Phylodynamic inference of contact network parameters through approximate Bayesian computation

Rosemary M. McCloskey and Art F.Y. Poon

February 23, 2016

Contents

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.

α preferential attachment power parameter in Barabási-Albert networks.

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

List of Abbreviations

ABC approximate Bayesian computation.

BA Barabási-Albert.

ER Erdős-Rényi.

GSL GNU scientific library.

HMM hidden Markov model.

i.i.d. independent and identically distributed.

IS importance sampling.

LTT lineages-through-time.

MAP maximum *a posteriori*.

MCMC Markov chain Monte Carlo.

ML maximum likelihood.

nLTT normalized lineages-through-time.

PCA principal components analysis.

pdf probability density function.

SI susceptible-infected.

SIR susceptible-infected-recovered.

SIS sequential importance sampling.

SMC sequential Monte Carlo.

SVM support vector machine.

WS Watts-Strogatz.

Chapter 1

Introduction

1.1 Phylogenetics and phylodynamics

1.1.1 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*) (Haeckel 1866). 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 (Harding 1971). 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* (Cavalli-Sforza and Edwards 1967).

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 map to the present day, as a reference point (Nee, Mooers, and Harvey 1992). 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 (Buneman 1974). These concepts are illustrated in Figure 1.1.1.

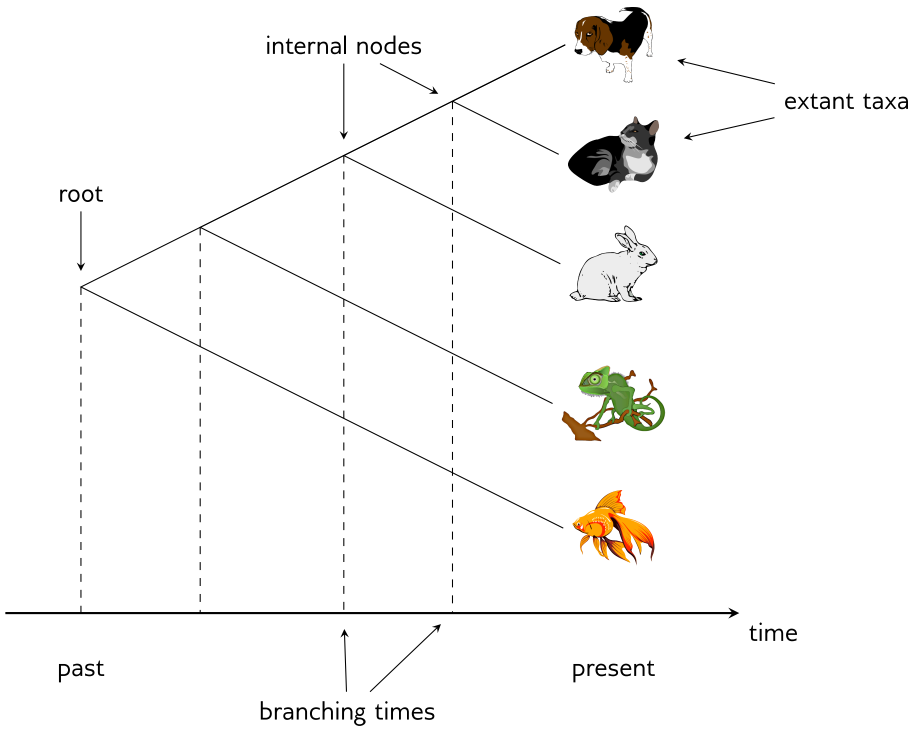


Figure 1.1: Illustration of a rooted, ultrametric, time-scaled phylogeny. The tips of the tree, which represent extant taxa, are placed at the present day on the time axis. Internal nodes, representing extinct common ancestors to the extant taxa, fall in the past. The topology of the tree indicates that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

1.1.2 Transmission trees and epidemiology

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 epidemiolgical 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 a root corresponding to the index case, and branching times indicating times of transmission. 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 (A. J. Drummond et al. 2003), in contrast to the *isochronous* taxa found in the phylogenies of macro-organisms. A transmission tree is illustrated in Figure 1.1.2.

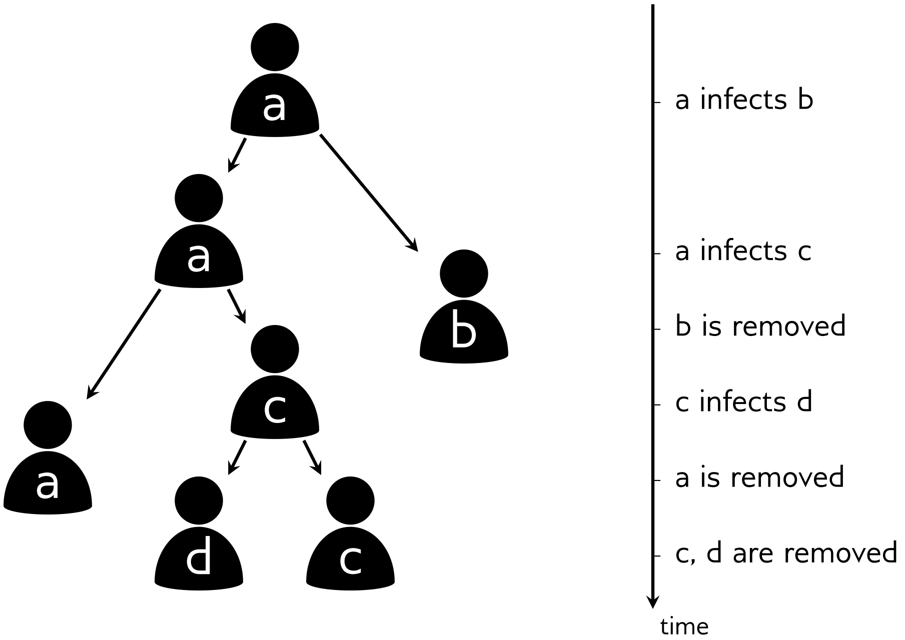


Figure 1.2: A transmission tree representing an epidemic's progress through four individuals. The index case, patient *a*, transmitted the virus to patients *b* and *c*; patient *c* went on to infect patient *d*.

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 (Nee, Mooers, and Harvey 1992), 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 (Holmes et al. 1995). Many more diverse epidemiological parameters have been investigated using transmission trees, such as the degree of clustering (Hughes et al. 2009) and the effect of elevated transmission risk in acute infection (Volz, Koopman, et al. 2012). However, in all but the most well-studied of epidemics, this is not possible to obtain through traditional epidemiological methods. The time and effort to conduct detailed interviews and contact 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.

1.1.3 Estimating transmission trees from 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 tips of the tree are labelled with both the virus and its host.

Is "strain" the best word here?

The crux of *phylodynamics* (Grenfell et al. 2004) is the fact that the processes generating these two types of tree - epidemiological processes, for transmission trees, and evolutionary processes, for viral phylogenies - occur on similar time scales for RNA viruses (A. J. Drummond et al. 2003). As a result, there is a close relationship between the two trees. In particular, the transmission process is quite similar to *allopatric speciation* (Coyne and Orr 2004), 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 causes the extinction of their viral lineage. 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 produce unrooted trees whose branch lengths measure genetic distance in units of expected substitutions per site, whereas transmission trees are rooted with branches measuring calendar time. We generally assume 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. Of course, there are an infinite number of possible combinations of root and branch lengths which result in the correct placement of the tips, so we would also like to find values which are the most "reasonable" given our data. By reasonable, we mean 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 (Shankarappa et al. 1999; Korber et al. 2000; A. Drummond, Oliver, Rambaut, et al. 2003), which we apply in this work, and least-square dating (**to2015fast**). 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 with an outgroup (**li1988rates**) before time-scaling.

needs a reference

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. (1996) 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. (Nei and Kumar 2000) 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.

needs a reference

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

are tip dates part of the prior?

analyses include BEAST (**bouckaert2014beast**) and MrBayes (**ronquist2012mrbayes**).

Several other authors have developed methods tailor-made for inferring transmission trees. Didelot, Gardy, and Colijn (2014) 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. (2008) 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. A different approach is undertaken by Jombart et al. (2011), 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.

1.1.4 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 (Pybus and Rambaut 2009; Volz, Koelle, and Bedford 2013). Phylogenies are complex objects, and it is not immediately obvious how to extract useful information from them with respect to fitting a parameter. Standard statistical methods built for numeric data cannot be applied directly - for example, one cannot perform a regression of a parameter of interest against a phylogeny. Therefore, before we discuss exactly what phylodynamics can tell us about epidemics, and how it has been applied in the past, we will review some of the methods for quantifying the shapes of phylogenies and their similarity to each other.

The shape of a phylogeny consists of two components: the topology, and the distribution of branch lengths (Mooers and Heard 1997). Many tree summary statistics have been developed to assign numerical values to phylogenies based on their shapes. One of the most widely used is Sackin's index (Shao 1990), which measures the imbalance or asymmetry in a rooted tree. For the ith tip of the tree, we define N_i to be the number of branches between that tip and the root. The unnormalized Sackin's index is defined as the sum of all N_i. It is called unnormalized because it does not account for the number of tips in the tree. Among two trees having the same number of tips, the least-balanced tree will have the highest Sackin's index. However, among two equally balanced trees, the larger tree will have a higher Sackin's index. This makes it challenging to compare balances among trees of different sizes. To correct this, Kirkpatrick and Slatkin (1993) derive the expected value of Sackin' index under the Yule model (Yule 1925). Dividing by this expected value normalizes Sackin's index, so that it can be used to compare trees of different sizes.

An alternative to summary statistics are *distance measures* on trees. Rather than assigning a numerical value to each tree individually, a distance measure associates each pair of trees with a number, 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) (Janzen, Höhna, and Etienne 2015), which compares the LTT (Nee, Mooers, and Harvey 1992) 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 (Holmes et al. 1995), so large values may indicate that the trees being compared are the products of different epidemic trajectories (Janzen, Höhna, and Etienne 2015).

Another tree distance measure is the *phylogenetic kernel*, or "tree kernel" developed by Poon et al. (2013). Strictly speaking, the tree kernel is a *similarity measure*, rather than a distance measure, as it is maximized when the two trees being compared are the same. The basis of the tree kernel is the kernel trick originally developed for SVMs (Burges 1998). 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 (Collins and Duffy 2002) and did not incorporate branch length information. The version developed by Poon et al. (2013) 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 (Poon 2015). In that paper, Poon used the tree kernel as the distance function in approximate Bayesian computation (ABC) (see Section 1.4), an approach dubbed kernel-ABC, to fit epidemiological models to observed trees.

1.1.5 Applications of phylodynamics

1.2 Contact networks

1.2.1 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 a sexually transmitted pathogen like HIV, contacts would be sexual partnerships. A *contact network* is a graphical representation of a host population and the contacts among its members. The *nodes* in the network represent hosts, and *edges* represent contacts between them.

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, a blood-borne infection spread through transfusions would use undirected 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. In fact, it is possible to represent an undirected edge by two symmetric directed edges. For this reason, we consider only contact networks with directed edges in the sequel. A directed contact network is shown in Figure 1.2.1 (left).

The path an epidemic takes through a contact network determines the topology of the transmission tree relating the infected hosts. The initially infected node who introduces the epidemic 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. This correspondence is illustrated in figure 1.2.1. 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, there are generally several orders of infection which could lead to the same topology, since the labels on the internal nodes of the tree are not available to the researcher.

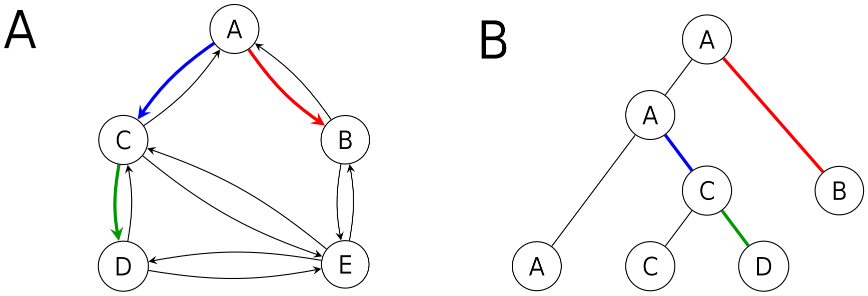


Figure 1.3: Illustration of epidemic spread over a contact network. On the left, a contact network with five hosts, labelled A through E. Directed edges indicate six symmetric contacts among the hosts. Coloured, bolded edges represent transmissions. The epidemic began with node A, who transmitted to nodes B and C. Node C further transmitted to node D, and node E was not infected. On the right, the transmission tree topology corresponding to this scenario. Note that individual B was sampled before the other infected individuals, resulting in a tree with heterochronous tips.

1.2.2 Models for contact networks

Throughout this section, the variable n will be used to refer to the number of nodes in the network.

An *Erdős-Rényi* (ER) network (Erdős and Rényi 1960), also known as a Bernoulli network, was the first network model described. It has a single parameter, p, which is the probability that any particular edge is present in the network. To construct an ER network, one simply chooses $p \times \binom{n}{2}$ distinct pairs of nodes and draws an edge between each pair. The average degree of a node in an ER network is equal to $p \times n$.

A *Barabási-Albert* (BA) network (Barabási and Albert 1999), incorporates the phenomenon of *preferential attachment* seen in real life networks. This means that new connections are more likely to involve nodes which already have a high number of connections - popular nodes tend to become more popular. In the formulation we consider here, BA networks have two parameters, m and α, which are used to construct the network via the following algorithm. The network begins as a single node, then the remaining $n - 1$ nodes are added one at a time. Each time a new node is added, it is connected to m other nodes. The probability of one of these connections involving an existing node of degree k is proportional to k^α. Note that in the original formulation (Barabási and Albert 1999), there was an additional parameter m_0 which indicated the number of initial vertices in the network, and the only value of α considered was 1. For our purposes, we fix $m_0 = 1$ but allow α to vary. We refer to α throughout this work as the *preferential attachment power* or just attachment power.

Networks produced by the BA model are *scale free*, which means that the probability of a node having k connections is proportional to k^γ for some constant γ. This is also called a *power law* degree distribution. An important remark on notation must be made here: in network literature, the Greek letter α has often been used to refer to the power law exponent of the degree distribution, not the preferential attachment power. However, the paper defining the network model (Barabási and Albert 1999) uses γ for the power law exponent, and α for the attachment power, and we shall do likewise. This must be noted because there are reports in the literature of the power law exponent being as large as 4, which would seem to conflict with our choice (see subsection XX) to bound α above by 2. The reader should rest assured that it is the attachment power we are bounding, not the power law exponent.

WS and BA networks differ from ER networks in an important way: it is possible to explicitly calculate the likelihood of the ER model parameter p given an observed network, but this cannot be done for the other two models. This is because the probability of a particular edge occurring is always equal to p in an ER network, regardless of the presence or absence of any other edges. In other words, the edges can be viewed as a series of $\binom{n}{2}$ Bernoulli trials with success probability p, so that the likelihood of p given an observed graph with k edges is distributed as Binomial$\left(\binom{n}{2}, p\right)$.

1.3 Sequential Monte Carlo

Sequential Monte Carlo (SMC) is a statistical inference method which samples from a sequence of probability distributions in a fixed order (Del Moral, Doucet, and Jasra 2006). 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 (2006), 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 (2006) and other reviews of SMC (Doucet, De Freitas, and Gordon 2001), 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 (2006) use the variable π for the target distributions, but Doucet, De Freitas, and Gordon (2001) use π to indicate the importance distributions. We again follow the former and denote target distributions by π and importance distributions by η.

The basis of SIS is importance sampling (IS), which is a method of estimating summary statistics of distributions which are known only up to a normalizing constant, and therefore cannot be sampled from directly. That is, if π is such a distribution and f is any real-valued function, IS is concerned with estimating

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

where the integral is over the space on which π is defined, $\gamma(x)$ is known, 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 η, called the *importance distribution*, from which we are able to sample. Define the *importance weight* as the ratio ratio $w(x) = \gamma(x)/\eta(x)$. We can express the normalizing constant Z in terms of the importance weight and distribution, $Z = \int w(x)\eta(x)\mathrm{d}x$, and in turn write the 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 η, then $\eta(x)$ can be approximated at each of them by an empirical distribution. Since the remaining quantities f, γ, and w can all be evaluated pointwise, these are all the ingredients we need to obtain an estimate of $\pi(f)$. Although this is a simple and elegant approach, the drawback is that the variance of the estimate is proportional to the variance of the importance weights (Liu 2008), which may be quite large if η and γ are very different. Therefore, the practical use of IS on its own is limited, since it depends on finding an importance distribution which is similar to π, which we usually know very little about *a priori*.

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

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

where f_{i+1} is a function which yields a distribution when multiplied with π_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 (2001) discuss the case when π_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 π_1 is known explicitly, or we have enough information about it to find an adequate importance distribution. In the HMM example, π_1 was taken to be the prior distribution on the starting state.

1.4 Approximate Bayesian computation

1.4.1 Overview and motivation

ABC is a statistical method designed to fit complex models, which cannot be fit using more conventional methods, to observed data (Marin et al. 2012; Sunnåker et al. 2013; Beaumont 2010) We shall make this precise below, but to fully describe and motivate ABC, it is necessary to first explain exactly what we mean by "fitting" a model. We will illustrate this concept with one of the simplest and most ubiquitous models: linear regression.

these are review articles, Beaumont is specific to popgen, the others are general

A linear regression is a model of the relationship between some data $\mathbf{x} = (x_1, \ldots, x_n)$ and outcomes $\mathbf{y} = (y_1, \ldots, y_n)$. This model assumes that the outcomes are linearly related to the data, modulo some noise introduced by, say, measurement errors and environmental fluctuations. It other words, there is a constant β such that $y_i = \beta x_i + \varepsilon_i$, where ε_i is the error associated with the outcome y_i. If we make the additional assumption that the errors are normally distributed, then the model takes the form

$$y_i = \beta x_i + \mathcal{N}(0, \sigma^2),$$

where $\mathcal{N}(0, \sigma^2)$ indicates a normally distributed random variable with mean 0 and variance σ. We will denote the probability density function (pdf) of this random variable by $f_\mathcal{N}(\cdot \mid \mu, \sigma)$. In this formulation, the coefficient β and the error variance σ are the two parameters of the model. If the y_i are all independent, then for particular choice of β and σ, we can write down the probability density of observing $\mathbf{y}$ given $\mathbf{x}$.

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

The higher the value of this probability density, the more likely the observations $\mathbf{y}$ are given $\mathbf{x}$ under the model. This gives us a natural criterion for choosing the parameters: we want to pick β and σ which define a model where the probability of our observed data is as high as possible. When performing such an optimization for fixed data, the density function just written is also called the *likelihood* of the parameters,

$$\mathcal{L}(\beta, \sigma \mid \mathbf{x}, \mathbf{y}) = f(\mathbf{y} \mid \mathbf{x}, \beta, \sigma).$$

The particular β and σ which optimize $\mathcal{L}(\beta, \sigma \mid \mathbf{x}, \mathbf{y})$ are called the maximum likelihood

(ML) estimates.

ML inference makes use only of the data at hand, in this case $\mathbf{x}$ and $\mathbf{y}$, to estimate the parameters β and σ. However, it is frequently the case that the investigator has some additional *prior* information or belief about what β and σ are likely to be. For example, the instrument used to measure the $\mathbf{y}$ may have a known margin of error, or the sign of the slope may be obvious from previous experiments. A Bayesian approach to fitting this model would make use of this information by codifying the investigator's beliefs as a *prior distribution*, denoted $f(\beta, \sigma)$. The optimal parameters are then those which maximize the product of the prior and the likelihood,

$$f(\mathbf{y} \mid \mathbf{x}, \beta, \sigma) f(\beta, \sigma).$$

Bayes' theorem tells us that this product is proportional to the *posterior distribution* $f(\beta, \sigma \mid \mathbf{y}, \mathbf{x})$. The parameters which optimize the posterior, and hence also optimize the above product, are called the maximum *a posteriori* (MAP) estimates.

Now that we have defined what it means to fit a linear regression, there remains the question of how to go about finding this optimal β value. Naïvely, we could simply try many different β and σ values, calculate the likelihood of each, and choose those which yield the highest value. This approach, though basic, falls under the umbrella of *numerical optimization* of the likelihood function. Of course, there is a whole field devoted to more sophisticated methods for exploring the parameter space, but they all boil down to the basic idea of trying out different values and choosing those which give the highest likelihood.

In the case of linear regression, there is another method we could use. We know from calculus that the maximum of $\mathcal{L}$ occurs either on the boundary of its domain, or at a point where its partial derivatives with respect to β and σ are both zero. Though we will not show it here, it turns out to be straightforward to find the β and σ values where the partial derivatives are zero. In fact, these values co-incide with the least-squares estimates of β and σ, which minimize the sum of the squared $x_i - \beta y_i$. However, the method of directly minimizing the loss function with calculus is only applicable to a narrow class of simple models. For the majority, the likelihood function is too complicated to set the partial derivatives with respect to the parameters to zero and solve.

An approximate Bayesian computation approach to least-squares regression might be as follows. We will need to make an additional assumption: that the errors ε_i are normally distributed. That is, our linear model is now of the form

$$y_i = bx_i + \mathcal{N}(0, \sigma),$$

where $\mathcal{N}(0, \sigma)$ indicates a normal distribution with mean zero and variance σ. We do not know σ in advance, so this is another parameter we will need to estimate. There is nothing special about normal distributions - we could have chosen any other distribution we deemed appropriate.

The main idea of ABC is to replace the traditional likelihood-based criteria for model "goodness" with one based on simulated data generated by the model. Good models should be able to simulate data which closely resemble reality. In ABC, we are no longer interested in the posterior distribution, but rather in the distribution of model parameters which produce data sufficiently "close" to the real data. To formalize this, let $d(\cdot, \cdot)$ be a distance function on datasets, and ε be a tolerance level. The objective of ABC is to recover the distribution…

1.4.2 Algorithms for ABC

Approximate Bayesian computation 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 sequential Monte Carlo (SMC).

All of these approaches require some common elements. First, as with all Bayesian methods, we are required to specify a *prior distribution*, denoted π, 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 η, which computes a vector of hopefully informative summary statistics on a data set. Third, we need a distance function ρ 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 σ 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 σ be independent, a and b be normally distributed, and σ 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 ρ we take the standard Euclidian distance.

Rejection ABC is the simplest method, and also the one which was first proposed (Rubin et al. 1984). 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 θ 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 θ 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 θ from the prior distribution $\pi(\theta)$. It then proceeds iteratively: a new parameter vector θ^* is chosen according to a proposal distribution $q(\theta^* \mid \theta)$. The proposal distribution q is often taken to be a Gaussian centred at θ. Then θ^* is accepted as the new θ with probability

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

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 θ 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 θ, 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) (Sisson, Fan, and Tanaka 2007). 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.

Chapter 2

Body of Thesis

2.1 Objective

Our objective is to develop a method to estimate structural parameters of the contact network underlying an observed transmission tree.

2.1.1 Prior work

The present study is closely related to three groups of work with distinct objectives, which will be reviewed in detail below.

The first and largest group of related studies are phylodynamic investigations of epidemiological parameters such as transmission rate, recovery rate, and basic reproductive number (Pybus and Rambaut 2009; Volz, Koelle, and Bedford 2013). Like our work, 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 (Kendall 1948) or coalescent (Kingman 1982) models. Stadler *et al.* (Stadler et al. 2011) 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 (Stadler et al. 2011). However, the birth-death model is cannot tell us anything about population structure, as it assumes that every individual becomes infected at the same rate. Volz (**volz2012complex**) 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.

A second, more recent group of studies has evaluated the effect of network structure on transmission tree shape. The use of models which assume a *panmictic* (that is, homogeneously mixed) population in phylodynamics has become widespread, so it is natural that some researchers have investigated the effects of this assumption. In particular, since phylodynamic methods use phylogenies as their data source, several studies have examined the shapes of phylogenies arising from non-panmictic populations. O'Dea and Wilke (2010) simulated epidemics over networks with four types of degree distribution. They then estimated the Bayesian skyride (Minin, Bloomquist, and Suchard 2008) 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.* (**volz2009phylodynamics**). 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. simulated transmission trees over Erdős-Rényi (ER), Watts-Strogatz (WS), Barabási-Albert (BA), and full networks with fixed number of nodes and mean degree. They calculated Sackin's index of the simulated trees while varying the epidemic, network, and sampling parameters, and found that the relationship between these parameters and Sackin's index varies considerably among the different network models. 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.

Finally, a third group of studies has used Bayesian methods to infer a structural network parameter from observed infection and recovery times. This third group is most similar in aims to our own work; rather than using epidemiological observations, we employ viral phylogenies. The pioneering study in this group was by Britton and O'Neill (2002), who developed a Bayesian method to infer the p parameter of an ER network, along with the β and γ parameters of susceptible-infected (SI) model. Their method is able to use either infection and removal times, or removal times only. However, it is designed for only a small number of observations (15 and 42 cases in their applications), and their estimates of p for real outbreaks were mostly uninformative (95% confidence intervals [0.11-0.96] and 0.055-0.96]; p is bounded in [0, 1]). Groendyke, Welch, and Hunter (2011) 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. They also assumed that the entire network was infected, so that the network size was exactly 188. It is unclear how well their method would scale to networks with thousands of nodes.

Several studies (Little et al. 2014; Wang et al. 2015; Leigh Brown et al. 2011; Schneeberger et al. 2004) have found that the preferential attachment model provides a good fit to MSM networks.

The phylogenetic network may show little to no agreement with a contact data obtained through epidemiological methods (Yirrell et al. 1998; Resik et al. 2007).

2.2 Methods

2.2.1 Computer program

I implemented the adaptive SMC algorithm for ABC developed by Del Moral, Doucet, and Jasra (2012). The program was written in the *C* programming language. The *igraph* library (Csardi and Nepusz 2006) was used to generate and store contact networks and phylogenies. Judy arrays (Baskins 2004) were used for hash tables and dynamic programming arrays. The GNU scientific library (GSL) (Gough 2009) was used to generate random draws from probability distributions, and to perform the bisection step in the ABC-SMC algorithm.

For ease of exposition, we simplifiy the notation of Del Moral, Doucet, and Jasra (2012) 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, α). 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 ε 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.

I implemented a Gillespie simulation algorithm (Gillespie 1976) for simulating epidemics, and the corresponding transmission trees, over static contact networks. This method has been independently implemented and applied by several authors (*e.g.* O'Dea and Wilke 2010; Robinson et al. 2013; Leventhal et al. 2012; Groendyke, Welch, and Hunter 2011). Groendyke, Welch, and Hunter (2011) 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. The individual nodes and edges of

G follow the dynamics of the susceptible-infected-recovered (SIR) model (Kermack and McKendrick 1927). Each directed edge $e = (u, v)$ in the network is associated with a transmission rate β_e, which indicates that, once u becomes infected, the waiting time until u infects v is distributed as Exponential(β_e). Note that v may become infected before this time has elapsed, if v has other incoming edges. v also has a removal rate γ_v, so that the waiting time until removal of v from the population is Exponential(γ_v). Removal may correspond to death or recovery with immunity, or a combination of both, but in our implementation recovered nodes never re-enter the susceptible population. We define a *discordant edge* as an edge (u, v) where u is infected and v has never been infected.

To describe the algorithm, we introduce some notation and variables. Let $\text{in}(v)$ 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 β be the total transmission rate over all discordant edges, and γ be the total removal rate of all infected nodes,

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

The variables S, I, R, D, β, and γ are all updated as the simulation progresses. When a node v becomes infected, it is deleted from S and added to I, any formerly discordant edges in $\in (v)$ are deleted from D, and edges in $\text{out}(v)$ 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. In both cases, the variables β and γ are updated to reflect the changes. Since these updates are straightforward, we do not write them explicitly in the algorithm.

The Gillespie simulation algorithm is given as Algorithm 2.2.1. The transmission tree T is simulated along with the epidemic. We keep a map called tip, which maps infected nodes in I to the tips of T. The simulation continues until either there are no discordant edges left in the network, or we reach a user-defined cutoff of time ($t_{\max}$) or number of infections ($I_{\max}$). We use the notation Uniform$(0, 1)$ to indicate a number drawn from a uniform distribution on $(0, 1)$, and likewise for Exponential(λ). The combined number of internal nodes and tips in T is denoted $|T|$.

Phylogenetic kernel and normalized lineages-through-time

The tree kernel developed in (Poon et al. 2013) 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. I implemented the fast algorithm developed in (Moschitti 2006), which first enumerates all pairs of subtrees with the same number of leaf children,

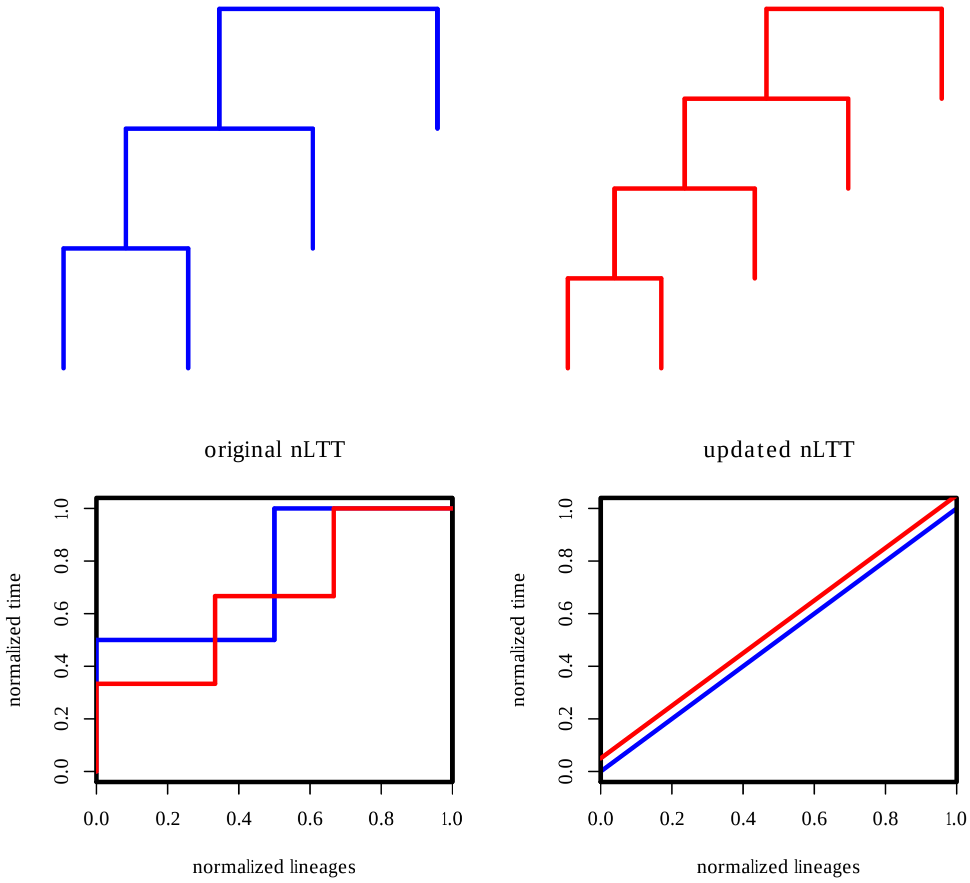
Algorithm 1 Simulation of an epidemic and transmission tree over a contact network

```
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 $T$            ▷ $|T|$ increased by 2
            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
```

and then computes the kernel by dynamic programming.

In addition, we implemented a modified version of the normalized lineages-through-time statistic developed in (Janzen, Höhna, and Etienne 2015), which uses piecewise linear functions instead of step functions for the lineages-through-time plots. This modification is to address a potential inconsistency for trees of different sizes, illustrated in Figure 2.1.

Figure 2.1: Comparison of original formulation of normalized lineages-through-time, developed in (Janzen, Höhna, and Etienne 2015), with our modified version using linear interpolation. Here, the red and blue trees both have uniformly spaced branching times. Using step functions (left), the nLTT of the two trees is non-zero due to the differing numbers of internal branches. Using linear interpolation, the nLTT is zero (right). The lines on the right graph have been offset for visibility.



Next, we calculate the next tolerance ε^*. 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 1.4.2), the idea of sequential Monte-Carlo is to begin with the prior distribution π, progress smoothly through a series of intermediate distributions $\pi_1, \ldots, \pi_{n-1}$, and eventually arrive at the target posterior distribution π_n. In the kth iteration, the distribution π_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 kernel reaches a user-defined threshold. Following a heuristic applied by the authors (Del Moral, Doucet, and Jasra 2012), we used the latter stopping criterion, accepting the SMC approximation to the posterior when the MCMC acceptance rate dropped below 1.5%.

2.2.2 Simulation experiments

Identification of separable parameters in kernel space

Recall that our approximate Bayesian computation approach to fitting contact network models involves simulating transmission trees under a wide variety of parameter values, and then comparing these simulated trees to the true transmission tree. Values which produce trees similar to the observed transmission tree are distinguished as more likely than values which produce trees very different from the truth. In order for this type of analysis to succeed, it is critical that different parameter values produce different looking trees. Otherwise, if many different values produce trees which are too similar to each other, it will be impossible to distinguish which value is most consistent with the real tree. Just as importantly, trees simulated with similar parameter values must be similar to each other. In mathematical terms, we require the trees simulated from distinct parameter values to be *separable* in tree space. The concept of separability is illustrated in Figure 2.2.

Before undertaking a complete ABC analysis, I analysed four simple contact network models to determine whether their parameters could be separated in tree kernel space. The four models are described in detail in subsection 1.2.2 of the introduction. Briefly, they are: random networks, where each possible contact has a fixed probability of occuring; preferential attachment networks, where highly connected nodes tend to attract more contacts; small world networks, where nodes are connected to their immediate neighbours and the occasional far-flung contact; and full networks, where every possible connection is present. I will describe here only the procedure and results for the preferential attachment networks. The details of the other three types of graph can be found in the supplemental materials. A graphical schematic of the analysis undertaken here is given in Figure 2.2.2.

The method of testing for separability was described previously in (Poon 2015), but I will reiterate it here for completeness. As a concrete example, consider the attachment power parameter α of the preferential attachment networks. This parameter describes the

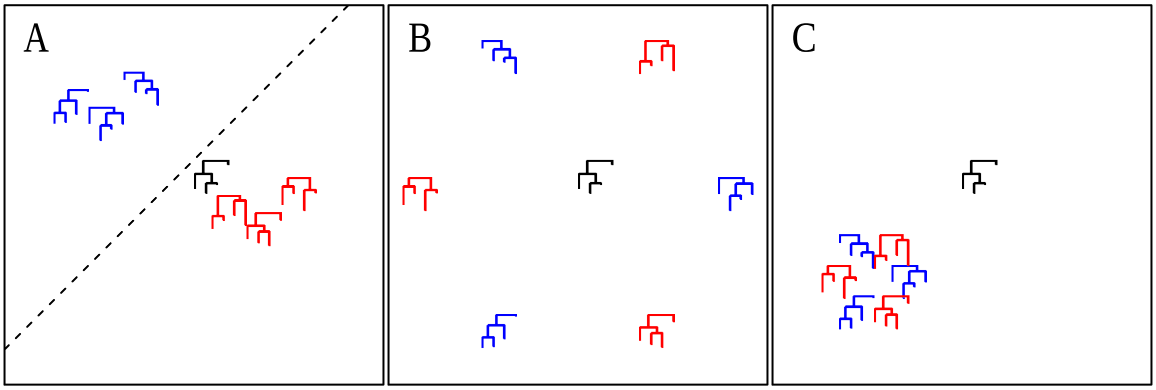


Figure 2.2: Separable versus non-separable parameters in kernel space. Trees have been simulated under three sets of parameters, represented as blue, red, and green. An observed tree is shown in black. Trees are layed out such that the distance between two trees corresponds to their similarity. In panel A, trees from the same parameter set are similar to each other, but different from other trees. The true tree is most consistent with the red parameters. In panels B and C, trees from the same parameter set are not similar to each other (B), or trees from different parameter sets are similar to each other (C). It's difficult to say which parameters the true tree is most consistent with.

strength of attraction to highly connected nodes and is bounded below by zero indicating no extra attraction. By qualitative observation, I determined that a power of 2.0, which produced networks with very few "hub" nodes with extremely high degree, was a suitable upper bound (see Figure 2.2.2). Therefore, I chose to test the values 0.5, 1.0, and 1.5 for separability. The other parameters were fixed: the number of nodes in the network was 5000, and the mean degree of each node in the network was four. As discussed in subsection 1.2.2, this is the smallest mean degree value for preferential attachment networks which produces networks which are more than trees.

For each of the values of α, I generated 100 networks on 5000 nodes. An epidemic was simulated over each network (see subsection X) until 1000 nodes were infected, and 500 of those infected nodes were sampled to form a transmission tree. This resulted in 300 total simulated transmission trees - 100 for each of the three values of α. The data generation steps are shown on the left side of Figure 2.2.2. Next, I computed the tree kernel (Poon et al. 2013) (see subsection 1.1.4) for each pair of trees. These values were placed into a 300 × 300 kernel matrix, where the value at the (i, j)th position was the tree kernel of the ith and jth trees.

The tree kernel provides a pairwise similarity score between two trees. The higher the kernel score, the more similar the trees are to each other. Therefore, to have separability, we need trees simulated with the same value of α to have high kernel scores with each other, but low kernel scores with trees from different α values. We can visually check whether or not this is true by laying out the trees as points on a graph, in such a way that trees

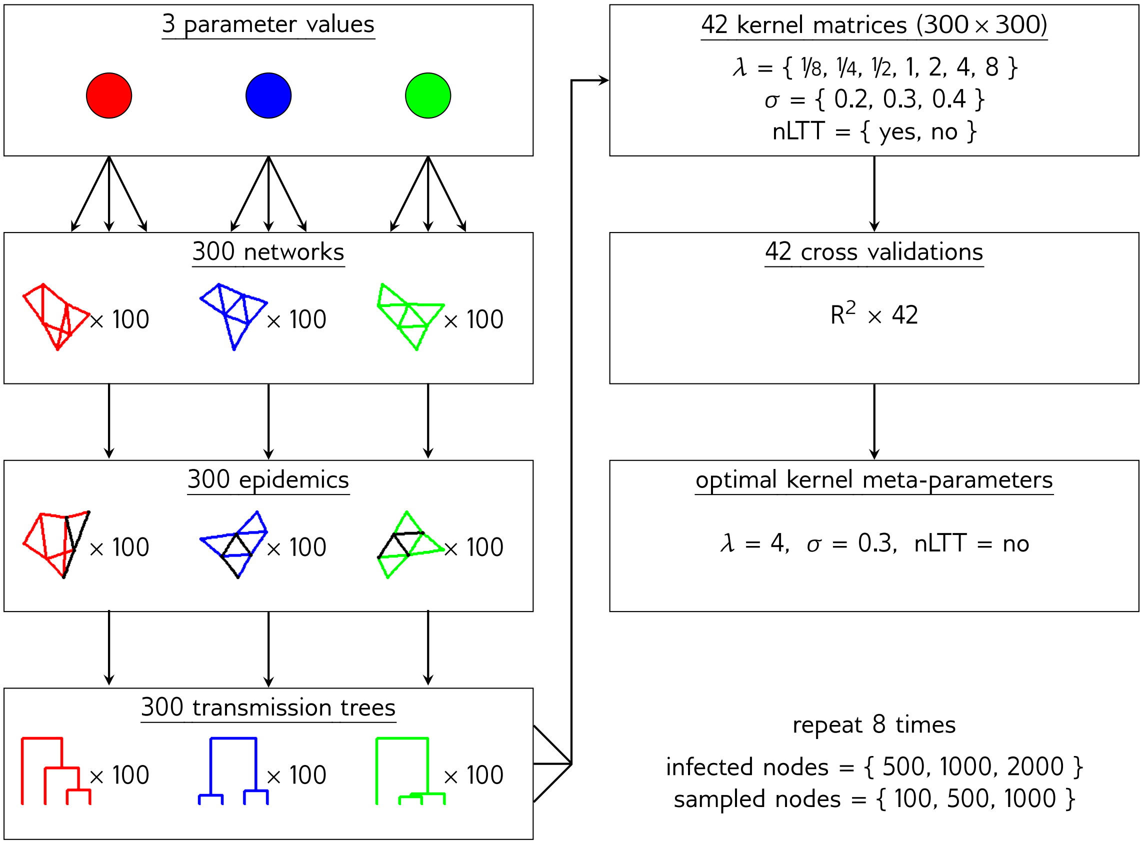


Figure 2.3: Schematic of first set of simulation experiments to determine separable parameters in kernel space, along with optimal tree kernel meta-parameters.

with high scores are close to each other, but trees with low scores are far apart. This is accomplished by performing a kernel principal components analysis (kPCA) (Schölkopf, Alexander Smola, and Müller 1998) on the kernel matrix. Briefly, ordinary principal components analysis (PCA) finds a lower dimensional representation of points in a high dimensional space which preserves as much of the variation in the data as possible. kPCA performs the same task, but using the dot products of each pair of points as input, instead of the data points themselves. A two-dimensional kPCA projection of the simulated trees is shown in Figure 2.3.1. The scenario just described - 500 samples from 1000 infected nodes - is the central panel. To ensure that the method could be used in a variety of contexts, the same analysis was performed with 500 and 2000 infected nodes, as well as 100 and 1000 sampled tips.

The next step was to quantify how well trees simulated with different α values could be distinguished from each other. As described previously (Poon 2015), this was done by assessing the accuracy of a support vector machine regression (SVR) (Alex Smola and Vapnik 1997). Briefly, an SVR operates by finding a hyperplane (in two dimensions, a line) such that the deviation of most of the data points from the line is less than some prescribed threshold. Points further away than this threshold are ignored in the model. I performed

1000 replicate 2-fold cross-validations of an SVR predicting preferential attachment power on the simulated trees, using the *ksvm* function from *kernlab* (Karatzoglou et al. 2004). That is, the SVR was trained on a random subset of 150 trees, and then used to predict α of the remaining 150 trees. The predictions were correlated against the true values of α to obtain an R^2, and this procedure was repeated 1000 times with different subsets of trees.

The cross-validation had the dual purpose of providing a means to selecting the optimal meta-parameters to the tree kernel (see subsection 1.1.4) - they are be those which provide the highest average R^2. To this end, the cross-validation was repeated for several values of λ and σ as shown in Figure 2.2.2 Each combination was evaluated with and without multiplying the tree kernel by the normalized lineages-through-time (nLTT) statistic, and was further repeated for the scenarios with differing numbers of infected and sampled nodes described above. To ensure that the tree kernel was the most appropriate similarity measure to use for ABC, we also computed the R^2 of α against Sackin's index, a widely used tree balance statistic (see subsection 1.1.4), by the same cross-validation procedure.

Grid search

The previous set of simulations were intended to investigate which contact network parameters could and could not be inferred by examining transmission trees. However, they tell us nothing about the accuracy or precision we might expect when inferring those parameters numerically. As illustrated in Figure 2.5, the ideal situation is one where we are accurate and precise, and the worst situation is when we are precise but not accurate.

Figure 2.6 shows a schematic of this experiment. A number of representative values, here denoted k, were chosen for the parameter of interest. In the case of preferential attachment power α, I chose $k = 8$ testing values, namely $0, 0.25, \ldots, 2.0$. For each of these values, ten networks on 5000 nodes were generated, an epidemic of 1000 nodes was simulated over each, and a transmission tree of 500 tips was sampled. The resulting $10 \times k$ simulated transmission trees were referred to as the *testing trees*. In addition, I chose a further $n \gg k$ training values spanning the range of the parameter. For α, these were $0, 0.01, \ldots, 2.0$. Fifteen trees were simulated in the same manner for each of these values, referred to here as *training trees*.

For each of the testing trees, I computed the tree kernel with all of the training trees. This resulted in 15 kernel scores per training value. The training value with the highest median kernel score was used as a point estimate of the testing value. The kernel scores were then normalized to lie in [0, 1], and an interval was found which contained 95% of the area under the curve and was of minimal width. This was used as a 95% confidence interval for the parameter.

Approximate Bayesian computation

Our final set of simulation experiments was to test the ABC-SMC algorithm on simulated data. For each replicate, we generated a network on 5000 nodes, simulated an epidemic over that network until 1000 nodes were infected, and sampled either 500 or all 1000 of those nodes to form a transmission tree. Priors were specified as follows: for the total number of nodes in the network, $\text{Uniform}(1000, 10000)$; for the number of infected nodes, $\text{Uniform}(500, 2000)$; for the preferential attachment power, $\text{Uniform}(0, 2)$. The mean degree of nodes in the network was fixed at either 4, 10, or 16. The algorithm was run with 1000 particles, and was stopped when the MCMC acceptance probability dropped below 1.5%.

2.2.3 Applications

HIV in British Columumbia

2.3 Results

2.3.1 Separable parameters

Our first simulation-based experiment was designed to determine which network model parameters were potentially estimable by kernel-ABC, and to find the best set of kernel meta-parameters to carry through to future experiments. To do this, we simulated trees under three distinct values of each parameter of the BA model, and evaluated the accuracy of a kernel-SVM classifier for the parameter.

In Figure 2.3.1, we show that trees simulated under different values of α are visibly quite distinctive. In particular, higher values of α result in networks with a small number of highly connected nodes (see Figure 2.2.2) which, once infected, are likely to transmit to many other nodes. This results in a more unbalanced, ladder-like structure in the phylogeny. Kernel-principal components analysis (PCA) projections show that all three α values are well separated from each other under several different sampling scenarios (Figure 2.3.1). With smaller trees, it becomes harder to visually disinguish $\alpha = 0.5$ from $\alpha = 1.0$ using the first two principal components.

With suitably chosen meta-parameters, the accuracy of the kernel-SVM classifier for α was very high under a variety of prevalence and sampling scenarios (Figure 2.3.1). 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. The fact that Sackin's index is informative of α was unsurprising, given the clear differences in tree balance observed under different α values. The nLTT statistic (Janzen, Höhna, and Etienne 2015) did not improve the classifier, but rather slightly reduced the R^2 in most cases.

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, α = 1.0, and m = 2. As shown in Figure 2.3.1, 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 2.3.1). Moreover, the use of the nLTT statistic improved classification accuracy by a small amount, in contrast to the results for α. 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.

2.3.2 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 α on the grid closest to the true values, as shown in Figure 2.12. However, there was a much stronger spike in kernel scores near the true value for α = 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 2.13, while the estimates are generally close to the true value, they are much closer for α = 1.25 than for the other values.

2.3.3 Accuracy of estimates with full ABC

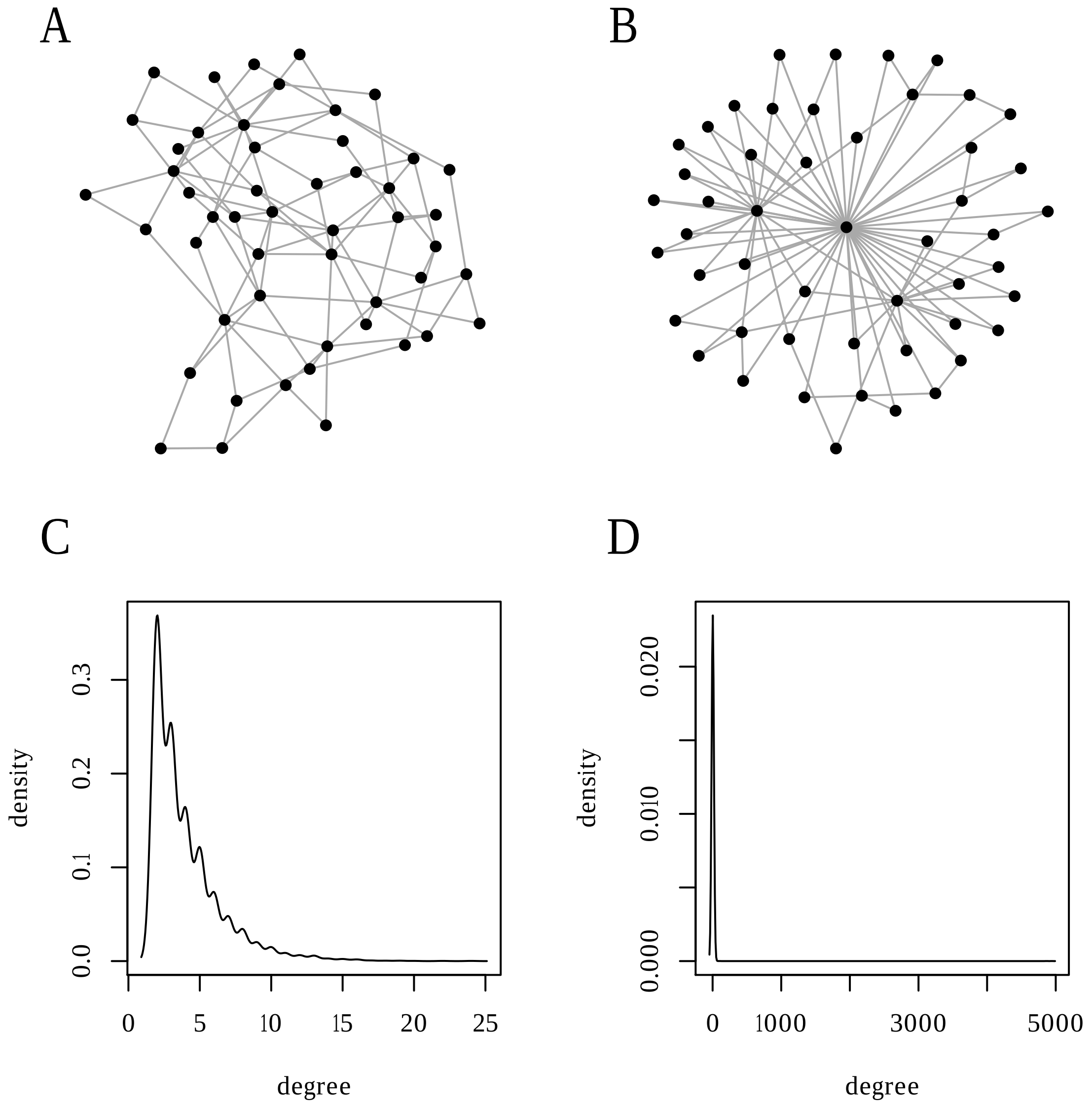


Figure 2.4: Qualitative justification for choice of zero and two as lower and upper bounds on preferential attachment power α. (A) Preferential attachment network on 50 nodes with $\alpha = 0$, the lower bound enforced by the model. (B) Preferential attachment network on 50 nodes with $\alpha = 2$, where a few nodes have very high degree but the majority have very low degree. (C) Density plot of node degrees in a 5000-node network with $\alpha = 0$. The maximum degree of a node in this network was 24. (D) Density plot of node degrees in a 5000-node network with $\alpha = 2$. The maximum degree of a node in this network was 4942.

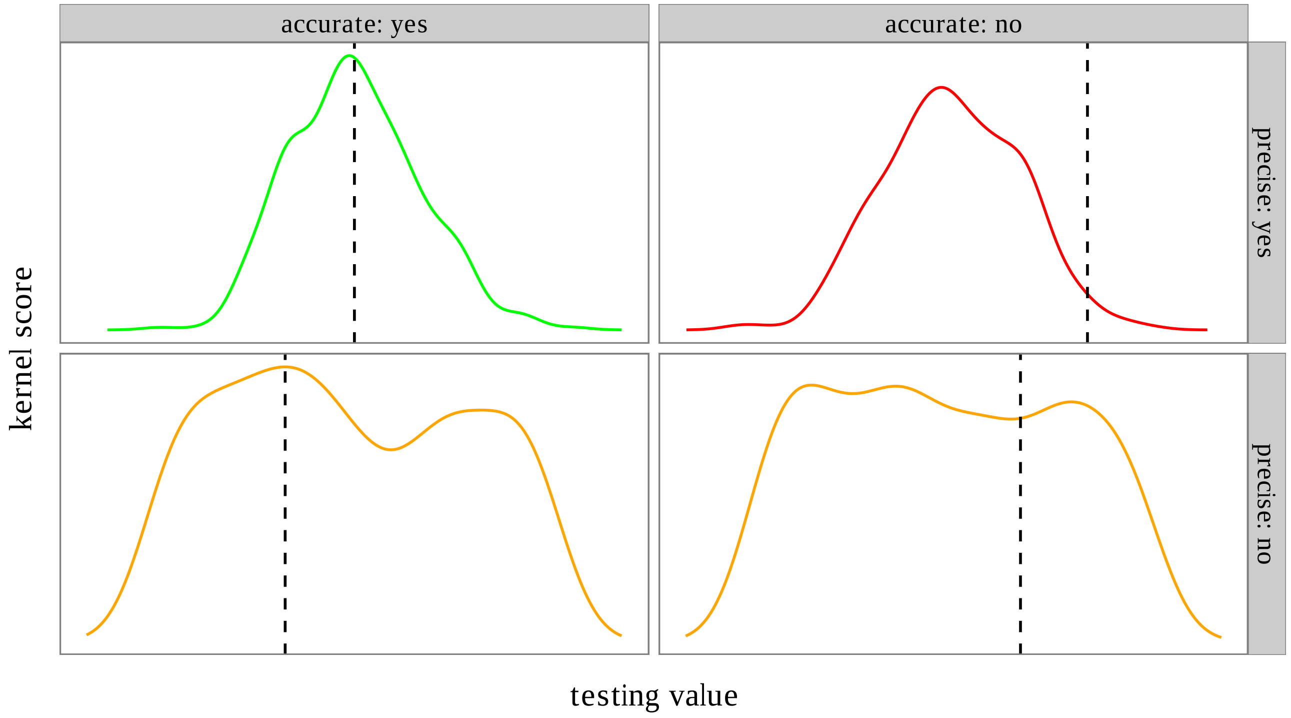


Figure 2.5: Illustration of accurate vs. precise kernel score estimates

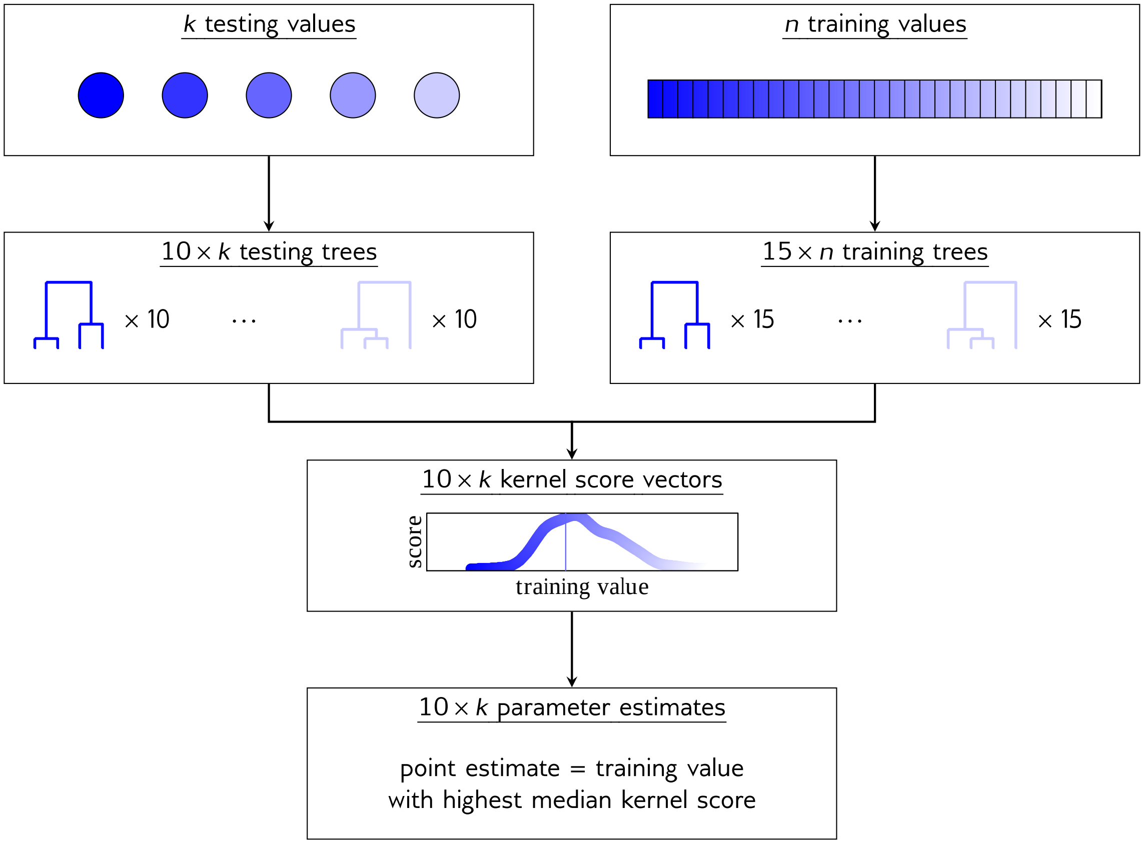

k testing values
n training values
10 × k testing trees
× 10
⋯
× 10
15 × n training trees
× 15
⋯
× 15
10 × k kernel score vectors
score
training value
10 × k parameter estimates
point estimate = training value
with highest median kernel score

Figure 2.6: Schematic of grid search simulation experiments.

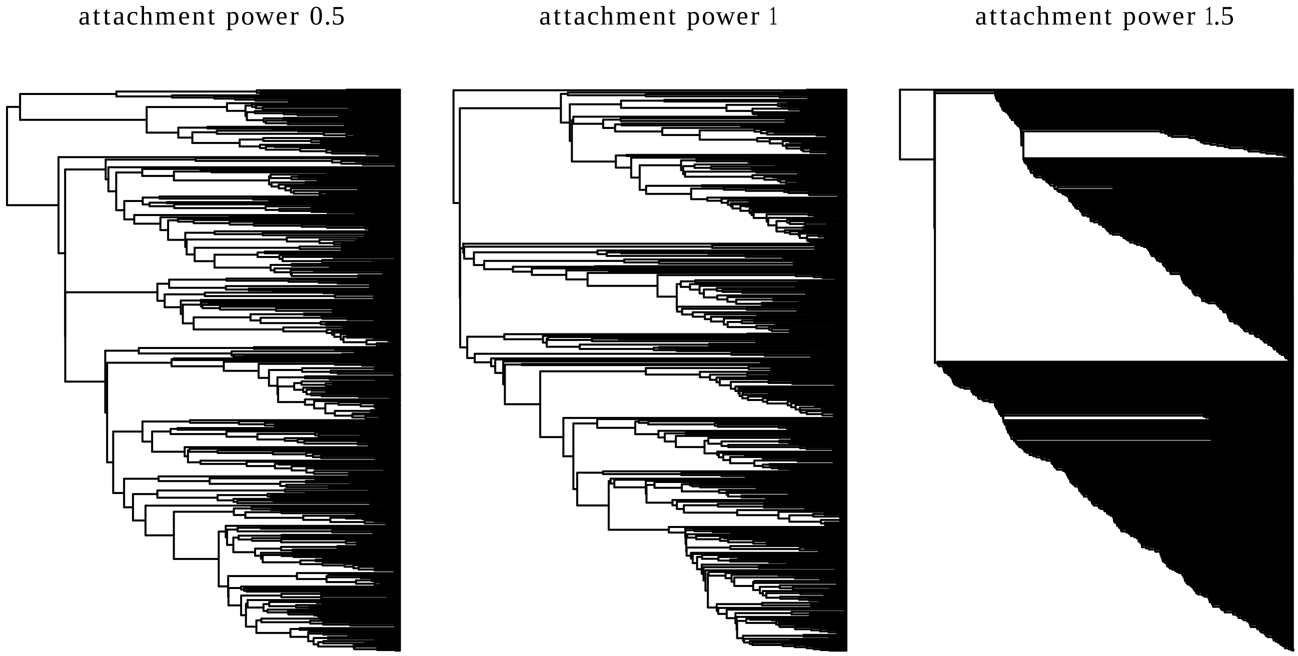


Figure 2.7: Epidemics of 1000 infected were simulated on BA networks of 5000 nodes, with α equal to 0.5, 1.0, or 1.5. Transmission trees were created by sampling 500 infected nodes. Higher α values produce networks with a small number of highly-connected nodes, which results in a highly unbalanced, ladder-like tree structure.

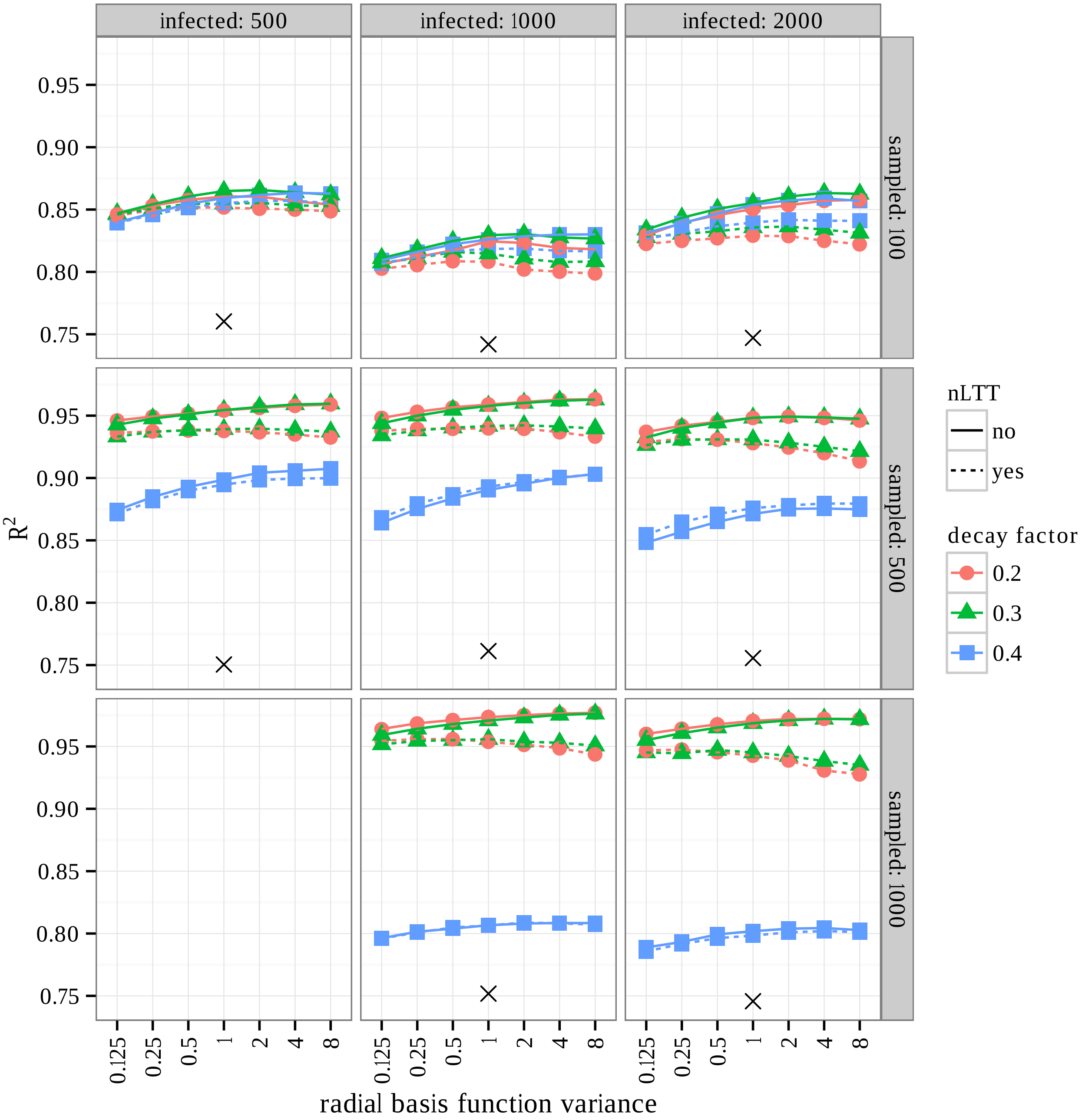


Figure 2.8: Cross-validation performance of kernel support vector machine classifier for preferential attachment power.

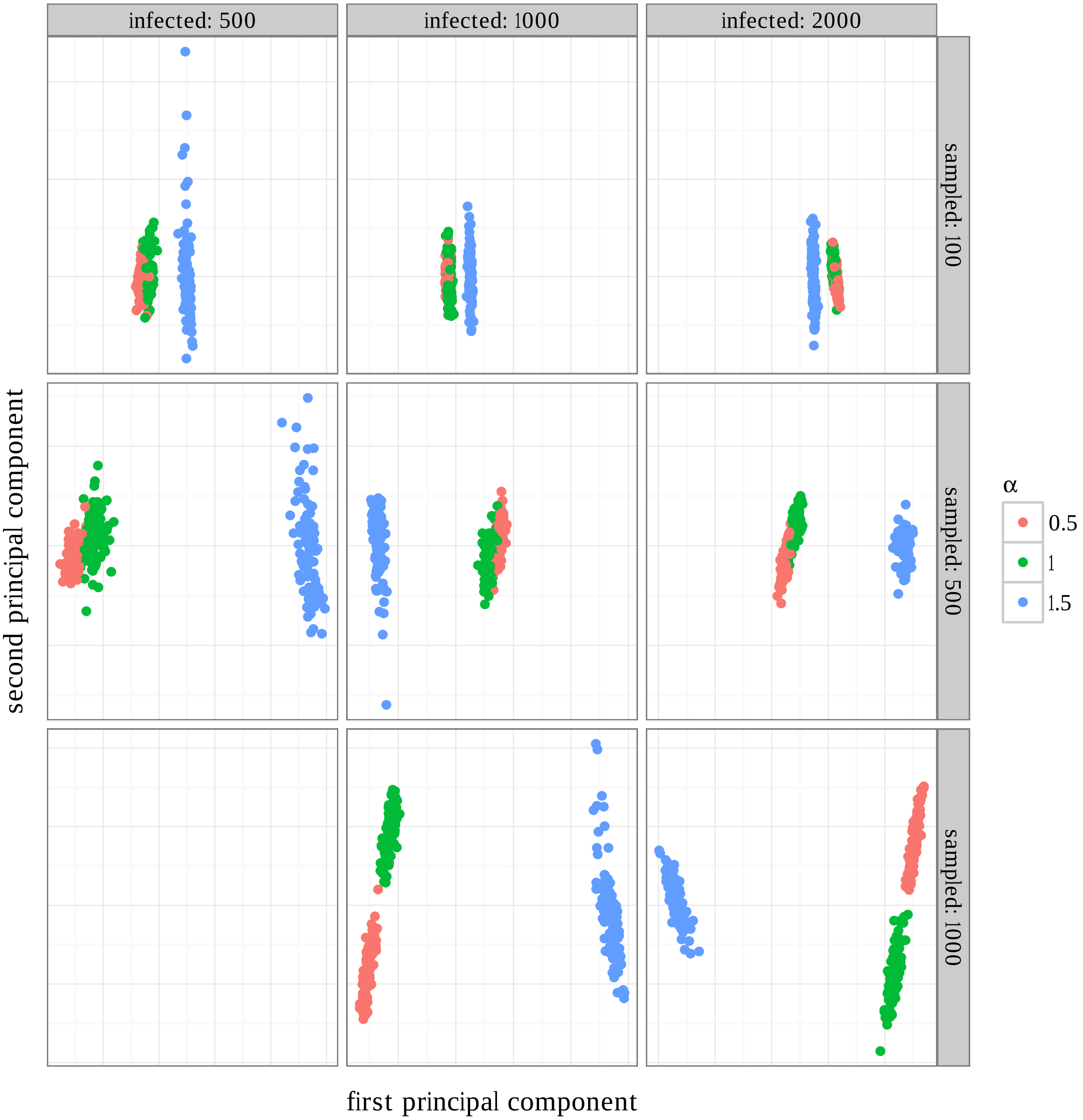


Figure 2.9: Projection of the kernel matrix for different preferential attachment power values onto its first two principal components, for eight simulation scenarios. Each point corresponds to a simulated transmission tree, and is coloured by preferential attachment power. Facets are number of infected 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.

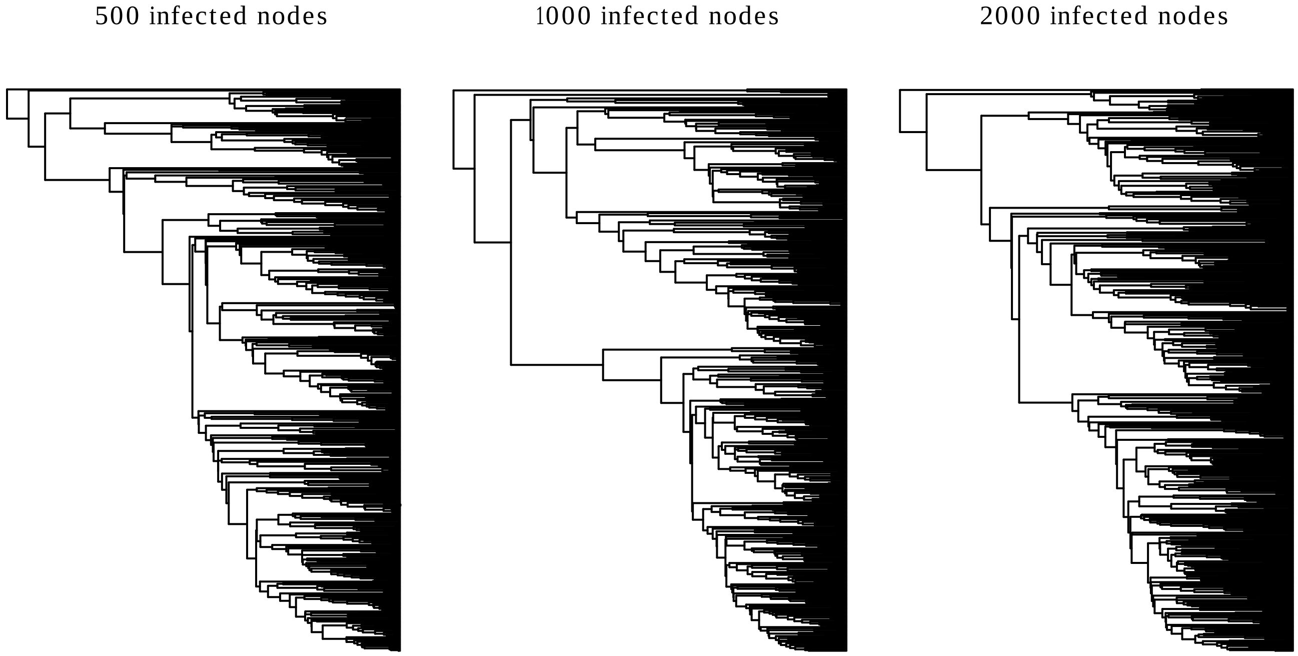


Figure 2.10: Epidemics were simulated on networks of size 5000 until $I = 500$, 1000, or 2000 nodes were infected. When scaled to the same height, there is no immediately visible distinction in shape between trees simulated under different I values.

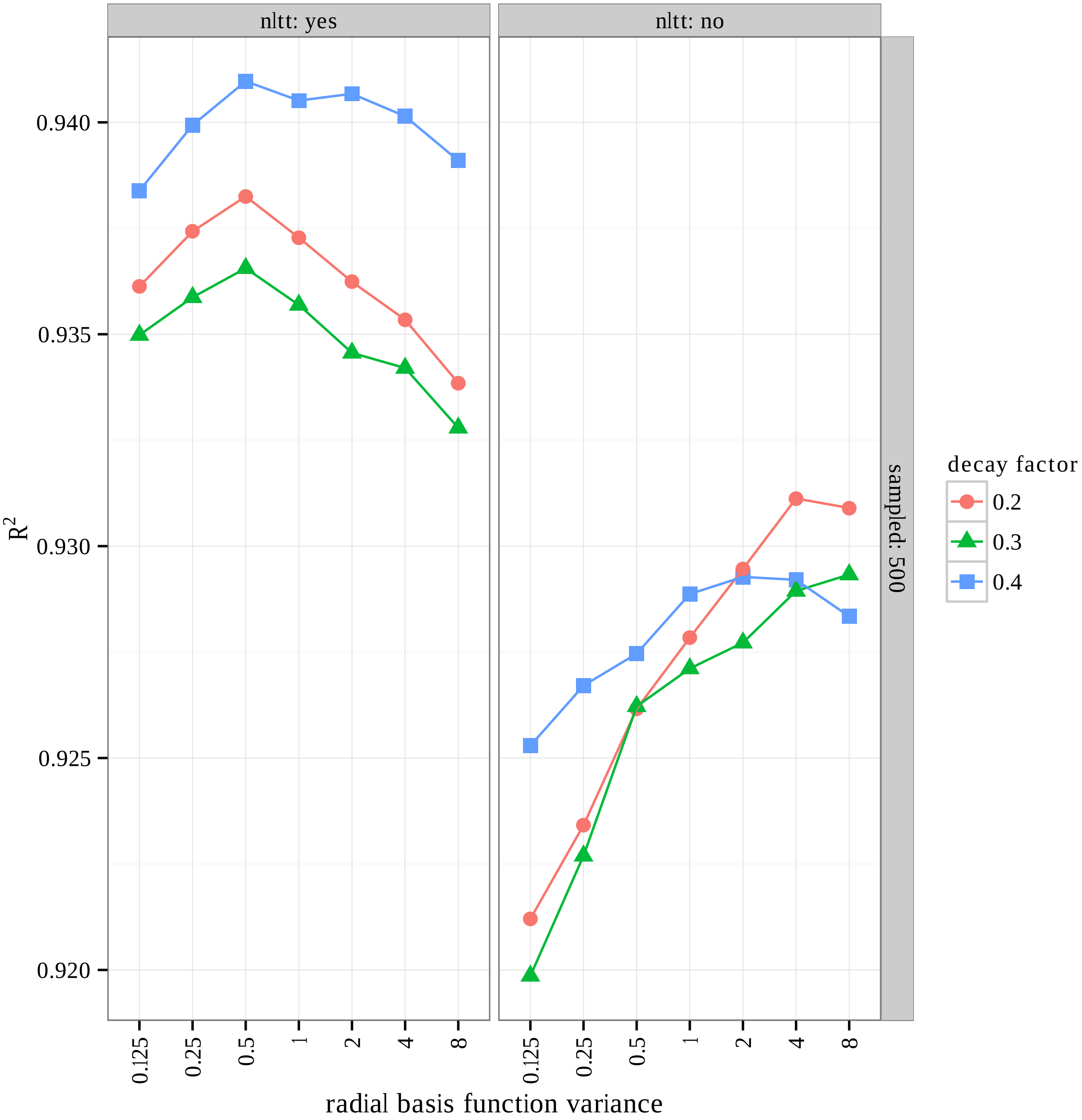


Figure 2.11: For BA networks of size $N = 5000$, a kernel-SVM classifier is very accurate at predicting I. Epidemics were simulated until $I = 500$, 1000, or 2000 nodes were infected, and either 100 or 500 nodes were sampled for inclusion in a transmission tree. The parameters of the BA network were $\alpha = 1.0$ and $m = 2$. A Sackin's index-based classifier performed very poorly on these data (R^2 = TODO for 100-tip trees and blah for 500-tip trees, not shown).

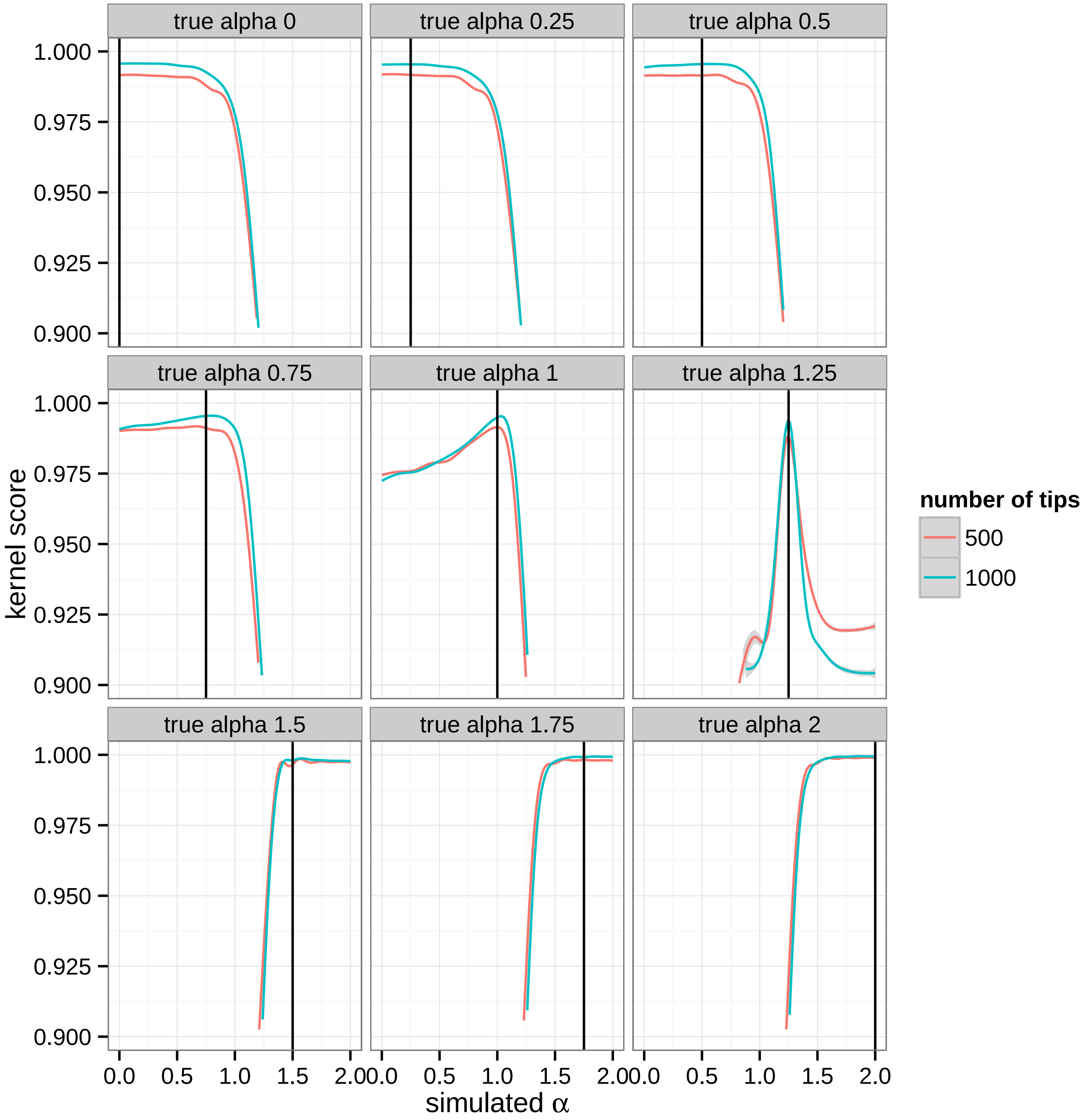


Figure 2.12: Grid search kernel scores for testing trees simulated under various α values. All epidemics had $I = 1000$ infected nodes, on BA networks of size $N = 5000$ with m fixed at 2. Colours indicate the number of sampled tips.

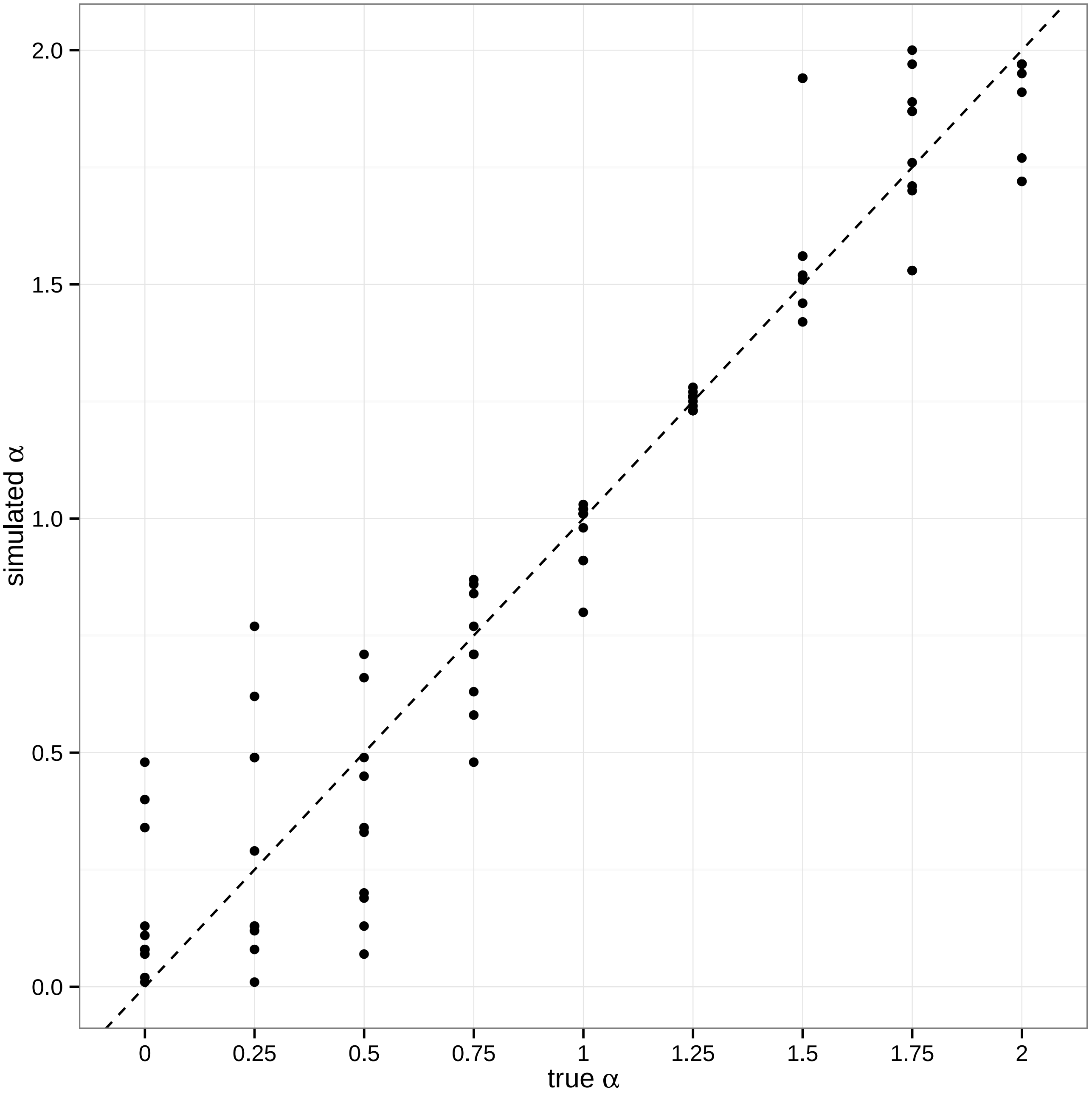


Figure 2.13: Marginal estimates of α obtained with grid search. Training trees were simulated on a narrowly spaced grid of α values, and compared to testing trees using the tree kernel. The α value in the grid with the highest median kernel score was taken as the point estimate for the testing tree. These point estimates are shown as black dots. The dashed line is the identity.

Chapter 3

Conclusion

Bibliography

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

Baskins, Doug (2004). *Judy arrays*.

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

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

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

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

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

Collins, Michael and Nigel Duffy (2002). "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, pp. 263–270.

Cottam, Eleanor M et al. (2008). "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, pp. 887–895.

Coyne, Jerry A and H Allen Orr (2004). *Speciation*. Vol. 37. Sinauer Associates Sunderland, MA.

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

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

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

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

Doucet, Arnaud, Nando De Freitas, and Neil Gordon (2001). "An introduction to sequential Monte Carlo methods". In: *Sequential Monte Carlo methods in practice*. Springer, pp. 3–14.

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

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

Erdős, Paul and Alfred Rényi (1960). "On the evolution of random graphs". In: *Publ. Math. Inst. Hungar. Acad. Sci* 5, pp. 17–61.

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

Gough, Brian (2009). *GNU scientific library reference manual*. Network Theory Ltd.

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

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

Haeckel, Ernst Heinrich (1866). *Generelle Morphologie der Organismen*. Vol. 2. Verlag von Georg Reimer.

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

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

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

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

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

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

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

Kermack, William O and Anderson G McKendrick (1927). "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, pp. 700–721.

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

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

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

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

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

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

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

Liu, Jun S (2008). *Monte Carlo strategies in scientific computing*. Springer Science & Business Media.

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

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

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

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

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

Nei, Masatoshi and Sudhir Kumar (2000). *Molecular evolution and phylogenetics*. Oxford University Press.

O’Dea, Eamon B and Claus O Wilke (2010). “Contact heterogeneity and phylodynamics: how contact networks shape parasite evolutionary trees”. In: *Interdisciplinary perspectives on infectious diseases* 2011.

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

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

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

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

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

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

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

Schölkopf, Bernhard, Alexander Smola, and Klaus-Robert Müller (1998). “Nonlinear component analysis as a kernel eigenvalue problem”. In: *Neural computation* 10.5, pp. 1299–1319.

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

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

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

Smola, Alex and Vladimir Vapnik (1997). “Support vector regression machines”. In: *Advances in neural information processing systems* 9, pp. 155–161.

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

Sunnåker, Mikael et al. (2013). “Approximate bayesian computation”. In: *PLoS Comput Biol* 9.1, e1002803.

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

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

Wang, Xicheng et al. (2015). “Targeting HIV Prevention Based on Molecular Epidemiology Among Deeply Sampled Subnetworks of Men Who Have Sex With Men”. In: *Clinical Infectious Diseases*, p. civ526.

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

Yule, G Udny (1925). “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, pp. 21–87.