# Phylogenetic estimation of contact network parameters with kernel approximate Bayesian computation

by

Rosemary Martha McCloskey

B.Sc., Simon Fraser University, 2014

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

Master of Science

in

The Faculty of Graduate and Postdoctoral Studies

Bioinformatics

The University of British Columbia
(Vancouver)

August 2016

# Abstract

# Preface

This work was conducted at the BC Centre for Excellence in HIV/AIDS under the supervision of Dr. Art Poon. I wrote the code, performed the experiments, and wrote the thesis. Dr. Richard Liang assisted with the development of the Gillespie simulation algorithm.

A version of chapters 2 and 3 has been submitted for publication in Molecular Biology and Evolution, with the title "Phylogenetic estimation of contact network parameters with kernel-ABC". A presentation with the same title was given at the 23rd HIV Dynamics and Evolution meeting on April 25, in Woods Hole, Massachussets, USA.

# Contents

# List of Tables

# List of Figures

# List of Symbols

$I$ number of nodes which are eventually infected.

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

$N$ number of nodes in the network.

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

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

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

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

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

# List of Abbreviations

**ABC** approximate Bayesian computation.

**BA** Barabási-Albert.

**ER** Erdős-Rényi.

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

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

**i.i.d.** independent and identically distributed.
**IS** importance sampling.

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

**LTT** lineages-through-time.

**MAP** maximum *a posteriori*.
**MCMC** Markov chain Monte Carlo.
**MH** Metropolis-Hastings.
**ML** maximum likelihood.

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

**ODE** ordinary differential equation.

**PCA** principal components analysis.

**SARS** severe acute respiratory syndrome.
**SI** susceptible-infected.
**SIR** susceptible-infected-recovered.
**SIS** sequential importance sampling.
**SMC** sequential Monte Carlo.
**SVM** support vector machine.

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

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

$\mathcal{G}$ be the set of all possible contact networks, and $\mathcal{N}$ be the set of all possible labellings of the internal nodes of $T$. We can write the likelihood as

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

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

Depending on the network model studied, it is possible that eq. (1.1) could be simplified into a tractable expression. However, a simpler alternative to likelihood-based methods, which would apply to any network model, is provided by ABC. All of the ingredients required to apply ABC to this problem are readily available. Simulating networks is straightforward under a variety of models. Epidemics on those networks, and the corresponding transmission trees, can also be easily simulated. As mentioned above, contact networks can profoundly affect transmission tree shape, and those shapes can be compared using a highly informative similarity measure. SMC has several advantages over other algorithms for ABC [1], including a recently-developed adaptive algorithm requiring minimal tuning on the part of the user [2]. In summary, our method to infer contact network parameters will combine the following: stochastic simulation of epidemics on networks, the tree kernel, and adaptive ABC-SMC. Since our distance measure is a kernel function, our method is a type of kernel-ABC. For ease of exposition, we will often use the term "kernel-ABC" to refer to our method specifically.

Empirical studies of sexual contact networks have found that these networks tend to be scale-free, meaning that their degree distributions follow a power law (although there has been some disagreement). Preferential attachment has been postulated as a mechanism by which scale-free networks could be generated. This makes the BA model, one of the simplest preferential

attachment models, a natural choice to explore with our method. The second aim of this work is to use simulations to investigate the parameters of the BA model, including whether they have a detectable impact on tree shape, and whether they can be accurately recovered using kernel-ABC.

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

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

## 1.2 Phylogenetics and phylodynamics

### 1.2.1 Phylogenetic trees

In evolutionary biology, a *phylogeny*, or *phylogenetic tree*, is a graphical representation of the the evolutionary relationships among a group of organisms or species (generally, *taxa*) [3]. The *tips* of a phylogeny, that is, the nodes without any descendants, correspond to *extant*, or observed, taxa, while the *internal nodes* correspond to their common ancestors. The edges or *branches* of the phylogeny connect ancestors to their descendants. Phylogenies may have a *root*, which is a node with no descendants distinguished as the most recent common ancestor of all the extant taxa [4]. 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* [5].

The branches of the tree may have associated lengths, representing either evolutionary distance or calendar time between ancestors and their descendants. The term "evolutionary distance" is used here imprecisely to mean any sort of quantitative measure of evolution, such as the number of differences between the DNA sequences of an ancestor its descendant, or the difference in average body mass or height. A phylogeny with branch lengths in calendar time units is often referred to as *time-scaled*. In a time-scaled phylogeny, the internal nodes can be mapped onto a timeline by using the tips of the tree, which usually correspond to the present day, as a reference point [6]. 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 [7]. These concepts are illustrated in fig. 1.1.

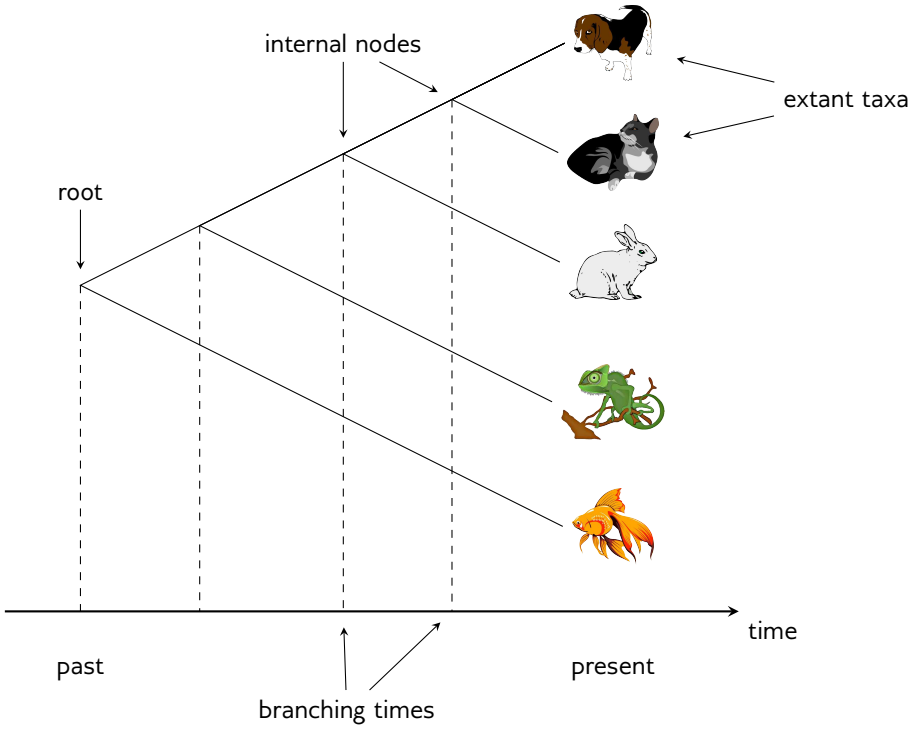



Figure 1.1: Illustration of a rooted, ultrametric, time-scaled phylogeny. The tips of the tree, which represent extant taxa, are placed at the present day on the time axis. Internal nodes, representing extinct common ancestors to the extant taxa, fall in the past. The topology of the tree indicates that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

### 1.2.2 Transmission trees

In epidemiology, a *transmission tree* is a graphical representation of an epidemic's progress through a population. 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 infected hosts, while internal nodes correspond to transmissions from one host to another. Transmission trees generally have branch lengths in units of calendar time, with branching times indicating times of transmission. The root of a transmission tree corresponds to the initially infected patient who introduced the epidemic into the network, also known as the *index case*. The internal nodes may be labelled with the donor of the transmission pair, if this is known. The tips of the tree, rather than being fixed at the present day, are placed at the time at which the individual was removed from the epidemic, such as by death, recovery, isolation, behaviour change, or migration. Consequently, the transmission tree may not be ultrametric, but may have tips located at varying distances from the root. Such trees are said to have *heterochronous* taxa [8], in contrast to the *isochronous* taxa found in most phylogenies of macro-organisms. A transmission tree is illustrated in fig. 1.2 (right).

Due to the internal nodes, each infected individual in an epidemic may appear in the transmission tree more than once. This is different from the transmission *network*, in which each infected individual appears exactly once, and edges are in one-to-one correspondence with transmissions [9, 10]. Transmission networks are discussed further in section 1.3, and the distinction between the two objects is illustrated in fig. 1.2. However, since transmission networks generally have no cycles (unless re-infection occurs), they are trees in the graph theoretical sense, and hence are sometimes also referred to as transmission trees [*e.g.* 11]. In this work, we reserve the term "transmission tree" for the objects depicted on the right side of fig. 1.2, following [*e.g.* 12]. The term "transmission network" is taken to mean the subgraph of the contact network along which transmissions occurred, following [9, 10].

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 [6], 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 [13]. Many more diverse epidemiological parameters have been investigated using transmission trees, such as the degree of clustering [14] and the effect of elevated transmission risk in acute infection [15]. However, in all but the most well-studied of epidemics, this is not possible to obtain through traditional epidemiological methods [9]. The time and effort to conduct detailed interviews and contact

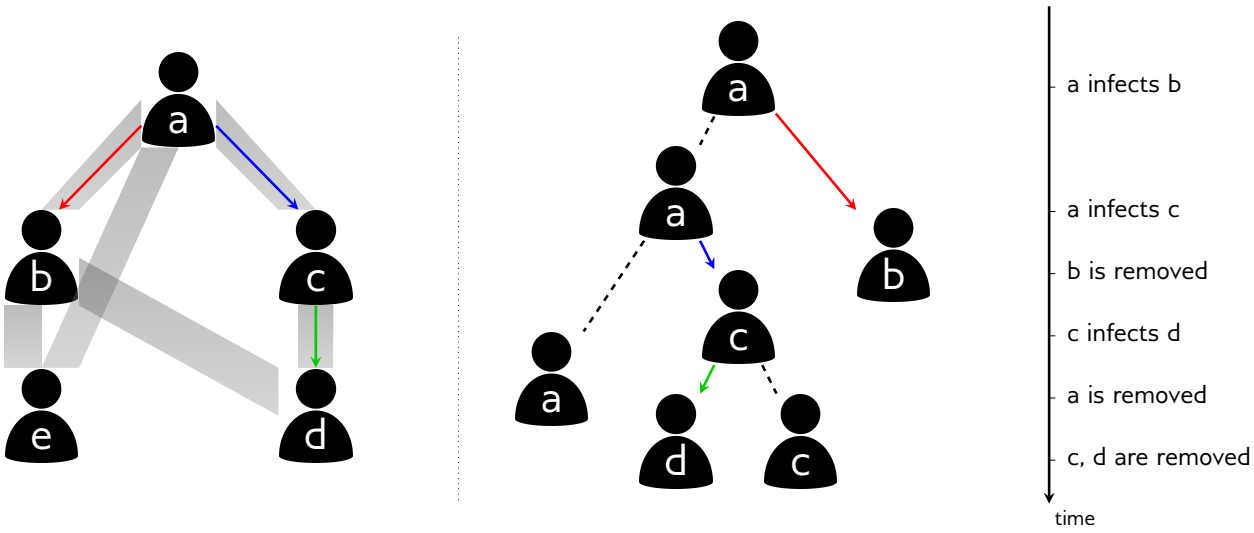



Figure 1.2: Illustration of epidemic spread over a contact network, and the corresponding transmission tree. (Left) A contact network with five hosts, labelled *a* through *e*. Thick shaded edges indicate symmetric contacts among the hosts. The transmission network is indicated by coloured arrows. The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

tracing of a sufficient number of infected individuals is usually prohibitive. Even when the resources for such methods are available, patients may not always recall whom they contacted and when, especially in the case of airborne transmission. Consequently, the transmission tree must be estimated using other methods. Most commonly, this is done by exploiting the relationship between transmission trees and viral phylogenies [16].

### 1.2.3 Relationship between transmission trees and viral phylogenies

In general, *viral phylogenies* are simply phylogenetic trees relating virus strains. In phylodynamics, we often consider *inter-host* phylogenies, which relate one viral genotype from each host in a population. The crux of *phylodynamics* [17] is the fact that the epidemiological processes recorded in transmission trees, and the evolutionary processes recorded in viral phylogenies, occur on similar time scales for RNA viruses [8]. As a result, there is a close relationship between the two types of tree. In particular, the transmission process is quite similar to *allopatric speciation* [18], where genetic divergence follows the geographic isolation of a sub-population of organisms. Thus, transmission, which is represented as branching in the transmission tree, causes branching in the viral phylogeny as well. Similarly, the removal of an individual from

the transmission tree causes the extinction of their viral lineage in the phylogeny. Due to these relationships, the topology of the viral phylogeny is often used as a proxy for the topology of the transmission tree. However, there are several complications and caveats which must be kept in mind when estimating the transmission tree in this manner.

First are the issues of rooting and time-scaling. Modern likelihood-based methods of phylogenetic reconstruction [*e.g.* 19, 20] 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 [21]. It is generally assumed that sampling a virus from the individual also corresponds to their removal from the transmission tree, so the positions of the tips in time are fixed. Therefore, to transform a viral phylogeny into an estimated transmission tree, we must find a root and an assignment of branch lengths such that the tips are placed at their proscribed times. In addition, the branch lengths should be chosen such that the variation in *evolutionary rate*, which is the ratio of a branch's length in genetic distance to its length in calendar time, is low across the tree. While there is some variation among hosts, due to immunological and other factors, we generally expect to observe globally similar evolutionary rates. Methods for time-scaling a phylogeny include root-to-tip regression [22–24], which we apply in this work, and least-square dating [25]. Both of these methods can be used to root the tree, by simply trying all possible root positions and choosing the one which minimizes a loss function ($1 - R^2$, or root-mean-square error, for root-to-tip regression; sum of squared errors for least-square dating). Alternatively, the tree may be rooted separately with an outgroup [26] before time-scaling.

A second, perhaps more insidious problem is the fact that the correspondence between the topologies of the viral phylogeny and transmission tree is not necessarily exact. Due to intra-host diversity, the viral strain which is transmitted may have split from another lineage within the donor long before the transmission event occurred. Hence, the branching point in the viral phylogeny may be much earlier than that in the transmission tree. Another possibility is that one host transmitted to two or more recipients, but the lineages they each received originated within the donor host in a different order than that in which the transmissions occurred. In this case, the topology of the transmission tree and the viral phylogeny will be mismatched. Although phylodynamics is quite new, these phenomena have been studied in evolutionary biology for some time. Viral phylogenies are a specific version of a more general class of trees called *gene trees*, which represent the evolutionary history of a section of genetic material. Transmission trees, on the other hand, are highly analogous to *species trees*, whose tips are species and internal nodes are common ancestors. This analogy derives from the functional similarity between transmission

and allopatric speciation. Hence, the potential discordance between transmission trees and viral phylogenies is the similar to that between gene and species trees, which is called *incomplete lineage sorting*. In practice, this discordance has not proven an insurmountable problem: for example, Leitner et al. [27] were able to accurately recover a known transmission tree using a viral phylogeny.

A final caveat is that the viral phylogeny itself is not known with certainty, so it must also be estimated from genetic data. Phylogenetic inference is a complex topic which we will not discuss in detail here (see e.g. [28] for a full review). Most modern analyses use model-based methods, which simultaneously estimate the phylogeny with branch lengths and the parameters of a model of evolution. Although they usually work well in practice, the estimated topology can vary based on the model used and, in the case of Bayesian analysis, the priors. In addition, intra-host viral populations are genetically heterogeneous, so choosing a single representative genotype per host is necessarily imprecise. One can use either the genotype of a specific virion sampled from the host, or a synthetic genotype, such as a consensus or reconstructed ancestral sequence.

What we have just discussed is a two-step procedure for estimating the transmission tree. First, a viral phylogeny is constructed from genetic sequence data, and then it is rooted and time-scaled into a transmission tree. This approach is straightforward, frequently used, and has the advantage of leveraging tried-and-true tools for phylogenetic inference. However, it also has drawbacks, perhaps the most obvious being the multiplication of errors produced by the separate steps. One commonly used alternative method is to directly estimate a time-scaled phylogeny by simultaneously inferring the tree topology, its root and branch lengths, and the parameters of a *molecular clock* model. A molecular clock is a hypothesis about the evolutionary rates along the branches of the tree, such as that they are all equal (a *strict clock*) or that they are independent and identically distributed (i.i.d.) from a common distribution (a *relaxed clock*). This inference is usually done in a Bayesian framework using Markov chain Monte Carlo (MCMC), so that prior information (including the tip dates) can be included in the analysis, and the so-called nuisance parameters of the molecular clock model can be marginalized out. Software packages for performing these analyses include BEAST [29] and MrBayes [30].

Several other authors have developed methods tailor-made for inferring transmission trees. Didelot, Gardy, and Colijn [31] develop a Bayesian version of the two-step approach which allows transmissions to occur anywhere along the branches of a transmission tree, rather than being constrained to the branching points in the viral phylogeny. The method requires sampling of every infected individual, although the authors indicate that it could be extended to

relax this assumption. Cottam et al. [32] describe a likelihood-based method which enumerates all transmission trees consistent with an established phylogeny, assigning each a likelihood based on other epidemiological data. This approach is novel in its integration of data from multiple sources, however because it enumerates a large portion of the tree space, it is unlikely to scale to larger epidemics. Ypma et al. [33] develop a joint likelihood function integrating temporal, geographic, and genetic observations, and use Bayesian MCMC to estimate both the tree and the parameters of the likelihood function. Their approach can handle missing data and produces high resolution transmission trees when multiple types of data are available. A different approach is undertaken by Jombart et al. [34], who describe a method to build transmission trees directly from sequence data, contingent on the common ancestors also being sampled. This makes the method attractive for slow-evolving pathogens, but less practical for viral outbreaks where samples from common ancestors are unlikely to be available.

### 1.2.4 Tree shapes

The aim of viral phylodynamics is to glean some kind of knowledge, about the epidemic, the virus, or its hosts and their behaviour, by studying a phylogeny, most often an estimated transmission tree [16, 21]. 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 [35]. Methods of quantifying tree shape fall into two categories: summary statistics, and pairwise measures.

Summary statistics assign a numeric value to each individual tree. One of the most widely used is Sackin's index [36], which measures the imbalance or asymmetry in a rooted tree. For the $i$th tip of the tree, we define $N_i$ to be the number of branches between that tip and the root. The unnormalized Sackin's index is defined as the sum of all $N_i$. It is called unnormalized because it does not account for the number of tips in the tree. Among two trees having the same number of tips, the least-balanced tree will have the highest Sackin's index. However, among two equally balanced trees, the larger tree will have a higher Sackin's index. This makes it challenging to compare balances among trees of different sizes. To correct this, Kirkpatrick and Slatkin [37] derive the expected value of Sackin' index under the Yule model [38]. Dividing by this expected value normalizes Sackin's index, so that it can be used to compare trees of different sizes.

Rather than assigning numbers to individual trees, pairwise measures associate a numeric value to each *pair* of trees, indicating how different the trees are from each other. Distance measures allow us to identify groups of related phylogenies, for example, local epidemics which are undergoing a similar pattern of expansion. One such distance measure is the normalized

lineages-through-time (nLTT) [39], which compares the LTT [6] 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 difference between the two plots is integrated between 0 and 1. In the context of infectious diseases, the LTT is related to the prevalence [13], so large values may indicate that the trees being compared are the products of different epidemic trajectories [39].

Another tree distance measure is the *phylogenetic kernel*, or "tree kernel" developed by Poon et al. [40]. As opposed to the nLTT, the tree kernel is maximized when the two trees being compared are the same. The basis of the tree kernel is the kernel trick originally developed for support vector machines (SVMs) [41]. The idea of the kernel trick is to compare objects by mapping them into a feature space of very high, possibly even infinite, dimension. The similarity between objects is taken to be their dot product in the feature space. It is called a "trick" because this dot product is computed using a *kernel function* without explicitly mapping the objects to the feature space, which would be computationally prohibitive. In the case of the tree kernel, the feature space is the space of all possible *subset trees*, which are subtrees that do not necessarily extend all the way to the tips. The subset-tree kernel was originally developed for comparing parse trees in natural language processing [42] and did not incorporate branch length information. The version developed by Poon et al. [40] 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 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 [43]. In that paper, Poon used the tree kernel as the distance function in ABC (see section 1.4), to fit epidemiological models to observed trees. This is an example of kernel-ABC [44], which will be discussed further in section 1.4.

### 1.2.5 Applications of phylodynamics

Phylodynamic methods have been used to investigate epidemiological parameters such as transmission rate, recovery rate, and basic reproductive number [16, 21]. These studies make inferences about epidemiological processes from the genetic diversity of virus populations, which is usually represented in the form of a phylogeny. The majority of these employ a Bayesian MCMC approach to infer parameters of an epidemiological model whose likelihood can be calculated, most often some variation of the birth-death [45] or coalescent [46] models. Stadler et al. [47] develop a formula for the likelihood of a phylogeny with heterochronous tips under the birth-death model, which has been used to estimate the basic reproductive number of several viral epidemics [47]. However, the birth-death model is cannot tell us anything about popula-

tion structure, as it assumes that every individual becomes infected at the same rate. Volz [48] writes down the likelihood of a heterochronous phylogeny under a coalescent model with arbitrarily complex population dynamics. This opens the door to more complex inferences about population structure, as the population can be partitioned into compartments with different transmission and recovery rates, but still assumes that each compartment is homogeneously mixed. In other words, the coalescent model can tell us about the *global* structure of a population, such as whether there exists a high-risk subgroup, but not about the *local* structure, such as the average number of contacts each individual has.

## 1.3 Contact networks

### 1.3.1 Overview

Epidemics spread through populations of hosts through *contacts* between those hosts. The definition of contact depends on the mode of transmission of the pathogen in question. For an airborne pathogen like influenza, a contact may be simple physical proximity, while for HIV, contact could be via unprotected sexual relations or blood-to-blood contact (such as through needle sharing). A *contact network* is a graphical representation of a host population and the contacts among its members [10, 49, 50]. The *nodes* in the network represent hosts, and *edges* or *links* represent contacts between them. A contact network is shown in fig. 1.2 (left). Contact networks are a particular type of *social network* [51, 52], which is a network in which edges may represent any kind of social or economic relationship. Social networks are frequently used in the social sciences to study phenomena where relationships between people or entities are important [for a review see 53].

Edges in a contact networks may be *directed*, representing one-way transmission risk, or *undirected*, representing symmetric transmission risk. For example, a network for an airborne epidemic would use undirected edges, because the same physical proximity is required for a host to infect or to become infected. However, an infection which may be spread through blood-to-blood contact through transfusions transfusions would use directed edges, since the donor has no chance of transmitting to the recipient. Directed edges are also useful when the transmission risk is not equal between the hosts, such as with HIV transmission, where acting as the receptive partner carries a higher risk of infection than acting as the insertive partner. In this case, a contact could be represented by two directed edges, one in each direction between the two hosts, with the edges annotated by what kind of risk they imply [53]. An undirected contact network is equivalent to a directed network where each contact is represented by two symmetric directed

edges. The *degree* of a node in the network is how many contacts it has. In directed networks, we may make the distinction between *out-degree* and *in-degree*, which count respectively the number incoming and outgoing edges. The *degree distribution* of a network denotes the probability that a node has any given number of links. The set of edges attached to a node are referred to as its *incident* edges.

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

From a public health perspective, knowledge of contact networks has the potential to be extremely useful. On a population level, network structure can dramatically affect the speed and pattern of epidemic spread [*e.g.* 57, 58]. For example, epidemics are expected to spread more rapidly in networks having the "small world" property, where the average path length between two nodes in the network is relatively low [59]. Some sexually transmitted infections would not be expected to survive in a homogeneously mixed population, but their long-term persistence can be explained by contact heterogeneity [56, 60]. Hence, the contact network can provide an idea of what to expect as an epidemic unfolds. In terms of actionable information, vaccination strategies which would eradicate an epidemic in a random network might not work if the network is scale-free [10, see section 1.3.2]. On a local level, contact networks can be informative about the groups or individuals who are at highest risk of acquiring or transmitting infection, and would therefore benefit most from public health interventions [61, 62].

Contact networks are a challenging type of data to collect, requiring extensive epidemiological investigation in the form of contact tracing [9, 10, 50]. Therefore, it has been necessary to explore less resource-intensive alternatives which still contain information about population structure. For instance, it is possible to obtain limited information about the contact network by individual interviews without contact tracing. Variables which can be estimated in this fashion are referred to as *node-level* measures [53]. One of the most well-studied of these is the degree distribution, which can be estimated by simply asking each person how many contacts they had in some interval of time [63–65].

An alternative approach has been the analysis of other networks, which can be estimated with phylogenetic methods from viral sequence data. Some work focuses on the *phylogenetic network*, in which two nodes are connected if the genetic distance between their viral sequences is below some threshold. Primarily, this work has focused on the detection of *phylogenetic clusters*, which are groups of individuals whose viral sequences are significantly more similar to each other's than to the general population's. The phylogenetic network is informative about "hotspots" of transmission and can be used to identify demographic groups to whom targeted interventions are likely to have the greatest effect [66]. However, this network may show little to no agreement with a contact data obtained through epidemiological methods [67–69], and therefore may be a poor proxy for the contact network. Other studies [70] have investigated the *transmission network*, which is the subgraph of the contact network consisting of infected nodes and the edges which led to their infections [9] (fig. 1.2, left). It is possible to estimate the transmission network phylogenetically, although the methods required for doing so are more sophisticated than for estimating the phylogenetic network [70]. These studies again mostly focusing on clustering, and also on degree distributions.

Other statistical methods have been developed to infer contact network parameters strictly from the timeline of an epidemic, using neither genetic data nor reported contacts. Britton and O'Neill [71] developed a Bayesian method to infer the $p$ parameter of an Erdős-Rényi (ER) network, along with the transmission and removal rate parameters of the susceptible-infected (SI) model, using observed infection and optionally removal times. However, it was designed for only a small number of observations, and was unable to estimate $p$ independently from the transmission rate. Groendyke, Welch, and Hunter [72] significantly updated and extended the methodology of Britton and O'Neill, and applied it to a measles outbreak affecting 188 individuals. They were able to obtain a much more informative estimate of $p$, although this data set included both symptom onset and recovery times for all individuals, and was unusual in that the entire contact network was presumed to be infected. Volz [58] developed differential

equations describing the dynamics of the SIR model on a wide variety of random networks defined by their degree distributions. Although the topic of estimation was not addressed in the original paper, Volz's method could in principle be used to fit such models to observed epidemic trajectories, similar to what is done with the ordinary SIR model. Volz and Meyers [55] later extended the method to dynamic contact networks and applied it to a sexual network relating 99 individuals investigated during a syphilis outbreak.

### 1.3.2 Scale-free networks and preferential attachment

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

Several properties of scale-free networks are relevant in epidemiology. The high-degree hub nodes are known as *superspreaders* [75], which have been postulated to contribute in varying degree to the spread of diseases such as HIV [12] and severe acute respiratory syndrome (SARS) [76]. Scale-free networks have no epidemic threshold [60], meaning that diseases with arbitrarily low transmissibility can persist at low levels indefinitely. This is in contrast with homogeneously mixed populations, in which transmissibility below the epidemic threshold would result in exponential decay in the number of infected individuals and eventual extinction of the pathogen (Anderson & May, I think).

One mechanism which has been shown to lead to scale-free networks is *preferential attachment* [73, 77]. Under this process, 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 [73], the partners of the new node are chosen with probability linearly proportional to their degree. However, Barabási and Albert suggest extending the model such that the probability of choosing a partner of degree $d$ is proportional to $d^{\alpha}$ for some constant $\alpha$, and we use this extension here.

There has been some contention of the idea that contact networks are scale-free. Handcock and Jones [78] fit several stochastic models of partner formation to empirical degree distributions derived from population surveys of sexual behaviour. They found that a negative binomial

distribution, rather than a power law, was the best fit to five out of six datasets, although the difference in goodness of fit was extremely small in four out of these five. Bansal, Grenfell, and Meyers [54] found that an exponential distribution, rather than a power law, was the best fit to degree distributions of six social or sexual networks.

### 1.3.3 Relationship between network structure and transmission trees

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

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

In summary, studies in this group have demonstrated that network structure profoundly influences tree shape, but have not attempted to quantitatively infer network parameters from

observed trees.

## 1.4 Approximate Bayesian computation

### 1.4.1 Model fitting

A *mathematical model* is a formal description of a hypothesized relationship between some observed data, $\mathbf{x} = \{x_1, \ldots, x_n\}$, and outcomes, $\mathbf{y} = \{y_1, \ldots, y_n\}$. A *parametric* model defines a family of possible relationships between data and outcomes, indexed by one or more numeric parameters $\theta$. A *statistical* model describes the relationship between data and outcomes in terms of probabilities. Statistical models define, either explicitly or implicitly, the probability of observing $\mathbf{y}$ given $\mathbf{x}$ and, if the model is parametric, $\theta$. In this context, the observed outcomes are taken to be realizations of random variables $\mathbf{Y} = \{Y_1, \ldots, Y_n\}$. Note that it is entirely possible to have no data $\mathbf{x}$, only observed outcomes $\mathbf{y}$. In this case, a model would describe the process by which $\mathbf{y}$ 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 each $x_i$ and $y_i$ is a real number. The linear model postulates that the outcomes are linearly related to the data, modulo some noise introduced by measurement error, environmental fluctuations, and other external factors. Formally, $y_i = \beta x_i + \varepsilon_i$, where $\beta$ is the slope of the linear relationship, and $\varepsilon_i$ is the error associated with measurement $i$. We can make this model a statistical one by hypothesizing a distribution for the error terms $\varepsilon_i$; most commonly, it is assumed that they are normally distributed with variance $\sigma$. In mathematical terms, $Y_i \sim \beta x_i + \mathcal{N}(0, \sigma^2)$, where "$\sim$" means "is distributed as". We can see from this formulation that the model is parametric, with parameters $\beta$ and $\sigma$. Moreover, we can write down the probability density of observing outcome $y_i$ given the parameters,

$$\Pr(Y_i = y_i \mid \beta, \sigma) = \Pr\left(\mathcal{N}(0, \sigma^2) = y_i - \beta x_i\right).$$

Note that we have followed the standard statistical abuse of notation and used "Pr" to refer to a probability *density*, rather than a probability. Assuming all the $y_i$ are independent, the probability density of the entire observed set of outcomes is the product of the probability density of each individual $y_i$,

$$\Pr(\mathbf{Y} = \mathbf{y} \mid \beta, \sigma) = \prod_{i=1}^{N} \Pr(Y_i = y_i \mid \beta, \sigma).$$

For a general model, the probability density of $\mathbf{y}$ given the parameters $\theta$ is also known as the *likelihood*, written $\mathcal{L}$, of $\theta$. That is, $\mathcal{L}(\theta \mid \mathbf{y}) = \Pr(\mathbf{y} \mid \theta)$. The higher the value of the likelihood, the more likely the observations $\mathbf{y}$ are under the model. Thus, the likelihood provides a natural criterion for fitting the model parameters: we want to pick $\theta$ such that the probability density of our observed outcomes $\mathbf{y}$ is as high as possible. The parameters which optimize the likelihood are known as the *ML* estimates, denoted $\hat{\theta}$. ML estimation is usually performed with numerical optimization. In the simplest terms, many possible values for $\theta$ are examined, $\mathcal{L}(\theta \mid \mathbf{y})$ is calculated for each, and the parameters which produce the highest value are accepted. Many sophisticated numerical optimization methods exist, although they may not be guaranteed to find the true ML estimates if the likelihood function is complex. Occasionally, as in the case of least squares, the ML estimates can be found explicitly by setting the likelihood function's derivatives to zero.

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

$$\Pr(\theta \mid \mathbf{y}) = \frac{\Pr(\mathbf{y} \mid \theta) \Pr(\theta)}{\Pr(\mathbf{y})}.$$

In principle, $\Pr(\mathbf{y} \mid \theta) \Pr(\theta)$ can be optimized numerically just like $\mathcal{L}(\theta \mid \mathbf{y})$, which would also optimize the posterior distribution. The resulting optimal parameters are called the maximum *a posteriori* (MAP) estimates. However, from a Bayesian perspective, $\theta$ is not a fixed quantity to be estimated, but rather a random variable with an associated distribution (the posterior). Therefore, the MAP estimate by itself is of limited value without associated statistics about the posterior distribution, such as the mean or credible intervals. Unfortunately, to calculate such statistics, it is necessary to evaluate the normalizing constant $\Pr(\mathbf{y})$, 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 $\theta_1$ to state $\theta_2$ is taken to be

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

where $q(\theta, \theta')$ is a symmetric *proposal distribution* used in the algorithm to generate the chain. The stationary distribution of this Markov chain is equal to the posterior distribution on $\theta$. Therefore, if a long enough random walk is performed on the chain, the distribution of states visited will be a Monte Carlo approximation of $\Pr(\theta \mid \mathbf{y})$, from which we can calculate statistics of interest. Actually performing this random walk is straightforward and is accomplished via the Metropolis-Hastings algorithm (algorithm 1).

**Algorithm 1** Metropolis-Hastings algorithm for Markov chain Monte Carlo.

Draw $\theta$ according to $\Pr(\theta)$
**loop**
  Propose $\theta'$ according to $q(\theta, \theta')$
  Accept $\theta \leftarrow \theta'$ with probability $\max\left(1, \frac{\Pr(\mathbf{y} \mid \theta')\Pr(\theta')q(\theta', \theta)}{\Pr(\mathbf{y} \mid \theta\ )\Pr(\theta\ )q(\theta, \theta')}\right)$
**end loop**

### 1.4.2 Overview and motivation for ABC

Most mathematical models are amenable to fitting via one or both of the approaches, ML or Bayesian inference, discussed above. However, there are some, particularly in the domain of population genetics, for which calculation of either the likelihood or the product of the likelihood and the prior may be infeasible. For example, one or both of these quantities may be expressible only as an intractable integral. Approximate Bayesian computation (ABC) is designed for such cases, where standard likelihood-based techniques for model fitting cannot be applied. Such models are particularly prevalent in population genetics [86, 87]. We will begin this subsection by describing ABC in general terms. Then, we will use linear regression as a toy problem to demonstrate how ABC could be applied. Finally, we will give an example of a model which is impractical to fit using likelihood-based methods, and show how its parameters were estimated using ABC.

Ordinarily, Bayesian inference targets the posterior distribution.

Of course, linear models can be fit more easily using likelihood-based methods, since the likelihood $\mathcal{L}(\theta \mid \mathbf{y})$ can be easily calculated (see **??**). One model where this is not the case is Kingman's coalescent [46].

As an example of such a model, we will review the first practical application of ABC by Tavaré et al. [88]. The aim of the investigation was to calculate the posterior distribution on the coalescence time $t$ of some observed sequences $D$ under the coalescent model [46] with the infinite-sites assumption [89]. Their estimation method relied on two key observations. First, the number $k$ of *segregating sites* in the data is informative about the coalescence time while being much easier to work with than the data itself. A segregating site is a position which is polymorphic in the observed sequences. Using this observation, the authors replaced their posterior distribution of interest, $\Pr(t \mid D)$, by

$$\Pr(t \mid k) \propto \Pr(t) \Pr(k \mid t),$$

where again we abuse Pr to refer to a probability density. Second, the authors observed that the expected number of segregating sites depends on the total branch length $l$ of the tree, rather than its height. This enabled them to write an alternative expression for the product of the likelihood and prior in terms of $l$,

$$\Pr(t) \Pr(k \mid t) = \int_0^\infty \Pr(t, l) \Pr(k \mid l) \mathrm{d}\, l.$$

While this integral may be difficult to evaluate analytically, the power of this formulation is the fact that it is straightforward to *simulate* values of $t$ and $l$ according to $\Pr(t, l)$, their joint probability under the coalescent model. One simply runs the (stochastic) coalescent process and observing the values $\hat{t}$ and $\hat{l}$ which result. It is straightforward to calculate $\Pr(k \mid \hat{l})$, the probability that $k$ segregating sites were observed in a tree with total branch length $\hat{l}$, using the Poisson distribution. Hence if $\hat{l}$ and $\hat{t}$, having been generated according to $\Pr(t, l)$, are sampled with probability proportional to $\Pr(k \mid \hat{l})$, they represent an unbiased sample from the above integral. Taking a large number of such samples allowed the authors to obtain a Monte Carlo estimate of $\Pr(t \mid k)$.

### 1.4.3 Algorithms for ABC

ABC does not refer to a particular procedure for model fitting. Rather, ABC refers to the general strategy of choosing model parameters based on the resulting model's propensity to generate data resembling the real data. There are three main classes of ABC algorithm which have been developed so far: rejection, MCMC, and SMC.

All of these approaches require some common elements. First, as with all Bayesian methods,

we are required to specify a *prior distribution*, denoted $\pi$, on the parameter space. The prior specifies what we already know or believe about the model parameters. Second, in order to compare simulated to observed data, we need to be able to summarize a data set in a numerical format. This is accomplished by a function, denoted $\eta$, which computes a vector of hopefully informative summary statistics on a data set. Third, we need a distance function $\rho$ which tells us how similar two data sets are to each other, based on their summary statistics.

Continuing with the linear model example, we need to specify a prior $\pi(a, b, \sigma)$ on the three parameters. We do not have much certain information about these parameters except that $\sigma$ has to be at least zero, but it seems reasonable to assume that extreme relationships are fairly rare, and that positive and negative correlations are equiprobable. Therefore, we will let $a$, $b$, and $\sigma$ be independent, $a$ and $b$ be normally distributed, and $\sigma$ be log-normally distributed. That is,

$$\pi(a, b, \sigma) = \begin{cases} \Pr[\mathcal{N}(0,1) = a] \cdot \Pr[\mathcal{N}(0,1) = b] \cdot \Pr[\mathcal{N}(0,1) = \log \sigma] & \sigma > 0 \\ 0 & \sigma \leq 0. \end{cases}$$

For the vector of summary statistics, we will use the mean and variance of the simulated data,

$$\eta(\mathbf{y}) = \langle \mathrm{E}[\mathbf{y}], \mathrm{Var}[\mathbf{y}] \rangle .$$

Finally, for the distance function $\rho$ we take the standard Euclidean distance.

Rejection ABC is the simplest method, and also the one which was first proposed [90]. Effectively, it comes down to the approach described in the previous subsection of guessing parameter values until one is close enough to the truth. More specifically, a possible set of parameters $\theta$ is drawn from the prior distribution, and a simulated data set $z$ is generated from the model with those parameters. If the distance between the simulated data set and the real data, $\rho(\eta(y), \eta(z))$, is small enough, then we accept $\theta$ as a sample from the posterior. This can be repeated until as many samples as desired are obtained.

The second method is ABC-MCMC. This is similar to ordinary Bayesian MCMC, except that a ratio of distances to the observed data replaces the likelihood ratio. The algorithm begins by sampling a single vector of parameter values $\theta$ from the prior distribution $\pi(\theta)$. It then proceeds iteratively: a new parameter vector $\theta^*$ is chosen according to a proposal distribution $q(\theta^* \mid \theta)$. The proposal distribution $q$ is often taken to be a Gaussian centred at $\theta$. Then $\theta^*$ is accepted as the new $\theta$ with probability

$$\max(1, \mathrm{TODO}) ,$$

or discarded otherwise. This process is iterated until some stopping criterion is reached, typically a simple limit on the number of steps. After some initial number of iterations, known as *burn-in*, parameters $\theta$ are routinely sampled. Since points in parameter space are visited in proportion to their posterior probability, these samples can be taken to approximate the posterior distribution on $\theta$, and can be used to calculate point estimates and confidence intervals.

The most recently developed class of algorithms for ABC is sequential Monte-Carlo (SMC) [91]. As with the other two classes, we want to eventually obtain a sample from the posterior distribution $f(\theta \mid D)$. The idea of SMC is to begin with a sample from a distribution we know, most often the prior, and approach the posterior smoothly by progressing through a series of intermediate distributions.

## 1.5 Sequential Monte Carlo

SMC is a statistical inference method which samples from a sequence of probability distributions in a fixed order [92]. This idea of *sequential sampling* is useful several contexts, but we consider here its application to model fitting in a Bayesian setting.

`TODO: move the basics of model fitting (likelihoods and priors) to here`

Our objective is to obtain samples from, and summary statistics of, the posterior distribution on the model of interest's parameters. SMC can be applied in this context by defining a sequence of distributions, starting from the posterior and ending at the prior, which constitute a "smooth" trajectory. The main SMC algorithm for this problem was developed by Del Moral, Doucet, and Jasra [92], based on an existing methodology called sequential importance sampling (SIS). SIS is also designed to sample from a sequence of distributions, but rather than all being defined on the same space (such as model parameters), they are defined on nested spaces of increasing dimension. Following Del Moral, Doucet, and Jasra [92] and other reviews of SMC [93], we will begin this section by describing SIS, and then turn to the adaptation of the method to the case when the sequence of distributions are all defined on the same space. Note that Del Moral, Doucet, and Jasra [92] use the variable $\pi$ for the target distributions, but Doucet, De Freitas, and Gordon [93] use $\pi$ to indicate the importance distributions. We follow the former and denote target distributions by $\pi$ and importance distributions by $\eta$.

### 1.5.1 Sequential importance sampling

The basis of SIS is importance sampling (IS), which is a method of estimating summary statistics of distributions which are known only up to a normalizing constant, and therefore cannot be

sampled from directly. That is, if $\pi$ is such a distribution and $f$ is any real-valued function, IS is concerned with estimating

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

where the integral is over the space on which $\pi$ is defined, $\gamma(x)$ is known, and $Z$ is the unknown normalizing constant, $Z = \int \gamma(x)\mathrm{d}\,x$. The posterior distributions of all but the simplest models' parameters fall into this category. Suppose we have at hand another distribution $\eta$, called the *importance distribution*, from which we are able to sample. Define the *importance weight* as the ratio ratio $w(x) = \gamma(x)/\eta(x)$. We can express the normalizing constant $Z$ in terms of the importance weight and distribution, $Z = \int w(x)\eta(x)\mathrm{d}\,x$, and in turn write the original integral of interest as

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

If we sample a large number of points from $\eta$, then $\eta(x)$ can be approximated at each of them by an empirical distribution. Since the remaining quantities $f$, $\gamma$, and $w$ can all be evaluated pointwise, these are all the ingredients we need to obtain an estimate of $\pi(f)$. Although this is a simple and elegant approach, the drawback is that the variance of the estimate is proportional to the variance of the importance weights [94], which may be quite large if $\eta$ and $\gamma$ are very different. Therefore, the practical use of IS on its own is limited, since it depends on finding an importance distribution which is similar to $\pi$, which we usually know very little about *a priory*.

However, an ideal context for the use of IS is when we want to sample from a sequence of nested probability distributions $\pi_1, \ldots, \pi_n$. By *nested*, we mean that $\pi_{i+1}$ is defined on a space of dimension $i + 1$ and admits $\pi_i$ as a marginal. That is,

$$\pi_{i+1}(x_1, \ldots, x_{i+1}) = \pi_i(x_1, \ldots, x_i)f_{i+1}(x_1, \ldots, x_{i+1}),$$

where $f_{i+1}$ is a function which yields a distribution when multiplied with $\pi_i$. This situation may seem somewhat contrived, but it arises naturally when trying to infer the hidden sequence of parameters of a stateful model. For example, Doucet, De Freitas, and Gordon [93] discuss the case when $\pi_i$ is the posterior distribution over the first $i - 1$ states of a hidden Markov model (HMM), conditional on the observed data up to time $i - i$. We assume that either $\pi_1$ is known explicitly, or we have enough information about it to find an adequate importance distribution. In the HMM example, $\pi_1$ was taken to be the prior distribution on the starting state.

# Chapter 2

# Body of Thesis

## 2.1 Methods

### 2.1.1 Kernel-ABC method

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

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

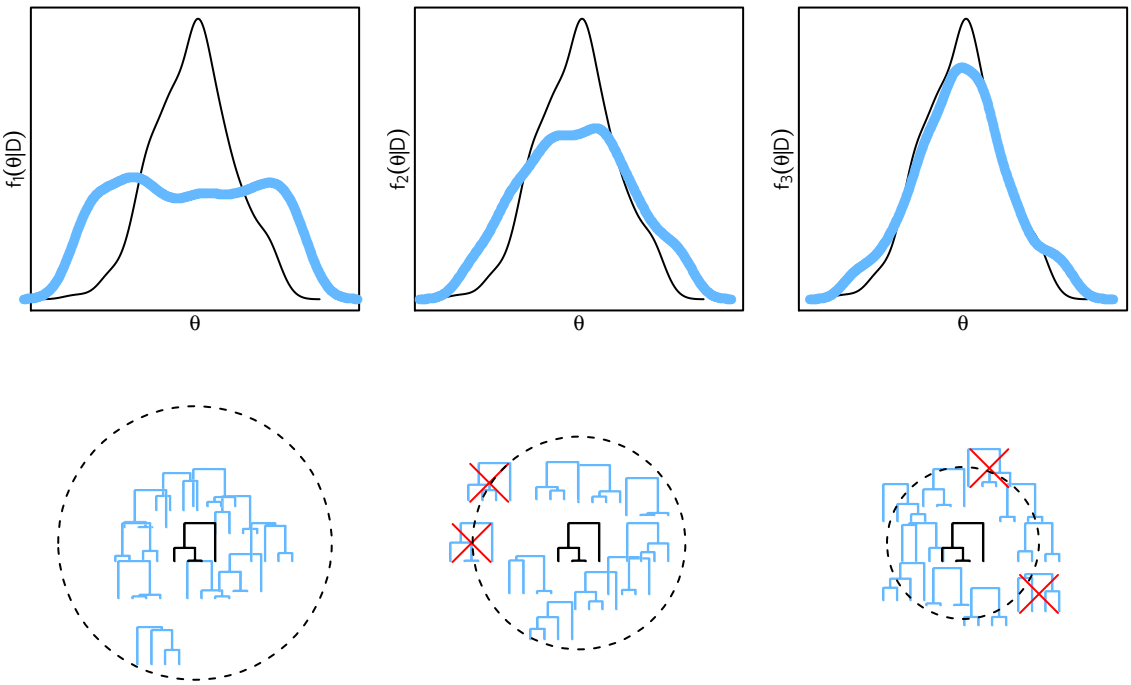



Figure 2.1: Graphical schematic of the ABC-SMC algorithm implemented in *netabc*. Particles are initially drawn from their prior distributions, making the initial population a Monte Carlo approximation to the prior. At each iteration, particles are perturbed, and a distance threshold around the true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

A graphical schematic of this algorithm is given in fig. 2.1.

The generation of contact networks from the model of interest, though critical to the program's function, is not included here as a component of *netabc*. Generation of networks is a model-specific task, whereas *netabc* is intended to be generic and work with any network model. For ease of use, generators for three basic network models, namely the BA, ER, and Watts-Strogatz (WS) models, are included in the program through the *igraph* library [95]. To investigate other models, additional graph generators will need to be added into the program.

*Netabc* is written in the *C* programming language. The *igraph* library [95] is used to generate and store contact networks and phylogenies. Judy arrays [96] are used for hash tables and dynamic programming matrices. The GNU scientific library (GSL) [97] is used to generate random draws from probability distributions, and to perform the bisection step in the adaptive ABC-SMC algorithm. Parallelization is implemented with POSIX threads [98]. In addition to the *netabc* binary to perform kernel-ABC, we provide three additional stand-alone utilities:

*treekernel*, to calculate the phylogenetic kernel; *nettree*, to simulate a transmission tree over a contact network; and *treestat*, to compute various summary statistics of phylogenies. The programs are freely available at `https://github.com/rmcclosk/netabc`.

### Epidemic simulation

The simulation of epidemics, and the corresponding transmission, trees over contact networks is performed in *netabc* using the Gillespie simulation algorithm [99]. This method has been independently implemented and applied by several authors [*e.g.* 69, 72, 80, 83, 85]. Groendyke, Welch, and Hunter [72] published their implementation as an *R* package, but since the SMC algorithm is quite computationally intensive, we chose to implement our own version in *C*.

Let $G = (V, E)$ be a directed contact network. We assume the individual nodes and edges of $G$ follow the dynamics of the SIR model [100]. Each directed edge $e = (u, v)$ in the network is associated with a transmission rate $\beta_e$, which indicates that, once $u$ becomes infected, the waiting time until $u$ infects $v$ is distributed as Exponential($\beta_e$). Note that $v$ may become infected before this time has elapsed, if $v$ has other incoming edges. $v$ also has a removal rate $\gamma_v$, so that the waiting time until removal of $v$ from the population is Exponential($\gamma_v$). Removal may correspond to death or recovery with immunity, or a combination of both, but in our implementation recovered nodes never re-enter the susceptible population. We define a *discordant edge* as an edge $(u, v)$ where $u$ is infected and $v$ has never been infected.

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

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

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

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

**Algorithm 2** Simulation of an epidemic and transmission tree over a contact network

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

### Phylogenetic kernel

The tree kernel developed by Poon et al. [40] provides a comprehensive similarity score between two phylogenetic trees. The kernel computes the dot-product of two feature vectors, corresponding to the two trees, in the infinite-dimensional feature space of all possible subset trees

with branch lengths (see ??). The kernel was implemented using the fast algorithm developed by Moschitti [101]. First, the number of leaf children of each node, also known as its *production rule*, is recorded. The nodes of both trees are ordered by production rule, and a list of pairs of nodes sharing the same production rule is created. These are the nodes for which the value of the tree kernel must be computed - all other pairs have a value of zero. The pairs to be compared are then re-ordered so that the child nodes are always evaluated before their parents. Due to its recursive definition, ordering the pairs in this way allows the tree kernel to be computed by dynamic programming. The complexity of this implementation is $O(|T_1||T_2|)$, where $|T|$ counts the number of nodes in the tree $T$.

**Adaptive sequential Monte Carlo for Approximate Bayesian computation**

I implemented the adaptive SMC algorithm for ABC developed by Del Moral, Doucet, and Jasra [2].

For ease of exposition, we simplify the notation of Del Moral, Doucet, and Jasra [2] by dropping the subscripts on the variables which indicate the current iteration number. Instead, we will add a prime $'$ to indicate a value which will be used in the next iteration (this should become clear later on).

In the algorithm, we keep track of a population of $n$ sets of model parameters, called *particles*, denoted $\{X_i\}_{i=1}^n$. For the BA model, the particles would be 4-tuples $(N, I, m, \alpha)$. Each particle $X_i$ is associated with a set of $M$ simulated datasets, denoted $\left\{X_{i,j}\right\}_{j=1}^M$, and a weight $W_i$.

The particles are initially drawn from the prior distribution.

Let $d$ be a distance measure on data sets, so that $d(x, y)$ is smaller the more similar $x$ and $y$ are to each other. Let $\varepsilon$ be the tolerance level which indicates whether a data set is "close" to the observed data. That is, if $d(x, y) < \varepsilon$, we will say that $x$ and $y$ are close, otherwise they are distant.

Next, we calculate the next tolerance $\varepsilon^*$. Before we explain how this is done, we first need to define how we adjust the weights on the particles. As explained above (see subsection 1.4.3), the idea of sequential Monte-Carlo is to begin with the prior distribution $\pi$, progress smoothly through a series of intermediate distributions $\pi_1, \ldots, \pi_{n-1}$, and eventually arrive at the target posterior distribution $\pi_n$. In the $k$th iteration, the distribution $\pi_k$ is approximated by the particles and their weights.

In contrast to most existing sequential Monte-Carlo methods, this algorithm does not require the user to specify a sequence of decreasing tolerances to approach the target posterior distribution. Rather, the tolerances are computed adaptively at each step, starting from infinity

at the first iteration.

The algorithm may be stopped when the tolerance reaches a user-defined final value, or when the rate of acceptance of the Metropolis-Hastings (MH) kernel reaches a user-defined threshold.

### 2.1.2 Analysis of Barabási-Albert model

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

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

#### Kernel classifiers

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

The attachment power parameter $\alpha$ was varied among three values: 0.5, 1.0, and 1.5. For each value, the *sample_pa* function in the *igraph* package was used to simulate 100 networks, with the other parameters set to $N = 5000$ and $m = 2$. This step yielded a total of 300 networks. An epidemic was simulated on each network using our *nettree* binary until $I = 1000$ nodes were infected, at which point 500 of them were sampled to form a transmission tree. A total of 300

transmission trees were thus obtained, comprised of 100 trees for each of the three values of $\alpha$. The trees were "ladderized" so that the subtree descending from the left child of each node was not smaller than that descending from the right child. Summary statistics, such as Sackin's index and the ratio of internal to terminal branch lengths, were computed for each simulated tree using our *treestat* binary. The trees were visualized using the *ape* package. Our *treekernel* binary was used to calculate the value of the kernel for each pair of trees, with the meta-parameters set to $\lambda = 0.3$ and $\sigma = 4$. These values were stored in a symmetric $300 \times 300$ kernel matrix. Similarly, we computed the nLTT statistic between each pair of trees using our *treestat* binary, and stored them in a second $300 \times 300$ matrix.

To investigate the effect of $\alpha$ on tree shape, we constructed classifiers for $\alpha$ based on three statistics. First, we used the *kernlab* package [102] to create a kernel support vector machine (kSVM) classifier using the computed kernel matrix. Second, we used the *e1071* package to create an ordinary SVM classifier using the pairwise nLTT matrix. Finally, we performed an ordinary linear regression of $\alpha$ against Sackin's index. Each of these classifiers was evaluated with 1000 two-fold cross-validations. We also performed a kernel principal components analysis (kPCA) projection of the kernel matrix, and used it to visualize the separation of the different $\alpha$ values in the tree kernel's feature space. A schematic of this experiment is presented in fig. 2.2.

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

| varied parameter | $N$ | $\alpha$ | $m$ | $I$ | tips | $\lambda$ | $\sigma$ |
|---|---|---|---|---|---|---|---|
| $N$ | 3000, 5000, 8000 | 1.0 | 2 | 500, 1000, 2000 | 100, 500, 1000 | 0.2, 0.3, 0.4 | 1/8, 1/4, 1/2, 1, 2, 4, 8 |
| $\alpha$ | 5000 | 0.5, 1.0, 1.5 | 2 | 500, 1000, 2000 | 100, 500, 1000 | 0.2, 0.3, 0.4 | 1/8, 1/4, 1/2, 1, 2, 4, 8 |
| $m$ | 5000 | 1.0 | 2, 3, 4 | 500, 1000, 2000 | 100, 500, 1000 | 0.2, 0.3, 0.4 | 1/8, 1/4, 1/2, 1, 2, 4, 8 |
| $I$ | 5000 | 1.0 | 2 | 500, 1000, 2000 | 100, 500 | 0.2, 0.3, 0.4 | 1/8, 1/4, 1/2, 1, 2, 4, 8 |

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

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

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

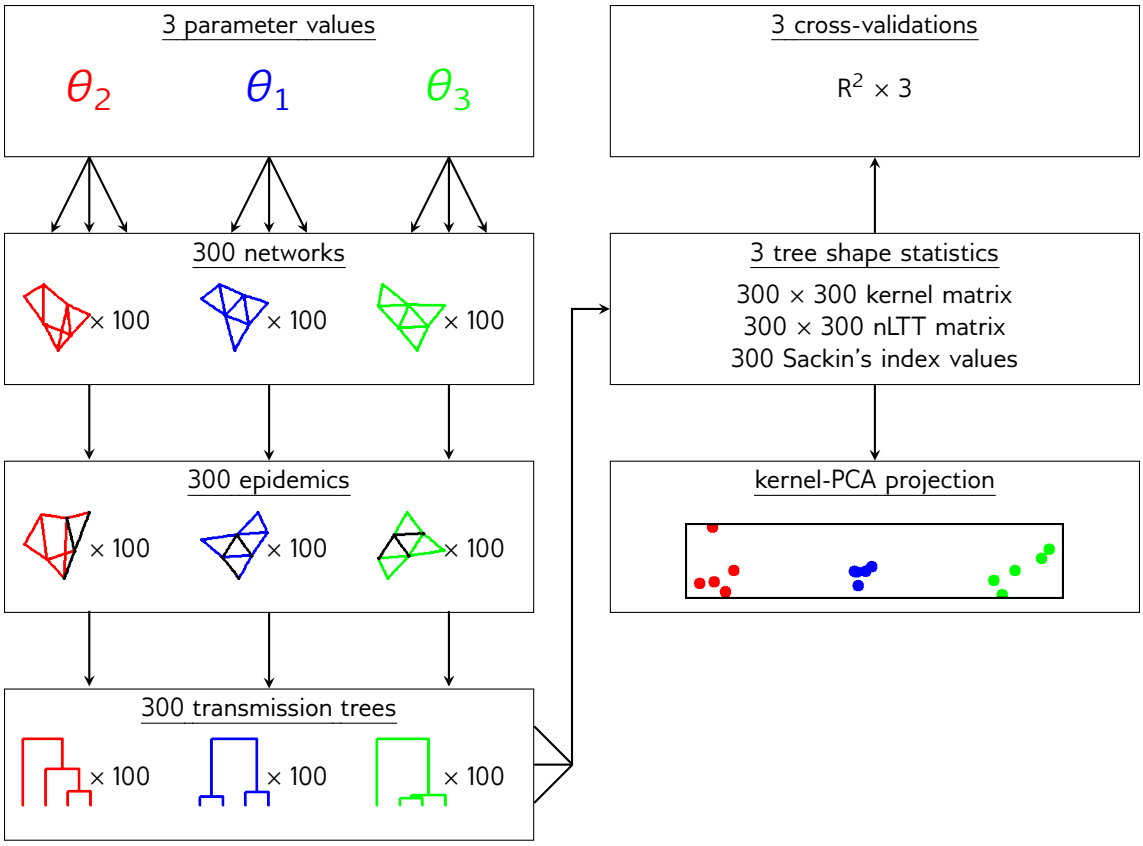



Figure 2.2: Schematic of investigation of BA model parameters using the tree kernel.

### Grid search

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

function in the *coda* package.

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

### Approximate Bayesian computation

To test the full kernel-ABC algorithm, we simulated four trees each under a variety of parameter values, and ran the *netabc* program to estimate posterior distributions for the parameters. The parameter values and priors used are listed in 2.3. The tree kernel meta-parameters were set to $\lambda = 0.3$ and $\sigma = 4$. The SMC algorithm was run with 1000 particles, five sampled datasets per particle, and the parameter controlling the tradeoff between speed and approximation quality Del Moral, Doucet, and Jasra [called "$\alpha$" in 2] set to 0.95. The algorithm was stopped when the acceptance rate of the MH kernel dropped below 1.5%, the same criterion used by Del Moral, Doucet, and Jasra. Approximate marginal posterior densities for each parameter were calculated using the *density* function in *R* applied to the final weighted population of particles. Credible intervals were obtained for each parameter using the *HPDinterval* function in the *coda* package.

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

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

### Characterization of power-law exponent in Barabási-Albert networks

Most studies of social network or transmission network parameters [e.g. 63, 64, 70, 103] report the coefficient $\gamma$ of the power law degree distribution. To make our results comparable to previous work, we used simulated networks to investigate the relationship between the BA model parameters and $\gamma$. A network was simulated for each combination of parameters listed in table 2.4. A power law distribution was fitted to the degree distribution of each simulated network using the *fit_power_law* function in *igraph* with the 'R.mle' implementation. We fitted

| parameter | values |
|---|---|
| $N$ | 500, 600, ..., 15000 |
| $\alpha$ | 0, 0.01, ..., 2 |
| $m$ | 1, 2, ..., 8 |

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

a GLM with Gamma-distributed errors and a log link function to the observed distribution of $\gamma$ values, with $\alpha$, $m$, $N$, and all possible interaction terms as predictors.

### 2.1.3 Real data experiments

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

We downloaded all sequences associated with each published study from GenBank. For the Novitsky et al. [104] data, each sequence was aligned pairwise to the HXB2 reference sequence (GenBank accession number HIVHXB2CG) and the hypervariable regions were clipped out with *BioPython* version 1.66+ [105]. Sequences were multiply aligned using *MUSCLE* version 3.8.31 [106], and alignments were manually inspected with *Seaview* version 4.4.2 [107]. Phylogenies were constructed from the nucleotide alignments by approximate maximum likelihood using *FastTree2* version 2.1.7 with the generalized time-reversible (GTR) model. Transmission trees were estimated by rooting and time-scaling the phylogenies by root-to-tip regression, using a modified version of Path-O-Gen (distributed as part of BEAST [108]) as described previously [43].

Three of the datasets [104, 109, and the BC data] were initially much larger than the others, containing 1265, 1299, and 7923 sequences respectively. To ensure that the analyses were comparable, we reduced these to a number of sequences similar to the smaller datasets. For the Li et al. and BC datasets, we detected clusters of size 280 and 399 respectively using a patristic distance cutoff of 0.02 as described previously [66]. Only sequences within these clusters were carried forward. For the Novitsky et al. [104] data, no large clusters were detected using the same cutoff,

so we analysed a subtree of size 180 chosen arbitrarily.

| Reference | Sequences ($n$) | Location | Risk group | Gene |
|---|---|---|---|---|
| [61] | 173 | Beijing, China | MSM | *pol* |
| [110] | 287 | Basque Country, Spain | mixed | *pol* |
| [111] [104] | 180 | Mochudi, Botswana | mixed | *env* |
| [109] | 280 | Shanghai, China | MSM | *pol* |
| [112] | 136 | Romaina | IDU | *pol* |
| N/A | 399 | British Columbia, Canada | IDU | *pol* |

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

## 2.2 Results

### 2.2.1 Kernel classifiers

Trees simulated under different values of $\alpha$ are visibly quite distinct (fig. 2.3). In particular, higher values of $\alpha$ produce networks with a small number of highly connected nodes which, once infected, are likely to transmit to many other nodes (**??**). This results in a more unbalanced, ladder-like structure in the phylogeny, compared to networks with lower $\alpha$ values. Sackin's index was significantly correlated with $\alpha$ (Spearman's rho 0.85, $p < 10^{-5}$). None of the other three parameters produced trees which were as easily distinguished from each other (figs. S1 to S3). However, the ratio of internal to terminal branch lengths was positively correlated with $N$ and negatively correlated with $I$ (Spearman's rho 0.18 and $-0.69$, both $p < 1e-5$).

Kernel-principal components analysis (PCA) projections show that all three $\alpha$ values are well separated from each other in feature space under several different sampling scenarios (Figure 2.4). With smaller trees, it becomes harder to visually distinguish $\alpha = 0.5$ from $\alpha = 1.0$ using the first two principal components.

With suitably chosen meta-parameters, the accuracy of the kernel-SVM classifier for $\alpha$ was very high under a variety of prevalence and sampling scenarios. Accuracy was highest, with $R^2$ values above 0.95, for the largest trees and complete sampling (bottom center panel). However, even for trees of size 100 sampled from an epidemic on 2000 nodes, the $R^2$ was above 0.8 (top right panel). In all cases, the accuracy of a Sackin's index-based classifier was also quite high, at about 0.75. An SVM classifier using only the nLTT statistic [39] was slightly worse than Sackin's index, with an $R^2$ of X.

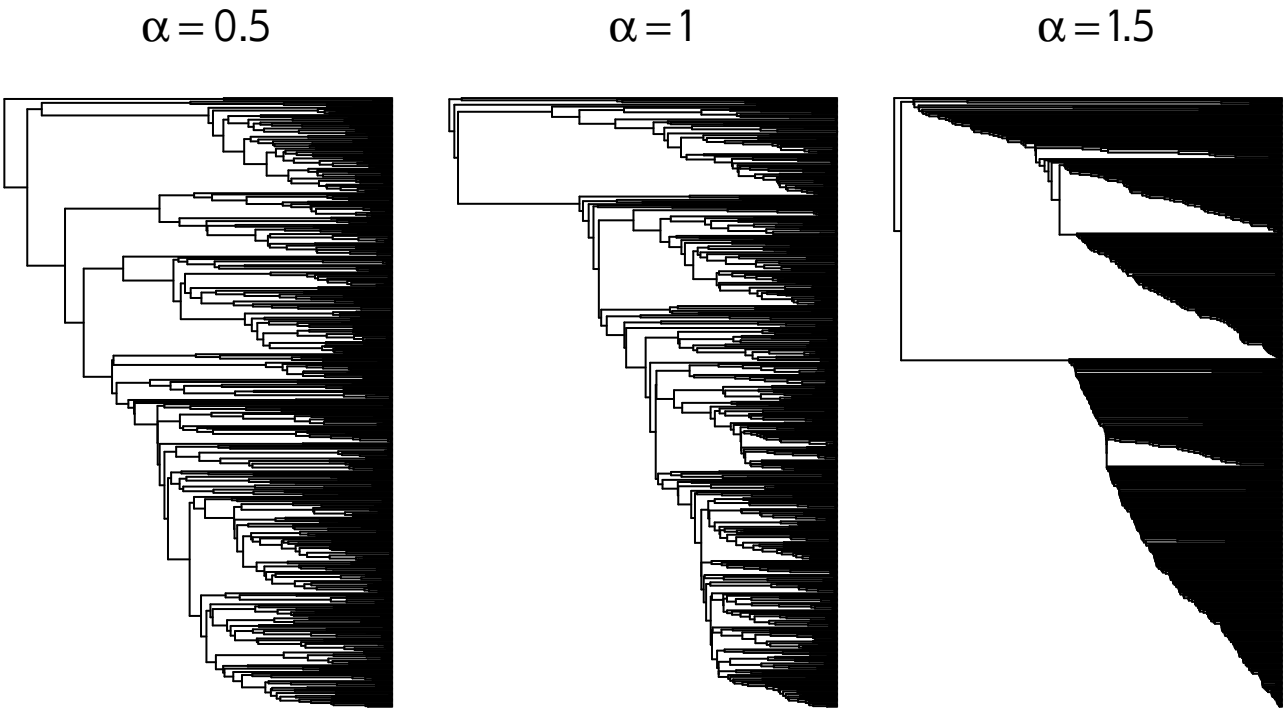



Figure 2.3: Epidemics simulated on BA networks of 5000 nodes, with $\alpha$ equal to 0.5, 1.0, or 1.5, until 1000 individuals were infected. Transmission trees were created by sampling 500 infected nodes. Higher $\alpha$ values produced networks with a small number of highly-connected nodes, resulting in highly unbalanced, ladder-like trees.

We also considered the possibility of inferring the number of infected nodes, or *prevalence*, under this model. All parameters except $I$ were fixed at the following values: $N = 5000$, $\alpha = 1.0$, and $m = 2$. As shown in Figure **??**, the prevalence had no obvious effect on the tree shape. However, a kernel-SVM classifier was able to distinguish the number of infected nodes with high accuracy ($R^2 > 0.9$; Figure **??**). Moreover, the use of the nLTT statistic improved classification accuracy by a small amount, in contrast to the results for $\alpha$. In contrast, a Sackin's index-based classifier displayed extremely poor performance ($R^2 < 0.1$, not shown).

It is important to note that, if we cut off the epidemic simulation when 500 nodes are infected, the resulting tree will be shorter (in calendar time) than if we continue until 2000 nodes are infected. However, this information is not used when building the classifier, since the branch lengths in each tree are scaled by their mean. Therefore, the high performance of the classifier is due to structural differences captured by the tree kernel, rather than the trees simply having different heights.

### 2.2.2 Accuracy of marginal estimates

We used grid search to obtain *marginal* estimates for each network parameter while holding all other parameters fixed. We observed that kernel scores were highest at the values of $\alpha$ on the grid closest to the true values, as shown in Figure 2.5. However, there was a much stronger spike in kernel scores near the true value for $\alpha = 1.0$ and 1.25. This is recapitulated when we look at the accuracy of point estimates obtained by taking the grid value with the highest median kernel score. As shown in Figure 2.6, while the estimates are generally close to the true value, they are much closer for $\alpha = 1.25$ than for the other values.

### 2.2.3 Accuracy of estimates with full ABC

We used kernel-ABC to estimate the parameters of the BA model on simulated trees where the true parameter values were known. Point estimates for each parameter are shown in Figure **??**. Of the four parameters, $\alpha$ was the most accurately estimated, with a median [IQR] absolute error of 0.11 [0.05-0.18]. The accuracy of the estimates was not significantly different between values of $m$ or $I$ (both one-way ANOVA, $p = 0.1$ and 0.25), although the errors when the true value of $\alpha$ was zero were significantly greater than the other values (Wilcoxon rank-sum test, $p = 6.41 \times 10^{-4}$). The error in the estimated value of $I$ was 305.66 [107.76-606.59]. Errors were significantly higher for $\alpha \geq 1$ (Wilcoxon rank-sum test, $p = 6.12 \times 10^{-4}$) and for $I = 2000$ ($p = 1.58 \times 10^{-6}$), but not for any values of $m$ (one-way ANOVA, $p = 0.33$). The $m$ parameter was estimated correctly in 37 % of simulations, with an error of one in 40 % and of two or more in 22 % (the only possible $m$ values were 2, 3, 4, or 5). The true values of $m$ and $I$ did not significantly affect the error (one-way ANOVA, $p = 0.5$ and 0.68), but the accuracy was significantly lower for integral than non-integral values of $\alpha$ (Wilcoxon rank-sum test, $p = 7.2 \times 10^{-3}$). Finally, the total number of nodes $N$ was consistently over-estimated by about a factor of two (error $6.59 \times 10^3$ [$4.21 \times 10^3$-$8.28 \times 10^3$]). No other parameters influenced the accuracy of the $N$ estimates (one-way ANOVA, $p \geq$ NA).

Figure **??** shows the ABC approximation to the posterior distribution on the BA parameters for one simulation (equivalent plots for all the simulations can be found in the supplemental materials). Highest posterior density (HPD) intervals around $\alpha$ and $I$ were narrow relative to the region of nonzero prior density, whereas the intervals for $m$ and $N$ were widely dispersed. Table **??** shows point estimates and 95% HPD intervals averaged over all simulations.

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

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

| | exp(Estimate) | Standard error | P-value |
|---|---|---|---|
| (Intercept) | 1.63 | $5.1 \times 10^{-3}$ | $< 10^{-5}$ |
| $\alpha$ | 1.77 | $4.4 \times 10^{-3}$ | $< 10^{-5}$ |
| $m$ | 1.03 | $1.0 \times 10^{-3}$ | $< 10^{-5}$ |
| $N$ | 1.00 | $5.8 \times 10^{-7}$ | $< 10^{-5}$ |
| $\alpha \times m$ | 1.00 | $8.7 \times 10^{-4}$ | $< 10^{-5}$ |
| $\alpha \times N$ | 1.00 | $5.0 \times 10^{-7}$ | $< 10^{-5}$ |
| $m \times N$ | 1.00 | $1.1 \times 10^{-7}$ | $< 10^{-5}$ |
| $\alpha \times m \times N$ | 1.00 | $9.9 \times 10^{-8}$ | $< 10^{-5}$ |

Table 2.7: Estimated GLM parameters for relationship between power-law exponent $\gamma$ and BA model parameters.

### 2.2.4 Characterization of power-law exponent in Barabási-Albert networks

Table 2.7 shows the estimated parameters for a log-link GLM fitted to the observed distribution of $\gamma$ values. The coefficients are interpretable as multiplicative effects.

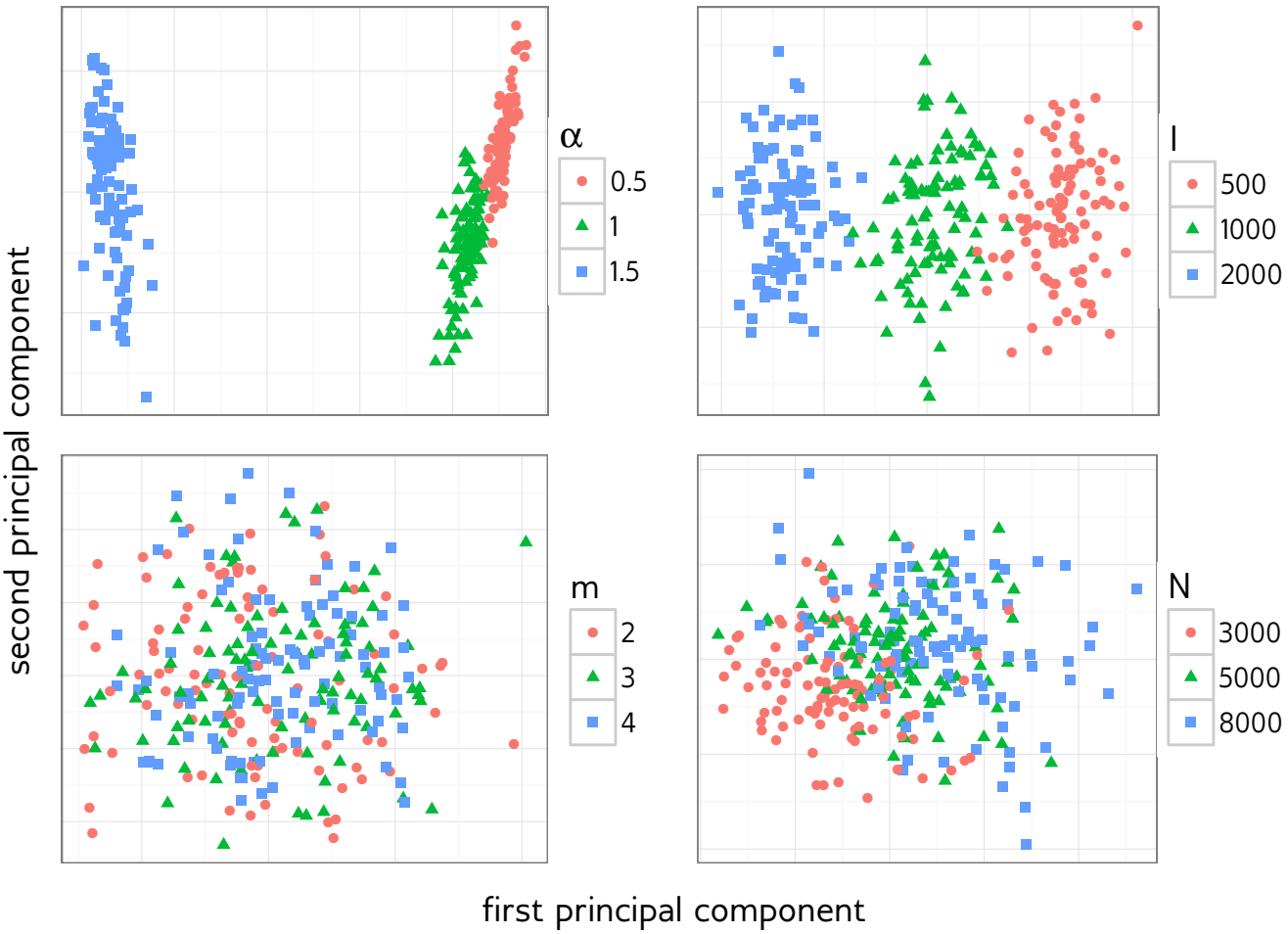



Figure 2.4: Projection of the kernel matrix for different preferential attachment power values onto its first two principal components, for eight simulation scenarios. Each point corresponds to a simulated transmission tree, and is coloured by preferential attachment power. Facets are number of infected nodes (horizontal), and number of sampled tips (vertical). The parameters to the tree kernel were $\lambda = 0.3$ and $\sigma = 4$, and the nLTT was not used. Qualitatively, trees with a larger number of tips are easier to separate in kernel space, regardless of what sampling proportion they represent. In all cases, the highest attachment power can be separated from the other two, but the two lowest values become hand to distinguish with in the 100-tip trees.

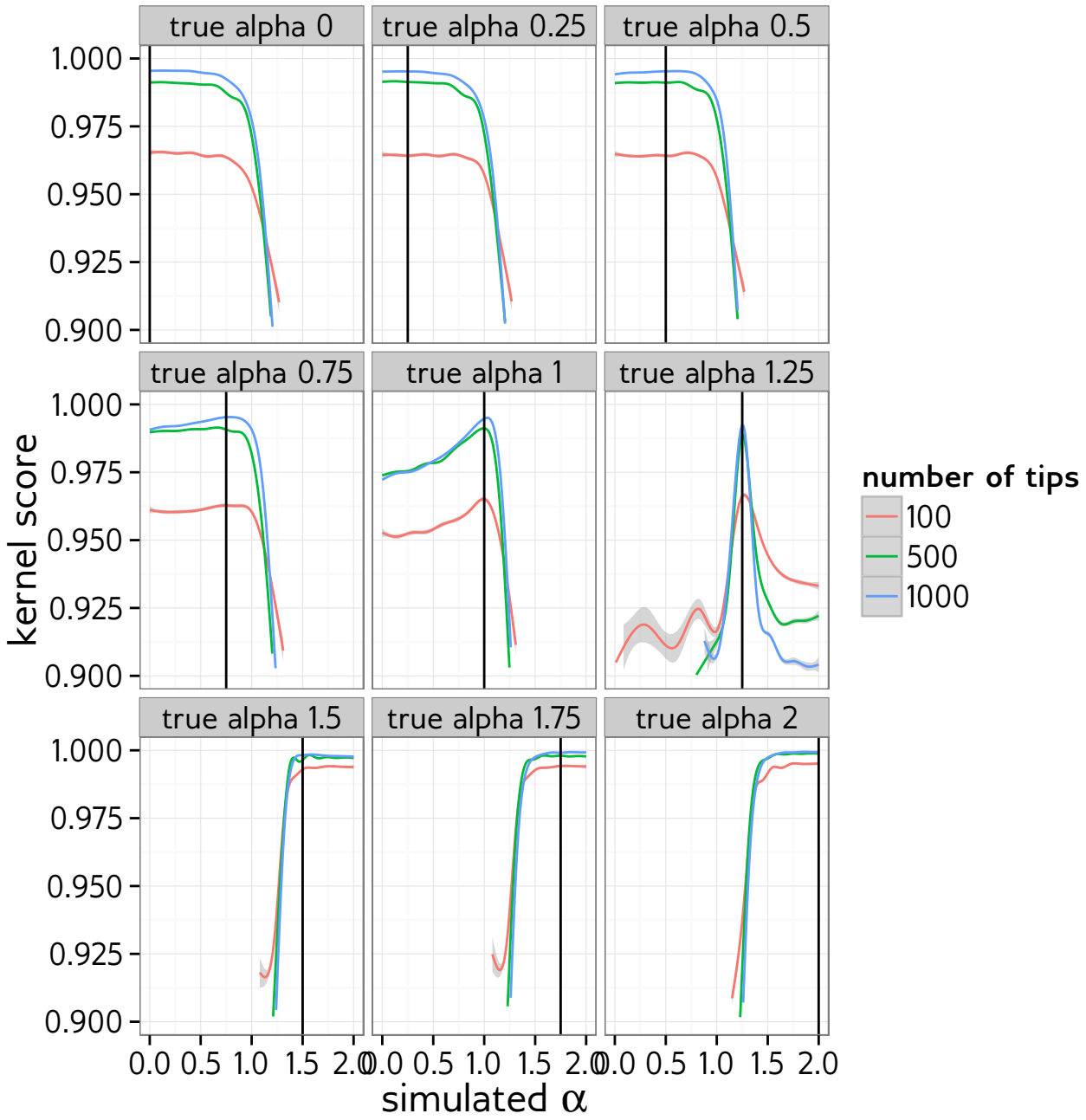



Figure 2.5: Grid search kernel scores for testing trees simulated under various $\alpha$ values. All epidemics had $I = 1000$ infected nodes, on BA networks of size $N = 5000$ with $m$ fixed at 2. Colours indicate the number of sampled tips.

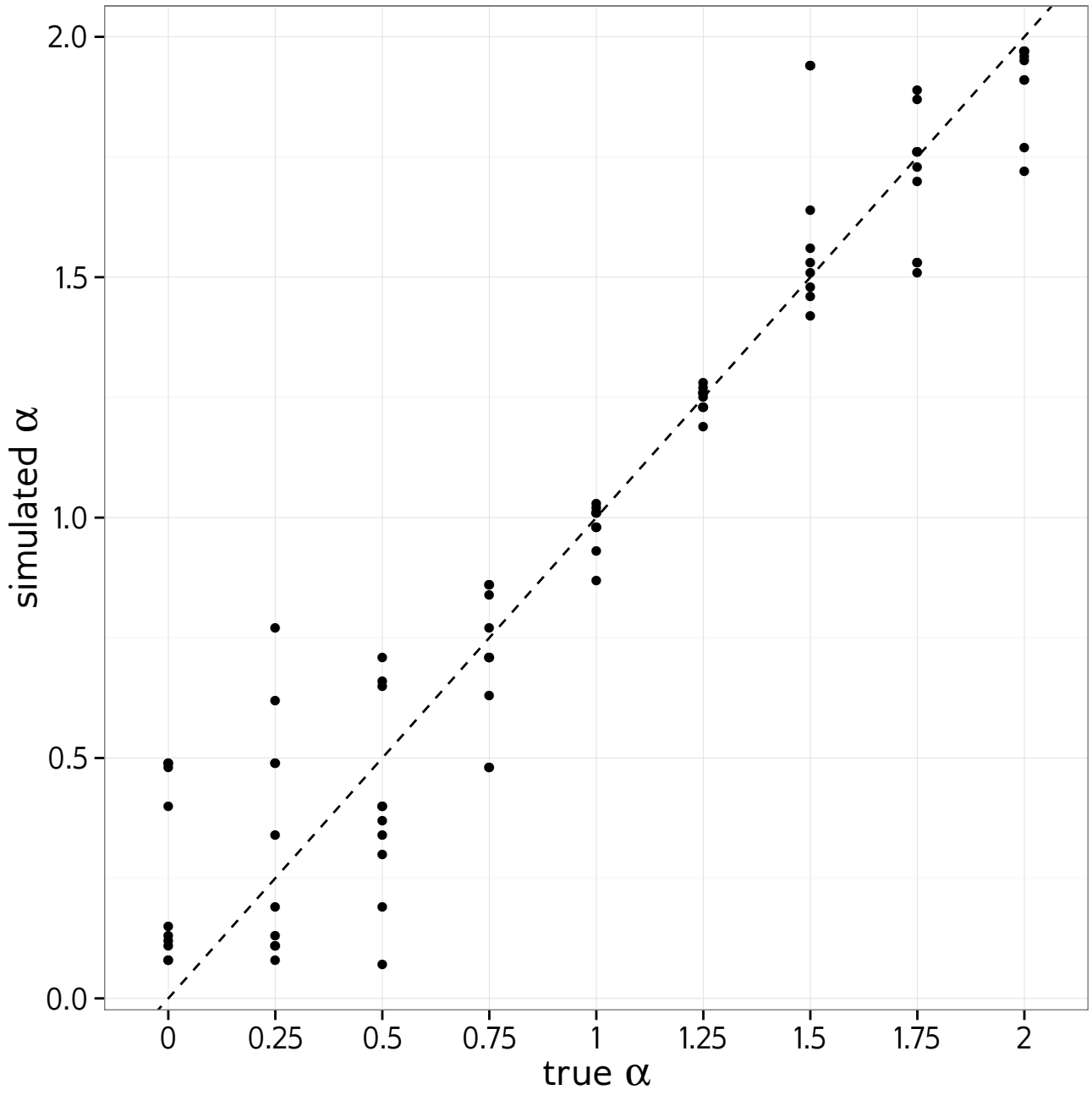



Figure 2.6: Marginal estimates of $\alpha$ obtained with grid search. Training trees were simulated on a narrowly spaced grid of $\alpha$ values, and compared to testing trees using the tree kernel. The $\alpha$ value in the grid with the highest median kernel score was taken as the point estimate for the testing tree. These point estimates are shown as black dots. The dashed line is the identity.

# Chapter 3

# Conclusion

Contact network structure has a substantial impact on epidemic trajectory [55, 57, 58], and a few methods have been developed to estimate network parameters from epidemiological data [55, 71, 72]. It is known that contact network structure can have a substantial impact on transmission tree shape [69, 80, 83, 85, 113].

Our work had three main aims. First, we developed a method to estimate contact network model parameters from phylogenetic data. This method widens the field of models which may be investigated in a phylodynamic context. Second, we investigated the parameters of the BA model. We determined through simulation studies that the preferential attachment power $\alpha$ and number of infected nodes $I$ had a substantial impact on transmission tree shape, and could be estimated using our method.

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

The two-step process of simulating a contact network and subsequently allowing an epidemic to spread over that network carries with it the assumption that the contact network is static over the duration of the epidemic. Clearly this assumption is invalid, as people make and break partnerships on a regular basis. Our work has not addressed this assumption, primarily due

to our desire to avoid the additional complexity required to address the dynamic nature of networks. This simplifying assumption is made by most studies using contact network models in an epidemiological context [9, 54]. However, in principle, kernel-ABC could be adapted to dynamic contact networks by using a method such as that developed by Robinson, Cohen, and Colijn [114] to simulate a dynamic contact network, while concurrently simulating the spread of an epidemic.

It is important to note that our kernel-ABC method takes a transmission tree as input, rather than a viral phylogeny. Thus, we have left the estimation of a transmission tree up to the user. There were two reasons for this choice. First and foremost, we wished again to avoid extra complexity and keep the number of estimated parameters small. In theory, it is possible to incorporate the process by which a viral phylogeny is generated along with a transmission tree into our method, for example by simulating within-host dynamics. Although this may be an avenue for future extension, we felt that it would obscure the primary purpose of this work, which is to study contact network parameters. Second, there are a number of different methods available for inferring transmission trees [31–34, 43], some of which incorporate geographic and/or epidemiological data not accommodated by our method. We therefore felt it would be best to allow researchers to use their own preferred tree building method.

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

# Bibliography


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

[2] 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.

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

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

[5] 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.

[6] 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.

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

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

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

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

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

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

[13] 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.

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

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

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

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

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

[19] 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.

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

[21] 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.

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

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

[24] 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.

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

[26] 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.

[27] 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.

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

[29] Remco Bouckaert et al. “BEAST 2: a software platform for Bayesian evolutionary analysis”. In: *PLoS Comput Biol* 10.4 (2014), e1003537.

[30] Fredrik Ronquist et al. “MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space”. In: *Systematic biology* 61.3 (2012), pp. 539–542.

[31] 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.

[32] 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.

[33] 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.

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

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

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

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

[38] 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.

[39] 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.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

[61] 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* (2015), p. civ526.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

[85] Luc Villandre et al. "Assessment of Overlap of Phylogenetic Transmission Clusters and Communities in Simple Sexual Contact Networks: Applications to HIV-1". In: *PloS one* 11.2 (2016), e0148459.

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

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

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

[89] GA Watterson. "On the number of segregating sites in genetical models without recombination". In: *Theoretical population biology* 7.2 (1975), pp. 256–276.

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

[91] 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.

[92] 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.

[93] 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.

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

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

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

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

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

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

[100] 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.

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

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

[103] 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.

[104] 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.

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

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

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

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

[109] 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.

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

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

[112] 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.

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

[114] 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.

# Appendix: Supplemental Figures

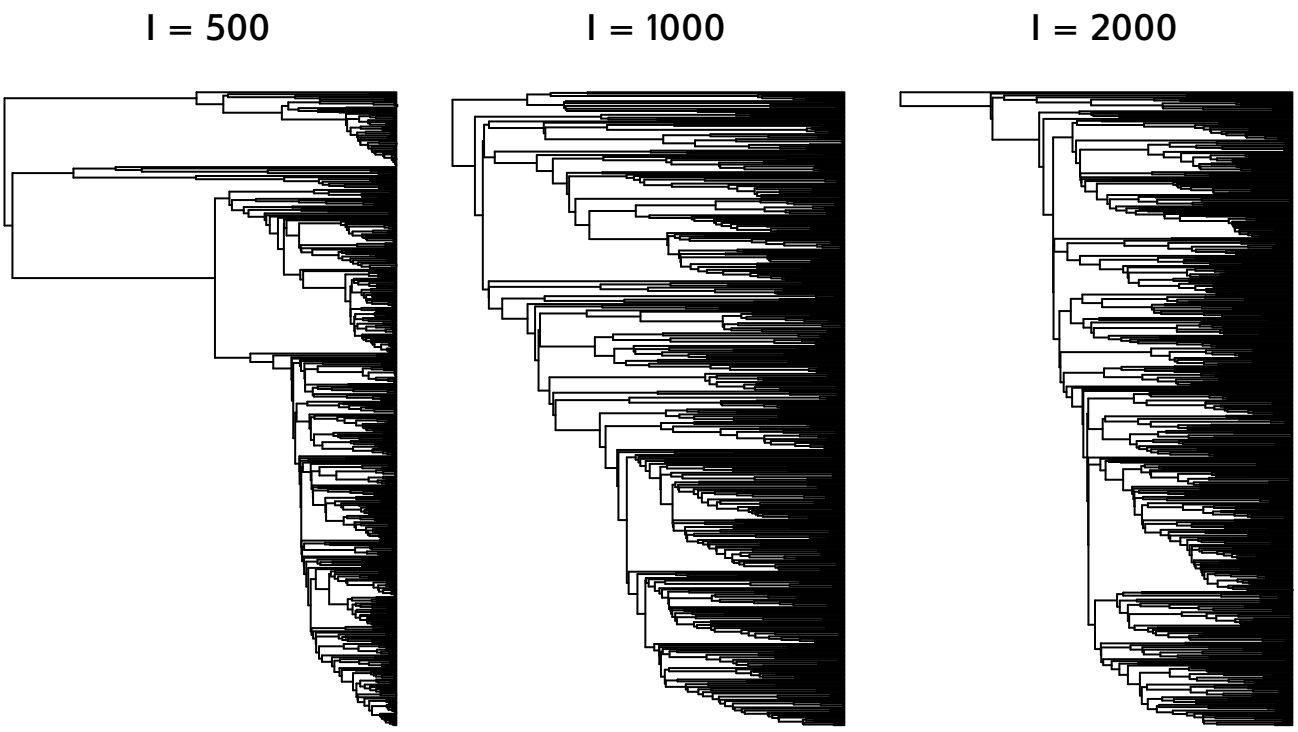



Figure S1: Transmission trees simulated over BA networks with varying values of $I$, the number of infected nodes when the epidemic simulation was stopped.

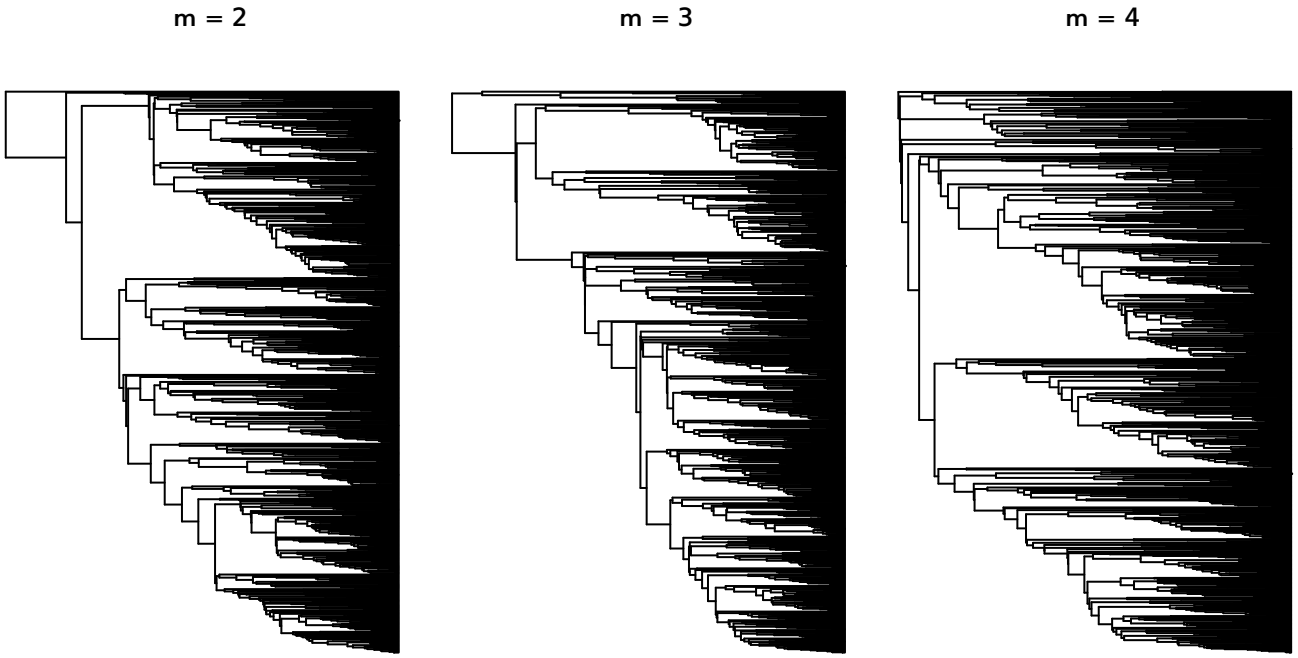



Figure S2: Transmission trees simulated over BA networks with varying values of $m$, the number of edges added per vertex.

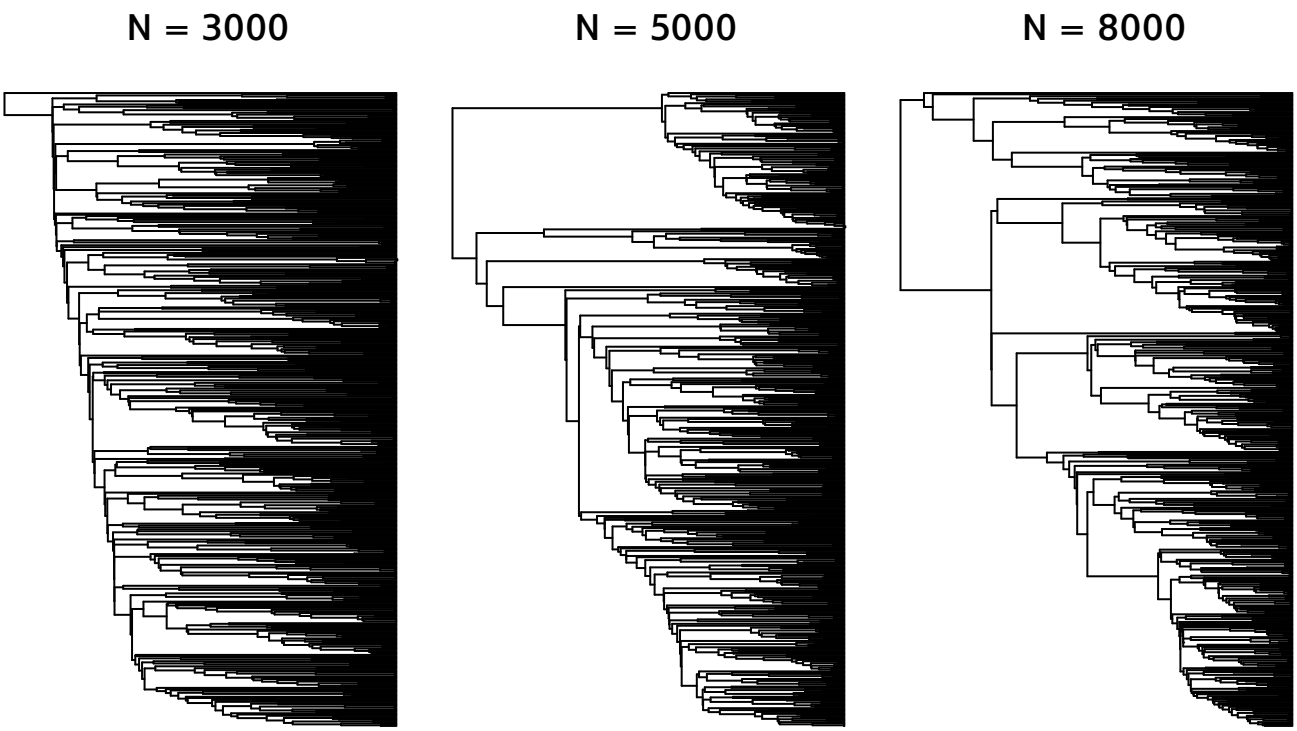
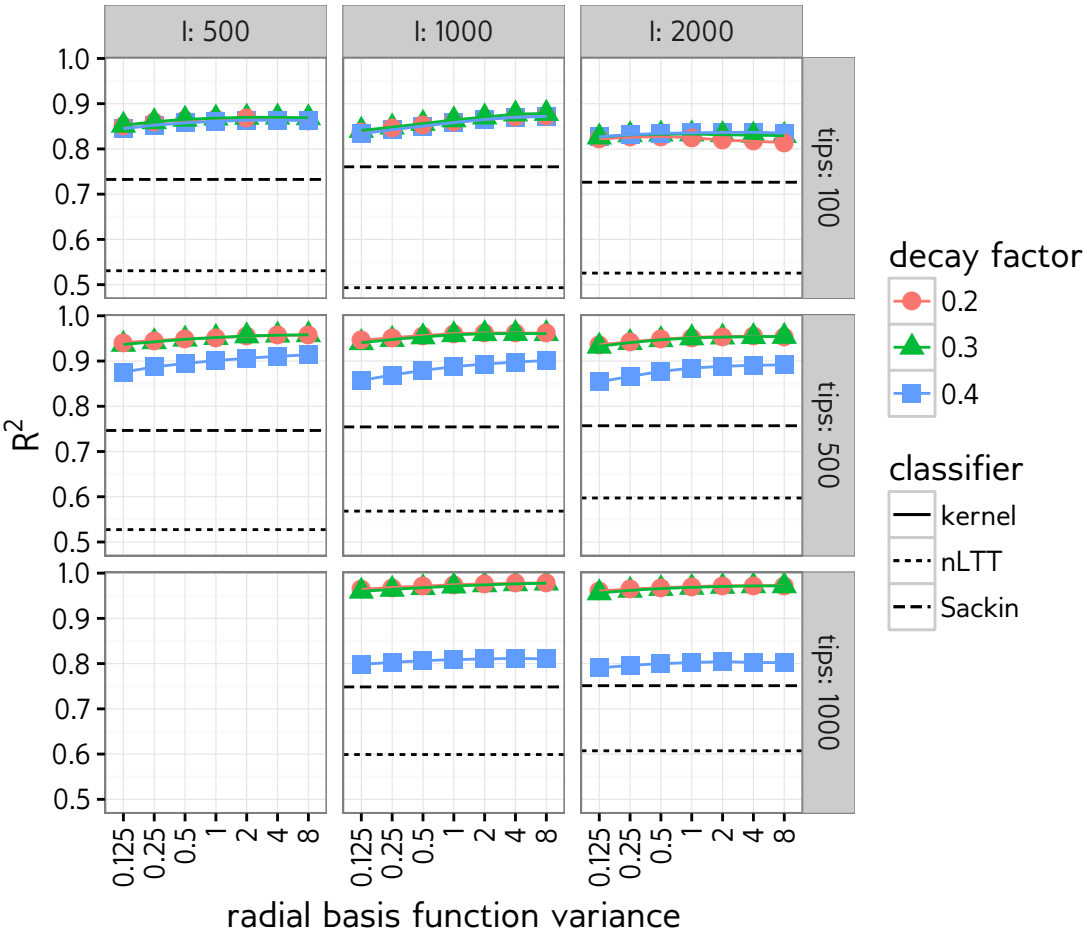
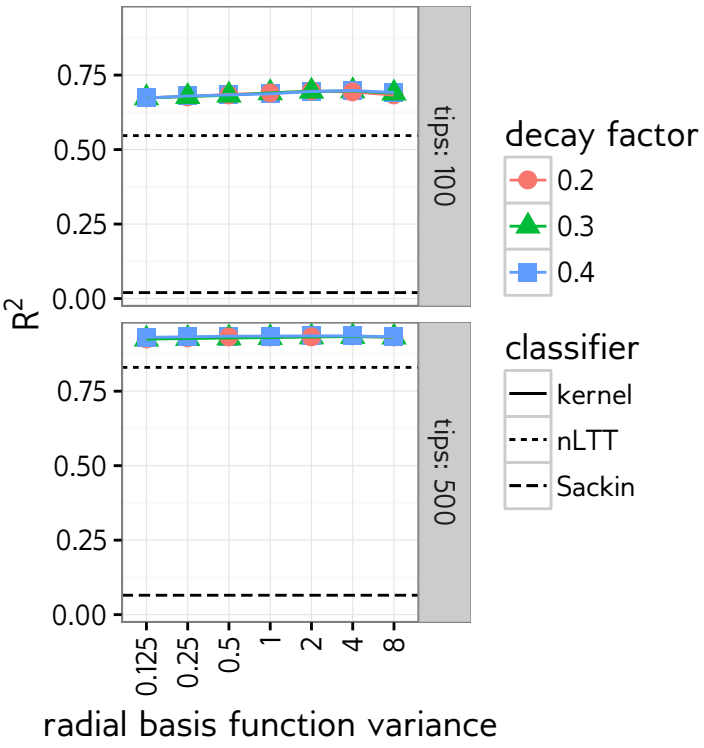
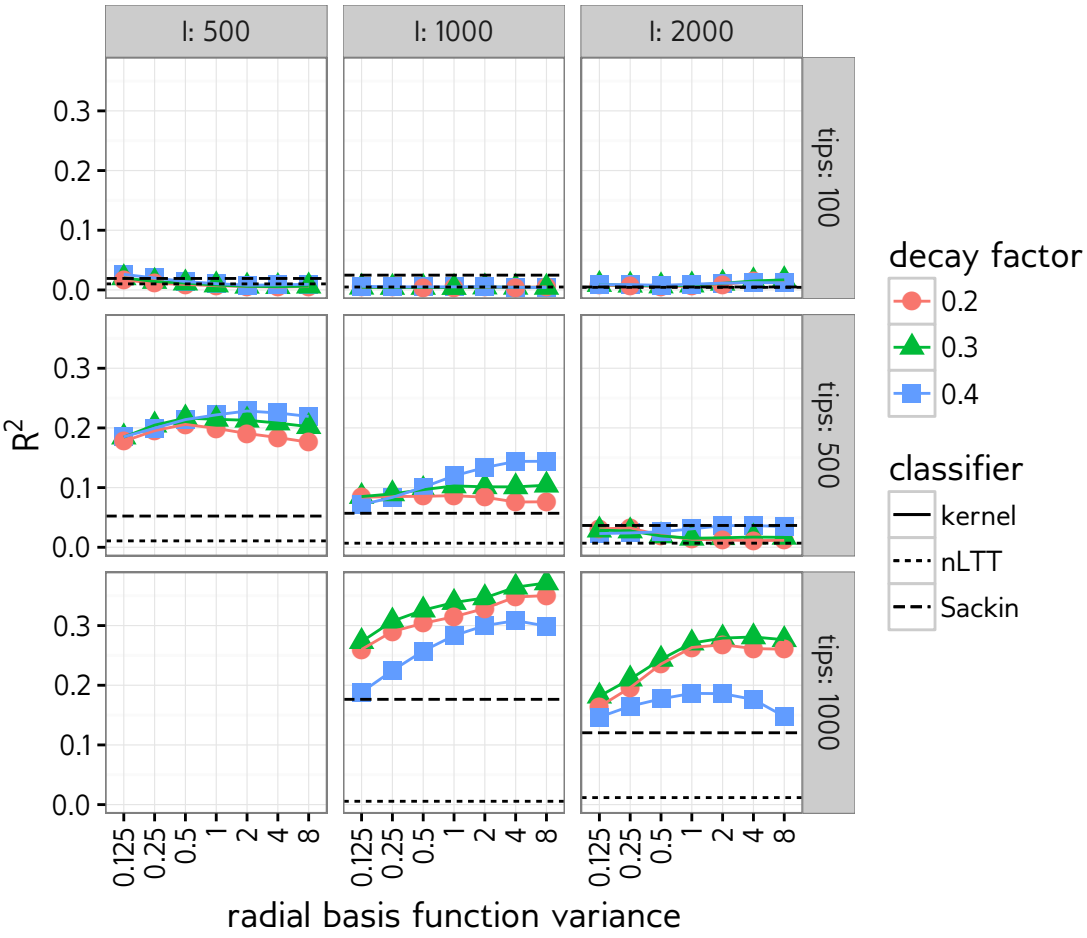
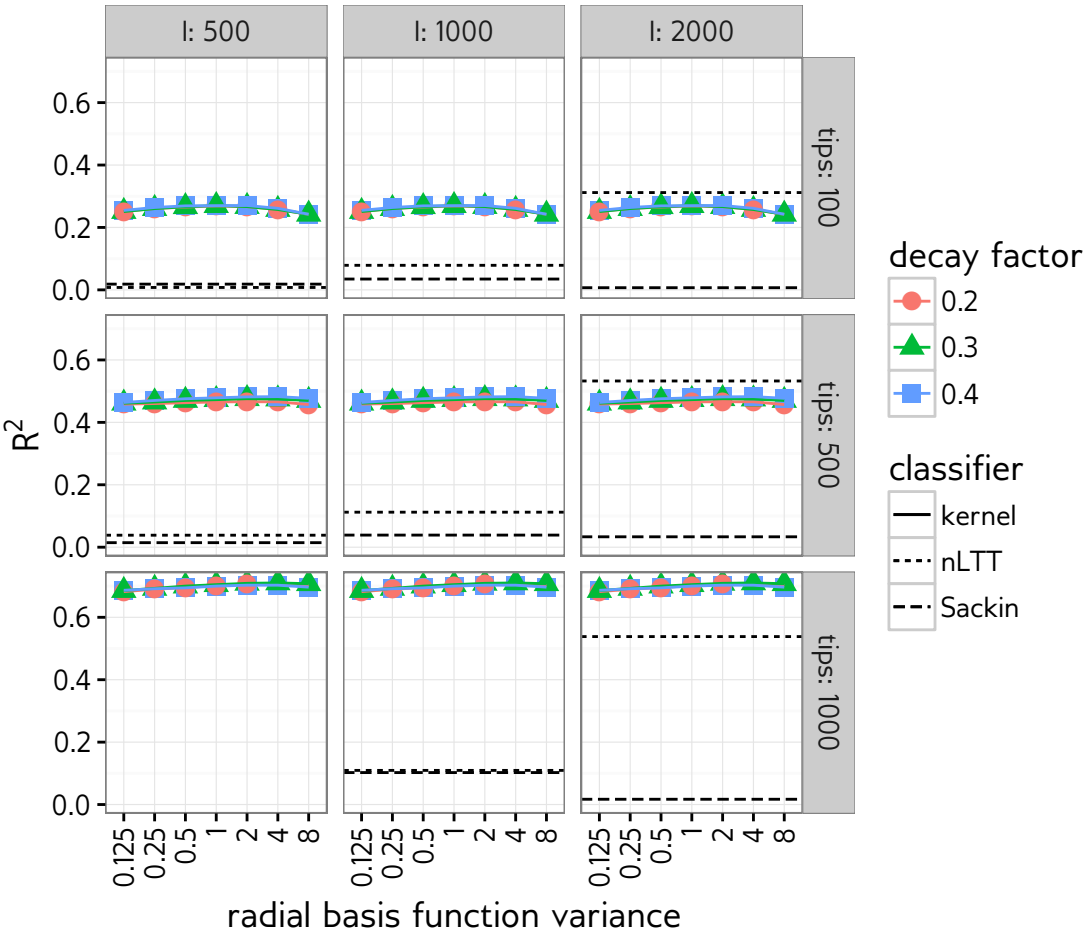


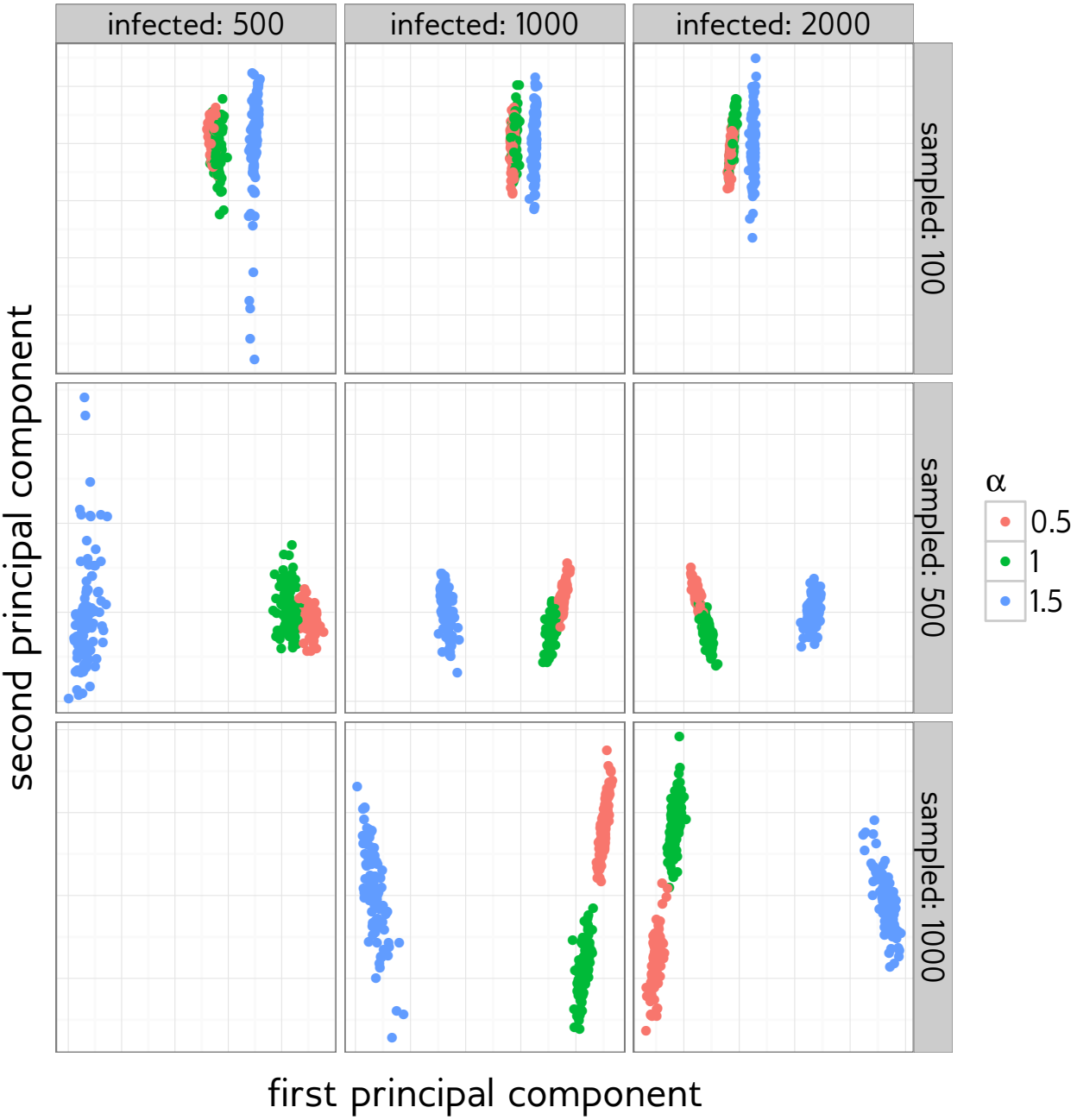

Figure S3: Transmission trees simulated over BA networks with varying values of $N$, the number of nodes in the network.

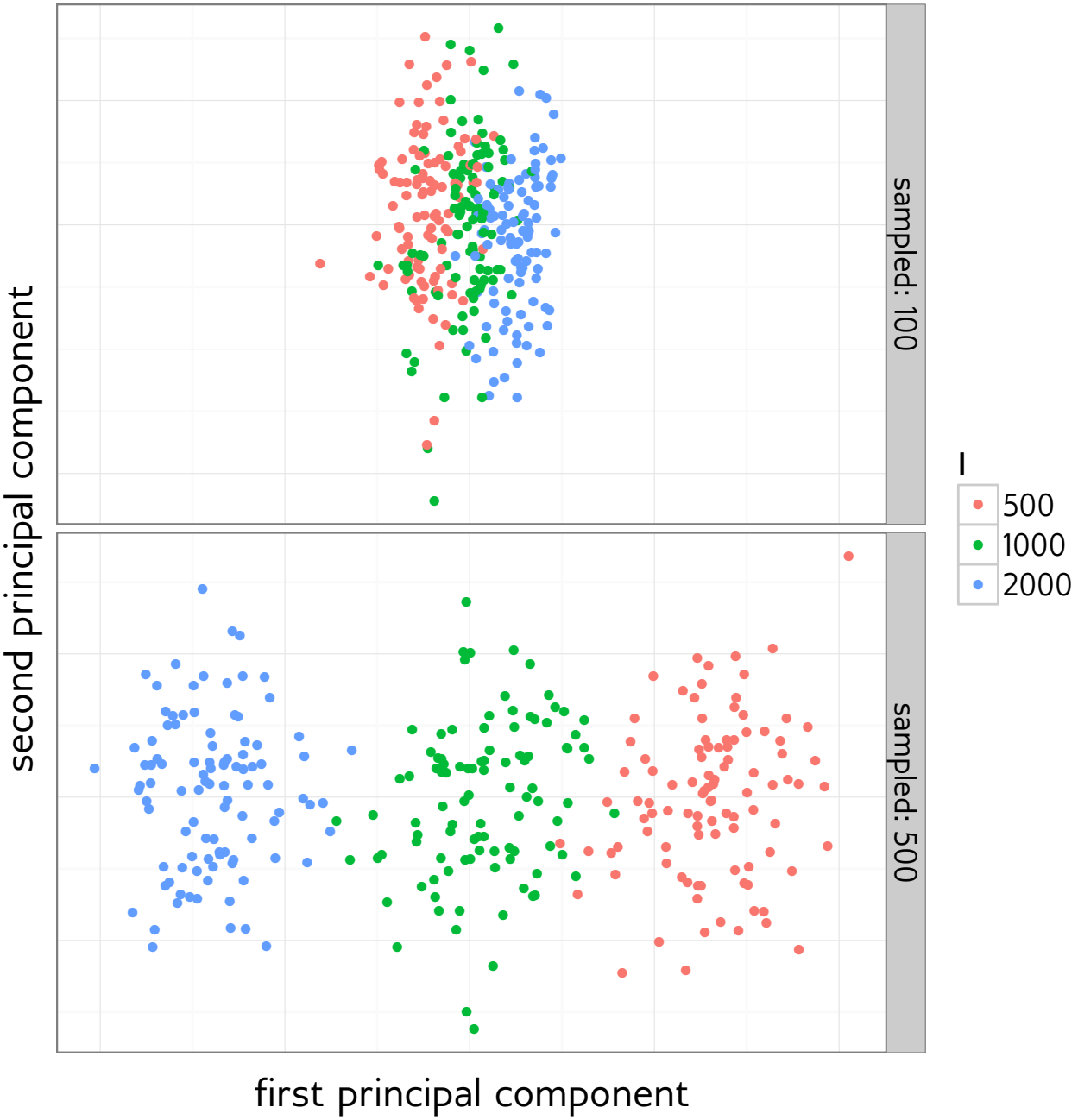

l: 500
l: 1000
l: 2000
tips: 100
tips: 500
tips: 1000
$R^2$
1.0
0.9
0.8
0.7
0.6
0.5
0.125
0.25
0.5
1
2
4
8
radial basis function variance
decay factor
0.2
0.3
0.4
classifier
kernel
nLTT
Sackin

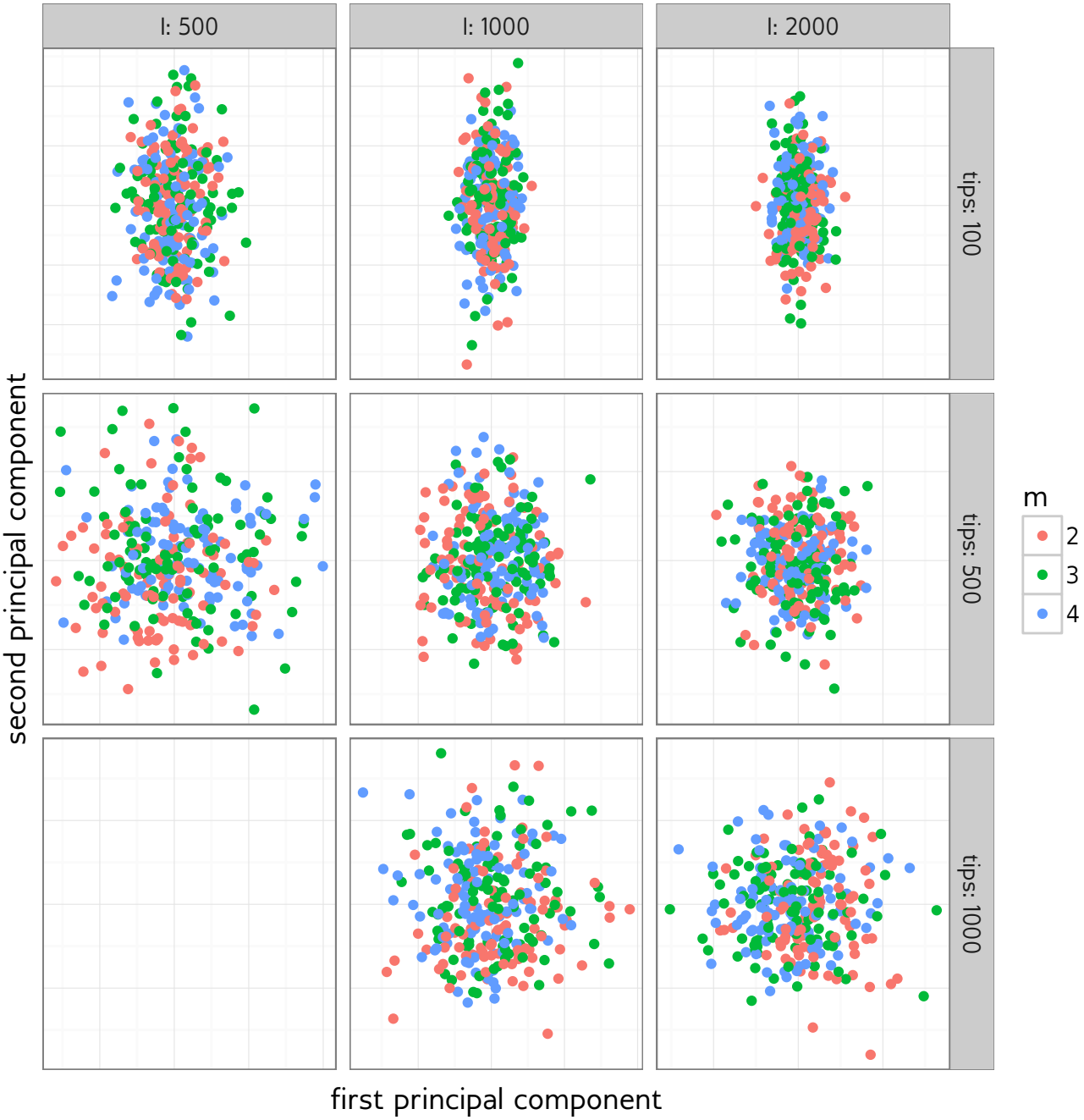

0.75
0.50
0.25
0.00
tips: 100
tips: 500
R2
0.125
0.25
0.5
1
2
4
8
radial basis function variance
decay factor
0.2
0.3
0.4
classifier
kernel
nLTT
Sackin

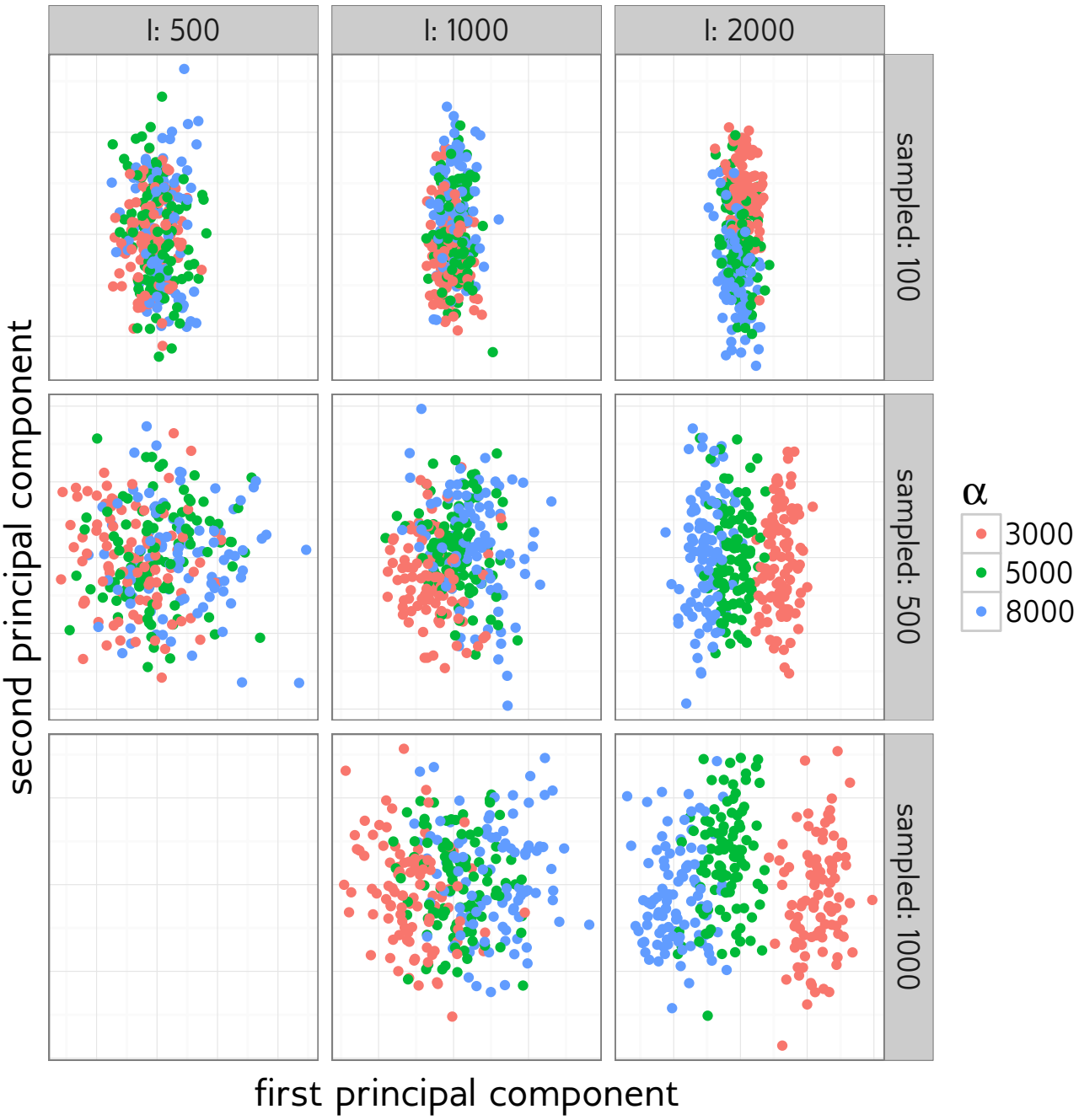

l: 500
l: 1000
l: 2000
tips: 100
tips: 500
tips: 1000
$R^2$
0.0
0.1
0.2
0.3
0.125
0.25
0.5
1
2
4
8
radial basis function variance
decay factor
0.2
0.3
0.4
classifier
kernel
nLTT
Sackin


l: 500
l: 1000
l: 2000
tips: 100
tips: 500
tips: 1000
$R^2$
0.0
0.2
0.4
0.6
0.125
0.25
0.5
1
2
4
8
radial basis function variance
decay factor
0.2
0.3
0.4
classifier
kernel
nLTT
Sackin


infected: 500
infected: 1000
infected: 2000
sampled: 100
sampled: 500
sampled: 1000
second principal component
first principal component
α
0.5
1
1.5


second principal component
first principal component
sampled: 100
sampled: 500
I
500
1000
2000


l: 500
l: 1000
l: 2000
tips: 100
tips: 500
tips: 1000
second principal component
first principal component
m
2
3
4


l: 500
l: 1000
l: 2000
sampled: 100
sampled: 500
sampled: 1000
second principal component
first principal component
α
3000
5000
8000