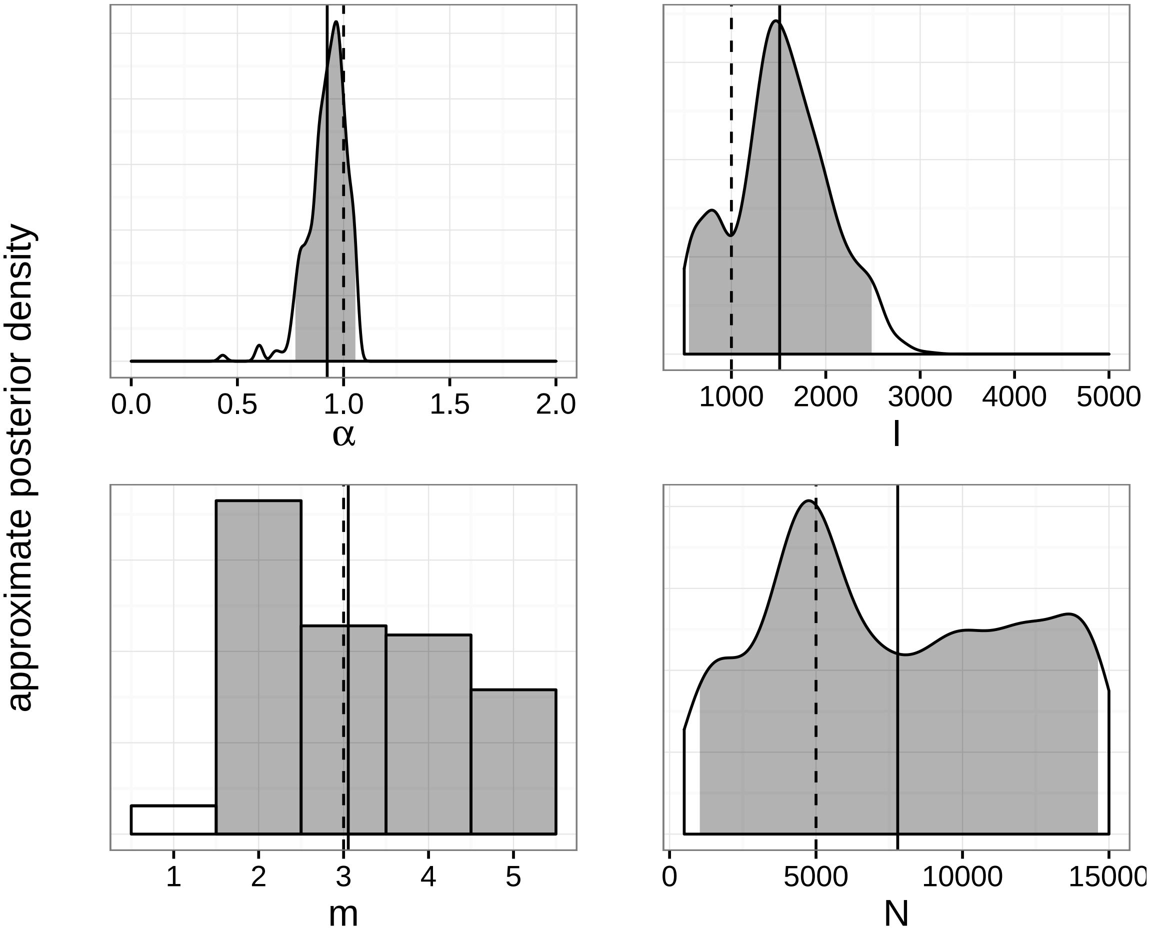

Figure A.46: Approximate marginal posterior distributions of BA model parameters obtained by applying *netabc* to a simulated transmission tree with BA parameter values $\alpha = 1.0$, $I = 1000$, $m = 3$, and $N = 5000$. Vertical dashed lines indicate true values. Shaded areas are 95% highest posterior density intervals. x-axes indicate regions of nonzero prior density.

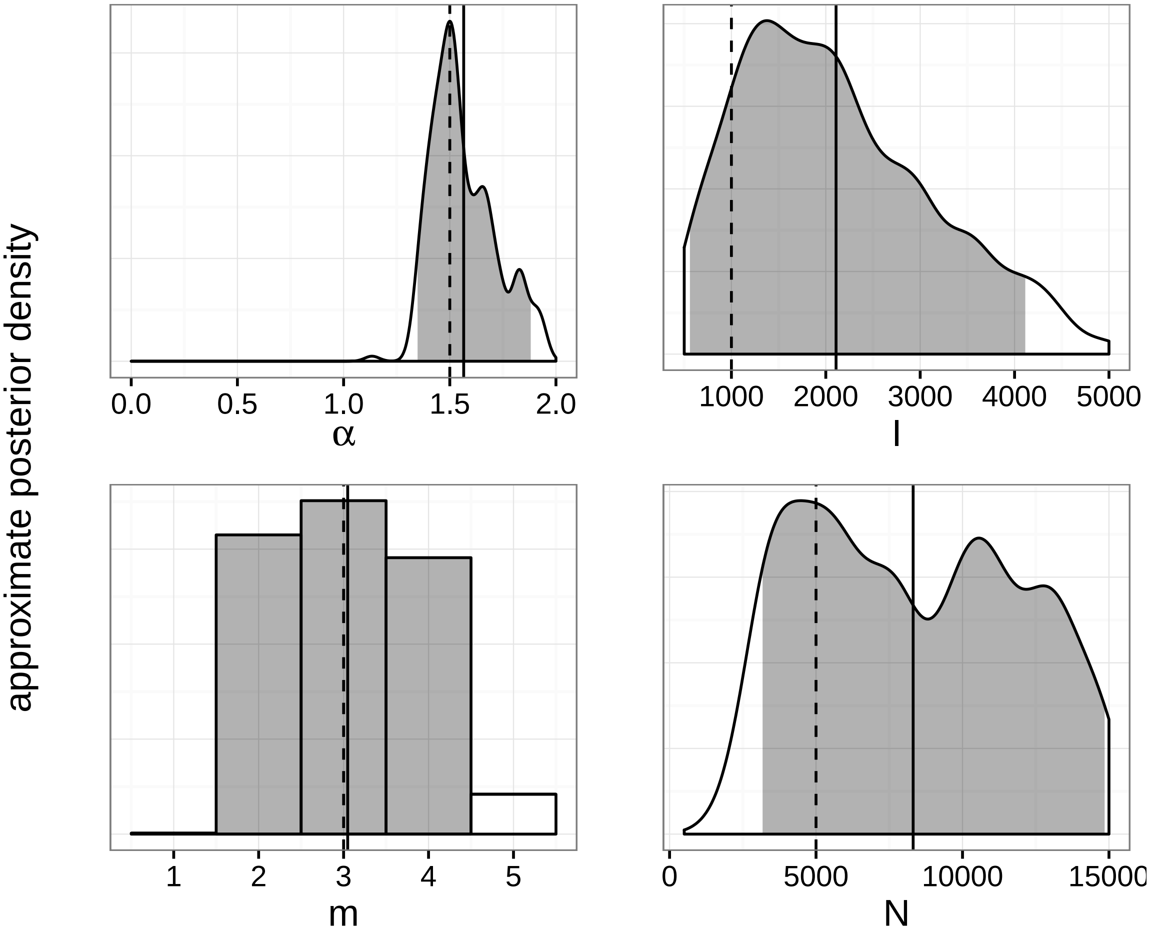

Figure A.47: Approximate marginal posterior distributions of BA model parameters obtained by applying *netabc* to a simulated transmission tree with BA parameter values $\alpha = 1.5$, $I = 1000$, $m = 3$, and $N = 5000$. Vertical dashed lines indicate true values. Shaded areas are 95% highest posterior density intervals. x-axes indicate regions of nonzero prior density.

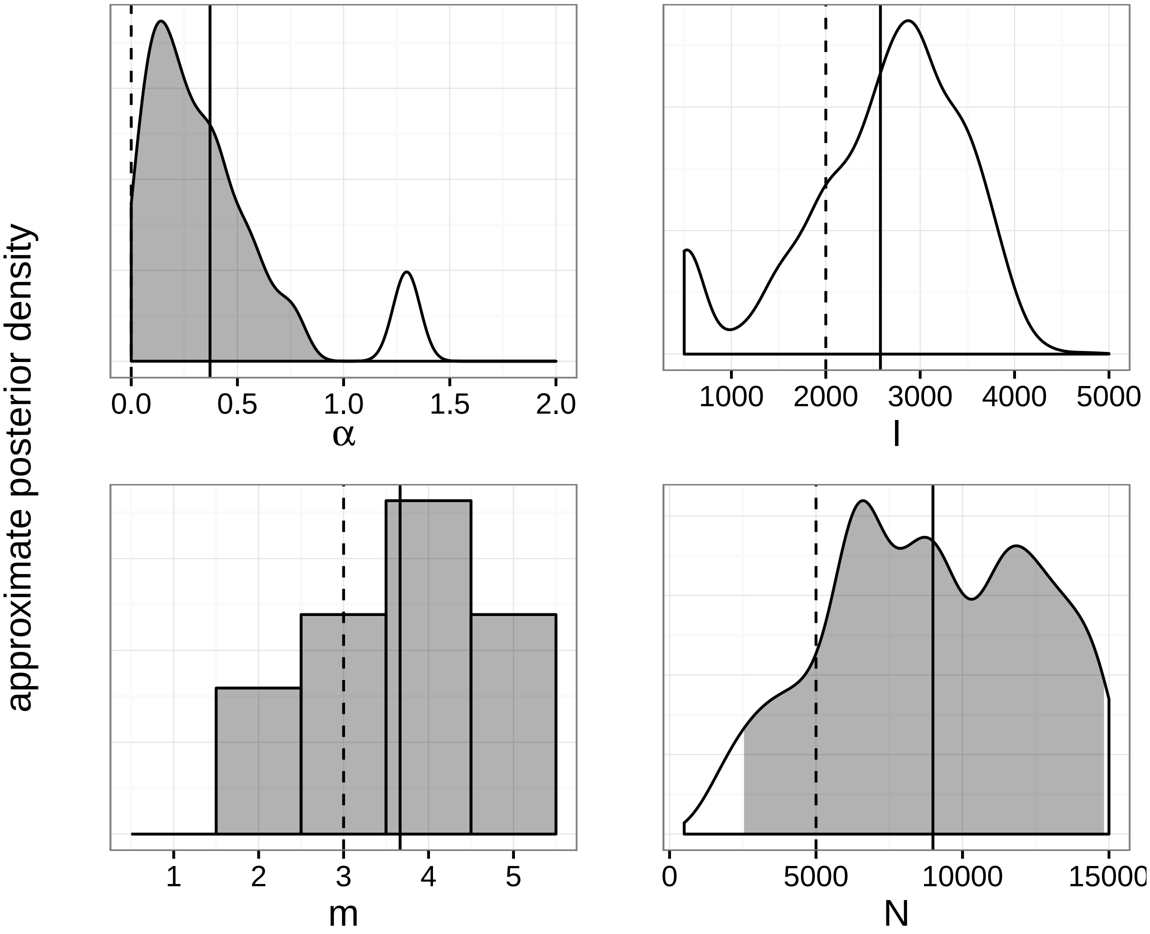

Figure A.48: Approximate marginal posterior distributions of BA model parameters obtained by applying *netabc* to a simulated transmission tree with BA parameter values $\alpha = 0.0$, $I = 2000$, $m = 3$, and $N = 5000$. Vertical dashed lines indicate true values. Shaded areas are 95% highest posterior density intervals. x-axes indicate regions of nonzero prior density.

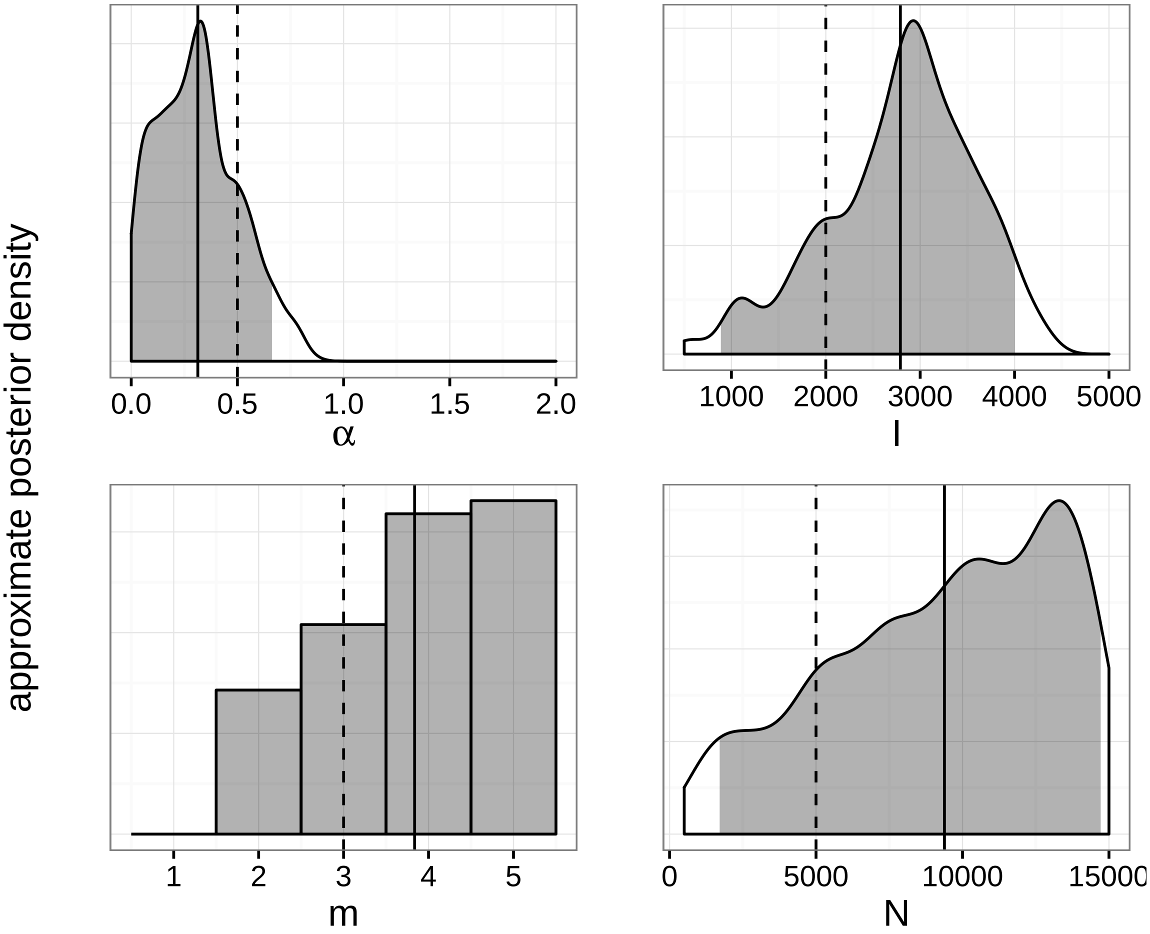

Figure A.49: Approximate marginal posterior distributions of BA model parameters obtained by applying *netabc* to a simulated transmission tree with BA parameter values $\alpha = 0.5$, $I = 2000$, $m = 3$, and $N = 5000$. Vertical dashed lines indicate true values. Shaded areas are 95% highest posterior density intervals. x-axes indicate regions of nonzero prior density.

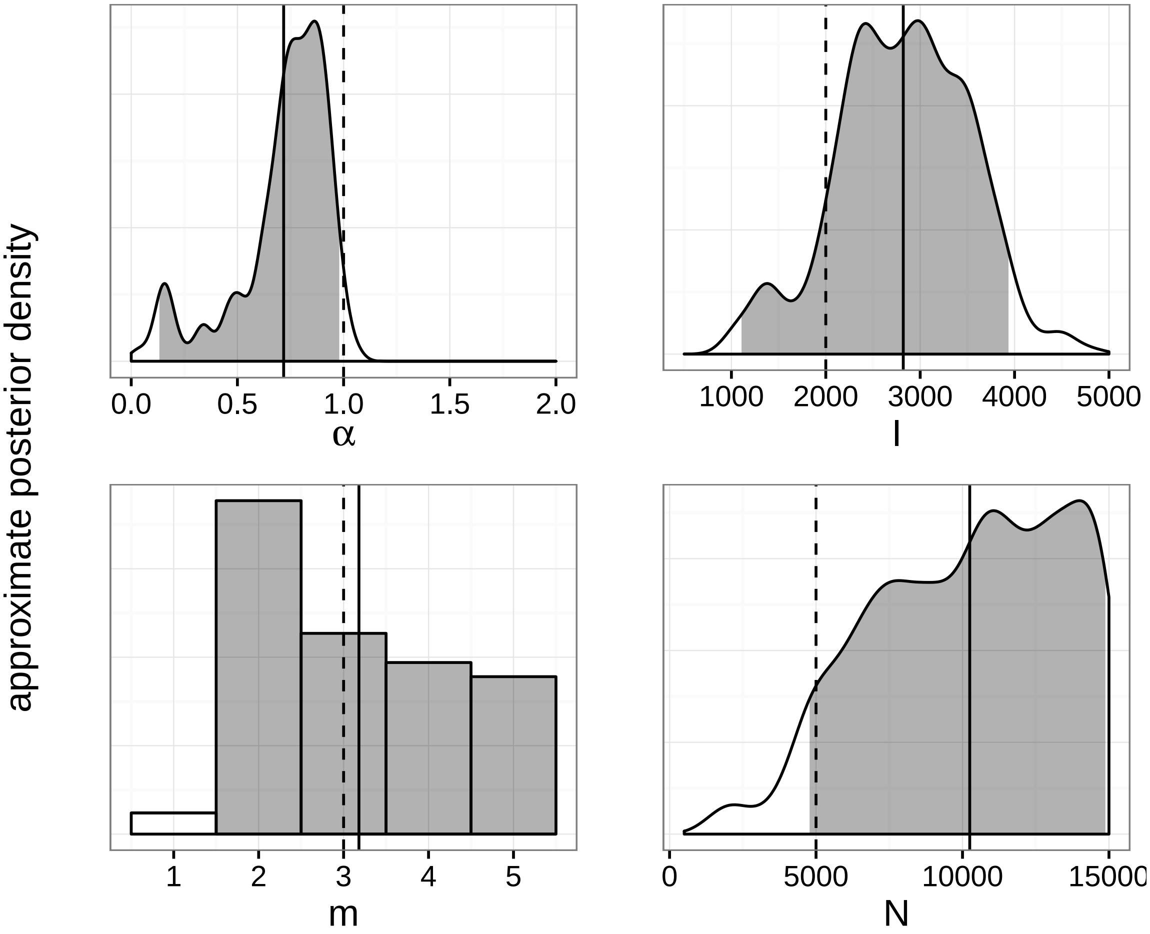

Figure A.50: Approximate marginal posterior distributions of BA model parameters obtained by applying *netabc* to a simulated transmission tree with BA parameter values $\alpha = 1.0$, $I = 2000$, $m = 3$, and $N = 5000$. Vertical dashed lines indicate true values. Shaded areas are 95% highest posterior density intervals. x-axes indicate regions of nonzero prior density.

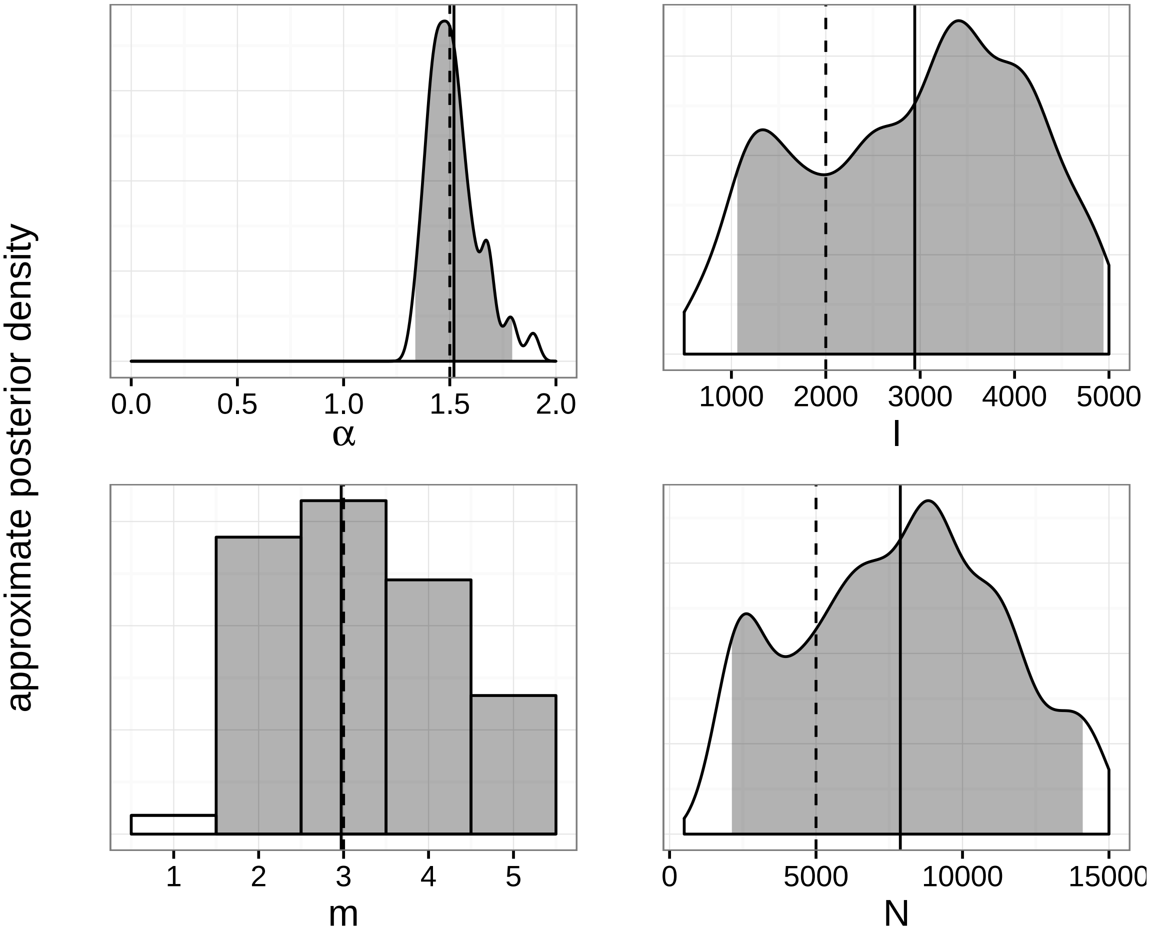

Figure A.51: Approximate marginal posterior distributions of BA model parameters obtained by applying *netabc* to a simulated transmission tree with BA parameter values $\alpha = 1.5$, $I = 2000$, $m = 3$, and $N = 5000$. Vertical dashed lines indicate true values. Shaded areas are 95% highest posterior density intervals. x-axes indicate regions of nonzero prior density.

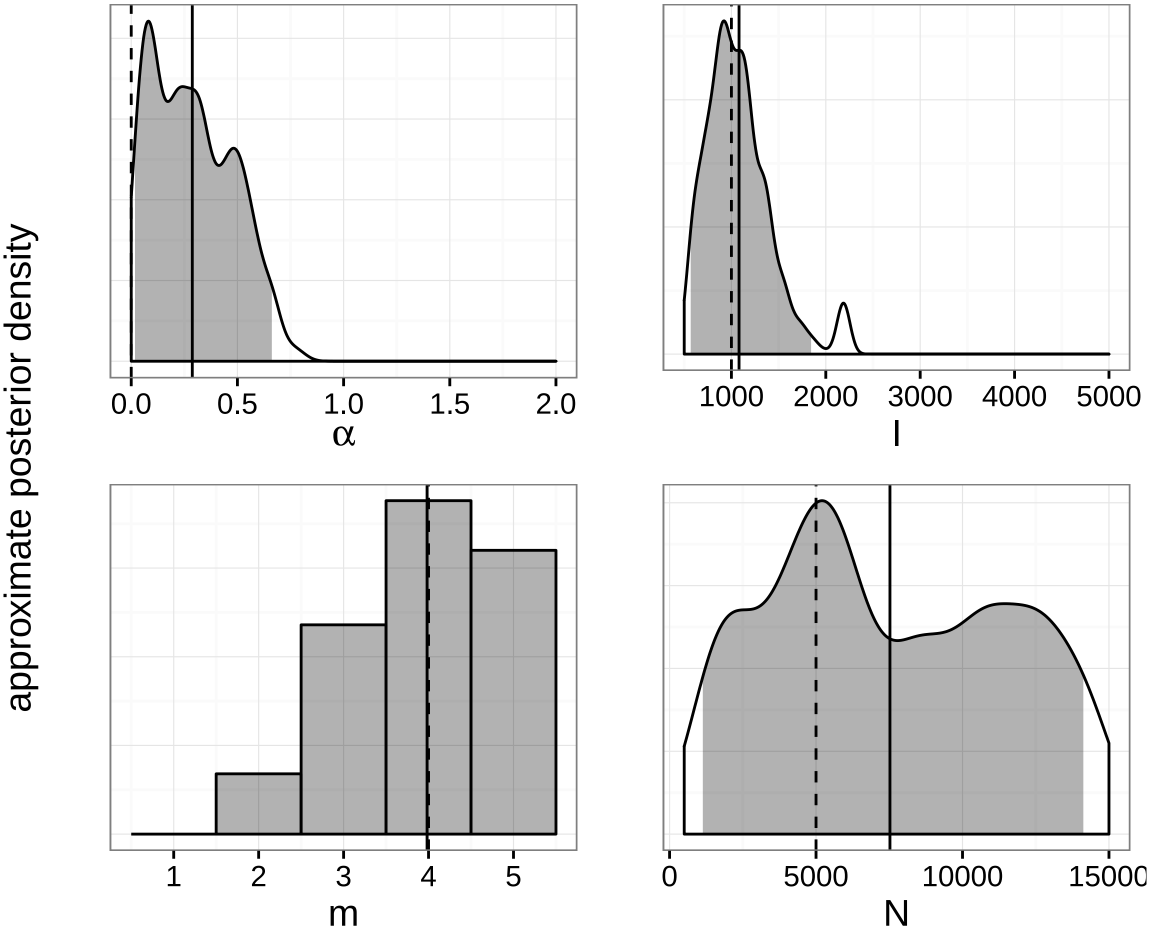

Figure A.52: Approximate marginal posterior distributions of BA model parameters obtained by applying *netabc* to a simulated transmission tree with BA parameter values $\alpha = 0.0$, $I = 1000$, $m = 4$, and $N = 5000$. Vertical dashed lines indicate true values. Shaded areas are 95% highest posterior density intervals. x-axes indicate regions of nonzero prior density.

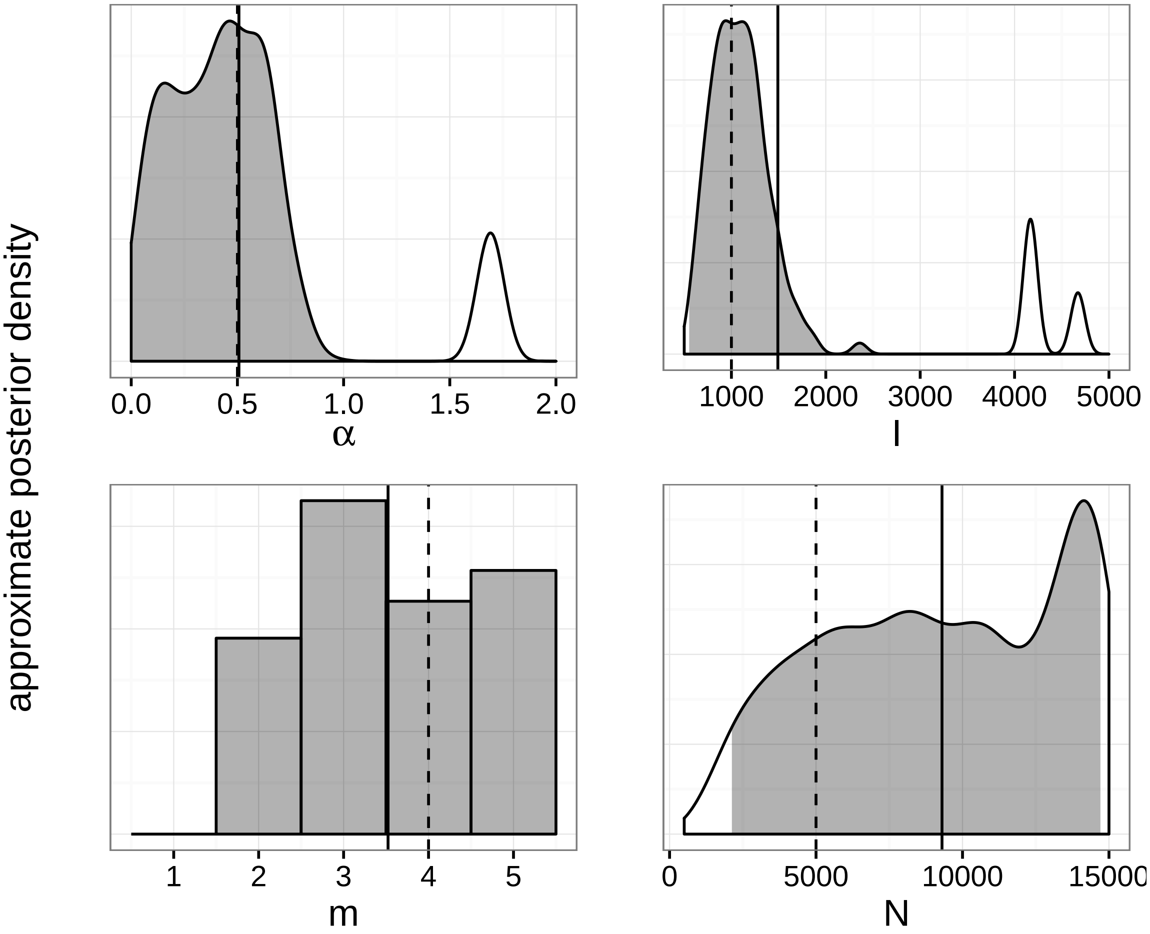

Figure A.53: Approximate marginal posterior distributions of BA model parameters obtained by applying *netabc* to a simulated transmission tree with BA parameter values $\alpha = 0.5$, $I = 1000$, $m = 4$, and $N = 5000$. Vertical dashed lines indicate true values. Shaded areas are 95% highest posterior density intervals. x-axes indicate regions of nonzero prior density.

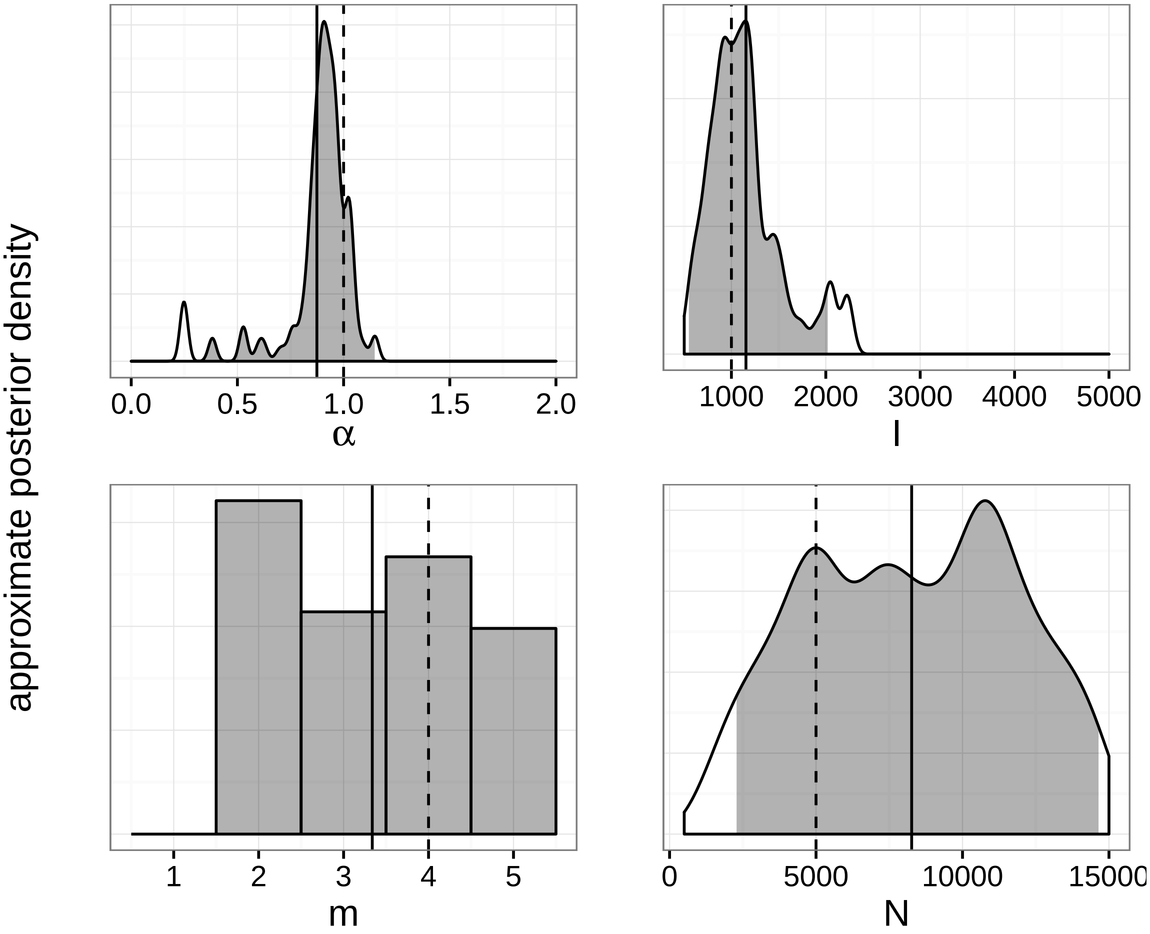

Figure A.54: Approximate marginal posterior distributions of BA model parameters obtained by applying *netabc* to a simulated transmission tree with BA parameter values $\alpha = 1.0$, $I = 1000$, $m = 4$, and $N = 5000$. Vertical dashed lines indicate true values. Shaded areas are 95% highest posterior density intervals. x-axes indicate regions of nonzero prior density.

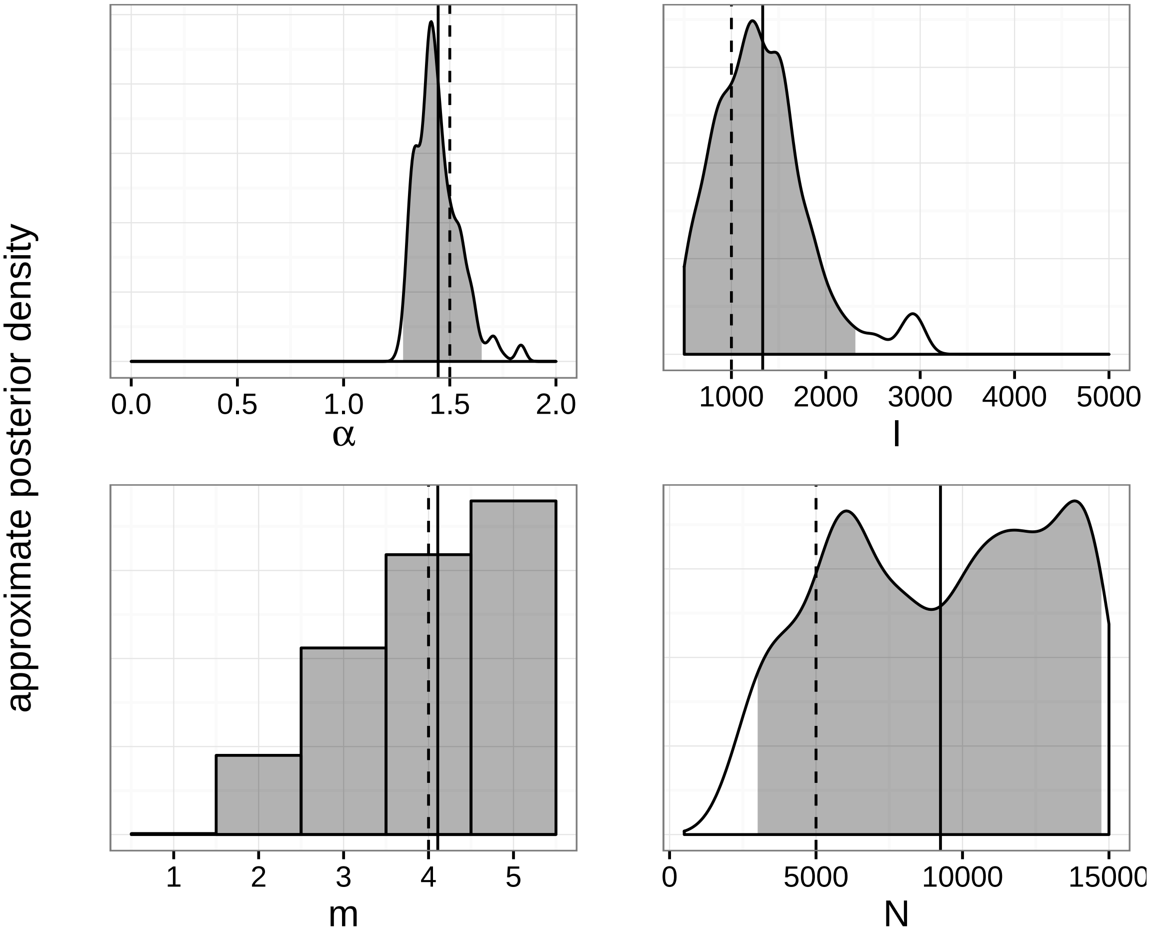

Figure A.55: Approximate marginal posterior distributions of BA model parameters obtained by applying *netabc* to a simulated transmission tree with BA parameter values $\alpha = 1.5$, $I = 1000$, $m = 4$, and $N = 5000$. Vertical dashed lines indicate true values. Shaded areas are 95% highest posterior density intervals. x-axes indicate regions of nonzero prior density.

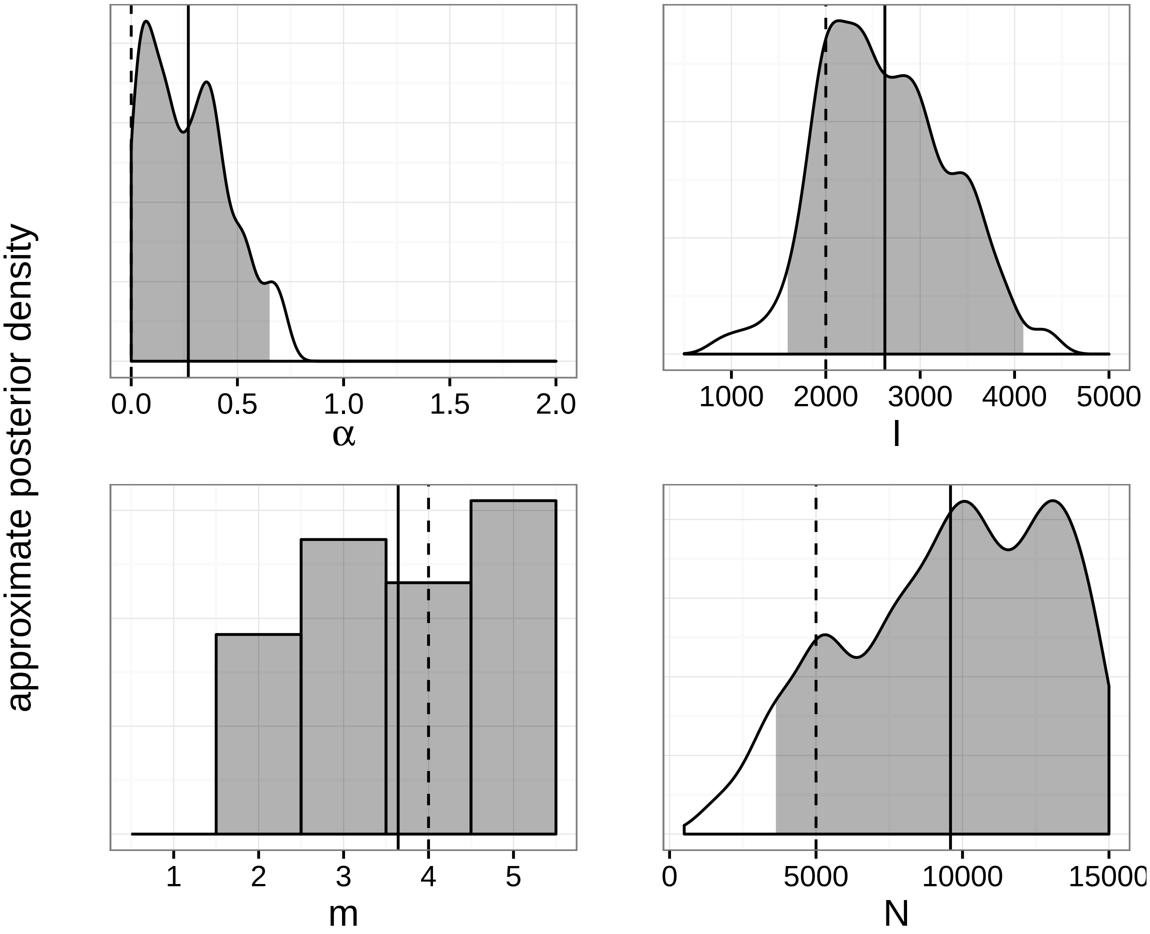

Figure A.56: Approximate marginal posterior distributions of BA model parameters obtained by applying *netabc* to a simulated transmission tree with BA parameter values $\alpha = 0.0$, $I = 2000$, $m = 4$, and $N = 5000$. Vertical dashed lines indicate true values. Shaded areas are 95% highest posterior density intervals. x-axes indicate regions of nonzero prior density.

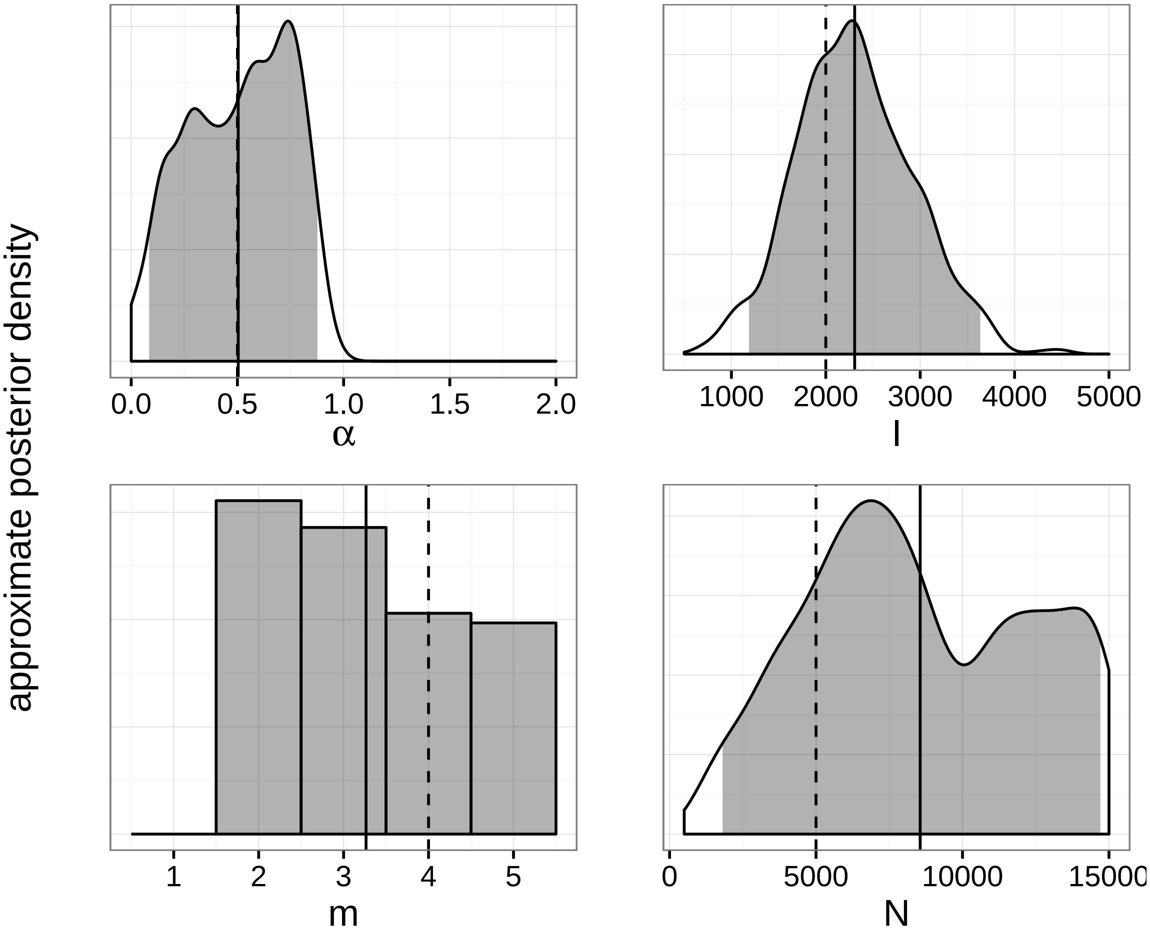

Figure A.57: Approximate marginal posterior distributions of BA model parameters obtained by applying *netabc* to a simulated transmission tree with BA parameter values $\alpha = 0.5$, $I = 2000$, $m = 4$, and $N = 5000$. Vertical dashed lines indicate true values. Shaded areas are 95% highest posterior density intervals. x-axes indicate regions of nonzero prior density.

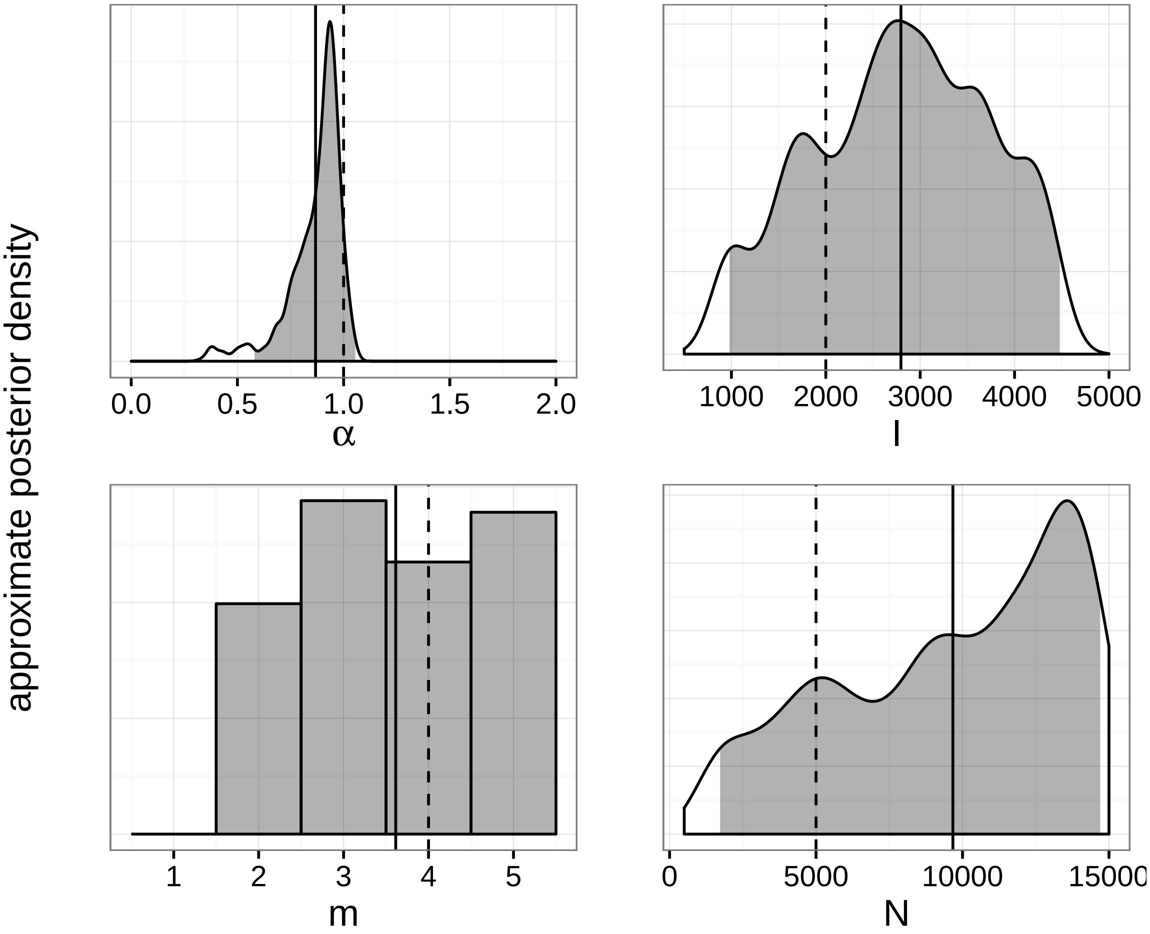

Figure A.58: Approximate marginal posterior distributions of BA model parameters obtained by applying *netabc* to a simulated transmission tree with BA parameter values $\alpha = 1.0$, $I = 2000$, $m = 4$, and $N = 5000$. Vertical dashed lines indicate true values. Shaded areas are 95% highest posterior density intervals. x-axes indicate regions of nonzero prior density.

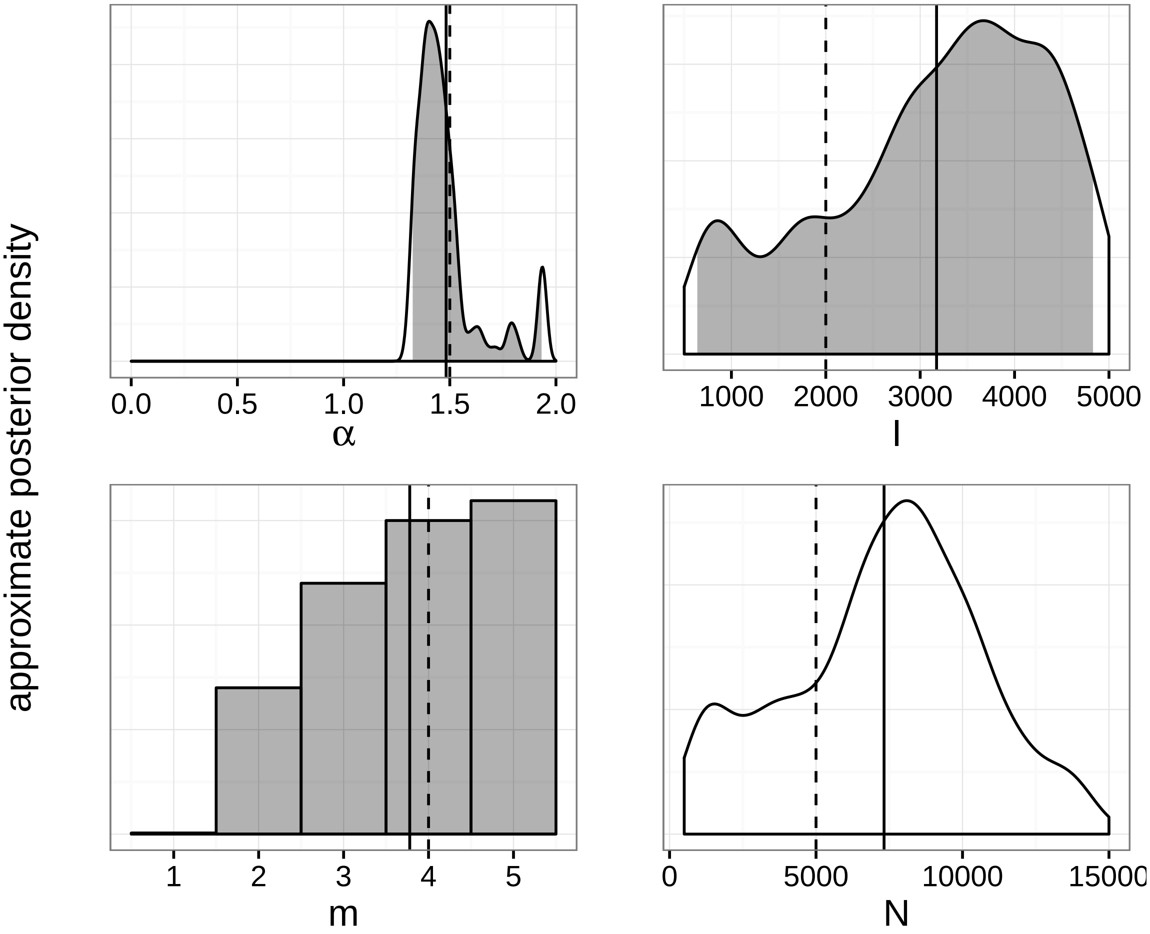

Figure A.59: Approximate marginal posterior distributions of BA model parameters obtained by applying *netabc* to a simulated transmission tree with BA parameter values $\alpha = 1.5$, $I = 2000$, $m = 4$, and $N = 5000$. Vertical dashed lines indicate true values. Shaded areas are 95% highest posterior density intervals. x-axes indicate regions of nonzero prior density.