PHYLOGENETIC ESTIMATION OF CONTACT NETWORK PARAMETERS WITH APPROXIMATE BAYESIAN COMPUTATION

by

Rosemary Martha McCloskey

B.Sc., Simon Fraser University, 2014

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF SCIENCE

in

The Faculty of Graduate and Postdoctoral Studies

(Bioinformatics Graduate Program)

THE UNIVERSITY OF BRITISH COLUMBIA
(Vancouver)

August 2016

Abstract

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

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

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

Finally, the Barabási-Albert model was fit to ~~six~~ eleven real world HIV datasets, and substantial heterogeneity in the parameter estimates was observed. ~~Point estimates~~ Posterior means for the preferential attachment power were all sub-linear and ranged from ~~0.06 to 1.05~~ 0.27 to 0.9. Point estimates of the strength of preferential attachment were higher in injection drug user populations, potentially indicating that high-degree "hub" nodes may play a role in epidemics among this risk group. Our results underscore the importance of considering contact structures when ~~performing phylodynamic inference~~ investigating viral outbreaks.

Preface

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

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

Use of the BC data is in accordance with an ethics application that was reviewed and approved by the UBC/Providence Health Care Research Ethics Board (H07-02559). Rosemary M. McCloskey completed the Tri-Council Policy Statement: Ethical Conduct for Research Involving Humans Course on Research Ethics (TCPS 2: CORE) tutorial on June 8, 2016.

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

Table of Contents

List of Tables

List of Figures

List of Symbols

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

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

N total number of nodes in a contact network.

Θ parameter space for a given mathematical model.

α preferential attachment power parameter in Barabási-Albert networks.

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

n number of particles used for sequential Monte Carlo.

List of Abbreviations

ABC approximate Bayesian computation.

BA Barabási-Albert.

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

HIV human immunodeficiency virus.
HMM hidden Markov model.

IDU injection drug users.
IS importance sampling.

LTT lineages-through-time.

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

nLTT normalized lineages-through-time.

PA preferential attachment.
POSIX Portable Operating System Interface.

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

TasP treatment as prevention.

WS Watts-Strogatz.

Acknowledgements

Chapter 1

Introduction

1.1 Objective

The spread of a disease is most often modelled by assuming either a homogeneously mixed population [1, 2], or a population divided into a small number of homogeneously mixed groups [3]. This assumption, also called *mass action* [4], or *panmixia*, implies that any two individuals in the same compartment are equally likely to come into contact making transmission possible at some predefined rate. Although this provides a reasonable approximation in many cases [5], the error introduced by assuming a panmictic population can be substantial when significant contact heterogeneity exists in the underlying population [6–8]. Contact network models provide an alternative to compartmental models which do not require the assumption of panmixia. In addition to more accurate predictions, the parameters of the networks themselves may be of interest from a public health perspective. For example, certain vaccination strategies may be more or less effective in curtailing an epidemic depending on the underlying network's degree distribution [9, 10]. Phylodynamic methods, which link viruses' evolutionary and epidemiological dynamics , have been used to fit many different types of models to phylogenetic data [11, 12]. However, these models generally assume a panmictic population. The primary objective of this work is *to develop a method to fit contact network models, and thereby relax the assumption of homogeneous mixing, in a phylodynamic framework.*

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

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

where $f(T \mid \theta)$ is the likelihood of the parameters given T, $\pi(\theta)$ is the prior on θ, and Θ is the space of possible model parameters. The denominator on the right-hand side is the marginal probability of T which acts as a normalizing constant on the posterior (see **??** for a review of mathematical modeling and Bayesian inference, including definitions of these concepts). As we shall show (**??**), estimating this distribution presents computational challenges beyond those usually encountered in Bayesian inference. Both the likelihood $f(T \mid \theta)$ and the normalizing constant seem to be intractable, which rules out the use of most common maximum likelihood and Bayesian methods.

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

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

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

Depending on the network model studied, it is possible that eq. (1.1) could be simplified into a tractable expression. An alternative to likelihood-based methods, which could be applied to any network model, is provided by approximate Bayesian computation (ABC) [14–17]. All of the ingredients required to apply ABC to this problem are readily available. Simulating networks is straightforward under a variety of models. Epidemics on those networks, and the corresponding transmission trees, can also be easily simulated. As mentioned above, contact networks can profoundly affect transmission tree shape. Those shapes can be compared using a highly informative similarity measure called the "tree kernel" [18]; similar kernel functions have been demonstrated to work well as distance functions in ABC [19] . ABC can be implemented with several algorithms, but sequential Monte Carlo (SMC) has advantages over others, including improved accuracy in low-density regions and parallelizability [20]. A recently-developed adaptive algorithm requiring minimal tuning on the part of the user makes SMC an even more attractive approach [21]. In summary, our method to infer contact network parameters will combine the following: stochastic simulation of epidemics on networks, the tree kernel, and adaptive ABC-SMC. Our method will expand on the framework developed by [22], who combined ABC with the tree kernel to infer parameters of population genetic models from viral phylogenies using Markov chain Monte Carlo (MCMC).

Empirical studies of sexual contact networks have found that these networks tend to be scale-free [23–26], meaning that their degree distributions follow a power law (although there has been some disagreement, see [6, 27]). Preferential attachment has been postulated as a mechanism by which scale-free networks could be generated [28]. The Barabási-Albert (BA) model [28] is one of the simplest

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

Due to its high global prevalence and fast mutation rate, human immunodeficiency virus (HIV) is one of the most commonly-studied viruses in a phylodynamic context. Consequently, a large volume of HIV sequence data is publicly available, more than for any other pathogen, and including sequences sampled from diverse geographic and demographic settings. At the time of this writing, there were $635,400$ HIV sequences publicly available in GenBank, annotated with 172 distinct countries of origin. Since HIV is almost always spread through either sexual contact or sharing of injection drug supplies, the contact networks underlying HIV epidemics are driven by social dynamics and are therefore likely to be highly structured [26]. Moreover, since no cure yet exists, efforts to curtail the progression of an epidemic have relied on preventing further transmissions through measures such as treatment as prevention (TasP) and education leading to behaviour change. The effectiveness of this type of intervention can vary significantly based on the underlying structure of the network and the particular nodes to whom the intervention is targeted [29, 30]. Due to this combination of data availability and potential public health impact, HIV is an obvious context in which our method could be applied. Therefore, the third and final aim of this work is *to apply ABC to fit the Barabási-Albert model to existing HIV outbreaks*.

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

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

1.2 Phylogenetics and phylodynamics

1.2.1 Phylogenetic trees

In evolutionary biology, a *phylogeny*, or *phylogenetic tree*, is a graphical representation of the evolutionary relationships among a group of organisms or species (generally, *taxa*) [31]. The *tips* of a phylogeny, that is, the nodes without any descendants, correspond to *extant*, or observed, taxa. The *internal nodes*

correspond to their common ancestors, usually extinct (although occasionally the internal nodes may be observed as well, eg. [32]). The edges or *branches* of the phylogeny connect ancestors to their descendants. Phylogenies may have a *root*, which is a node with no descendants distinguished as the most recent common ancestor of all the extant taxa [33]. When such a root exists, the tree is referred to as being *rooted*; otherwise, it is *unrooted*. The structural arrangement of nodes and edges in the tree is referred to as its *topology* [34].

The branches of the tree may have associated lengths, representing either evolutionary distance or calendar time between ancestors and their descendants. The term "evolutionary distance" is used here imprecisely to mean any sort of quantitative measure of evolution, such as the number of differences between the DNA sequences of an ancestor and its descendant, or the difference in average body mass or height. A phylogeny with branch lengths in calendar time units is often referred to as *time-scaled*. In a time-scaled phylogeny, the internal nodes can be mapped onto a timeline by using the tips of the tree, which usually correspond to the present day, reference points [35]. The corresponding points on the timeline are called *branching times*, and the rate of their accumulation is referred to as the *branching rate*. Rooted trees whose tips are all the same distance from the root are called *ultrametric* trees [36]. These concepts are illustrated in fig. 1.1.

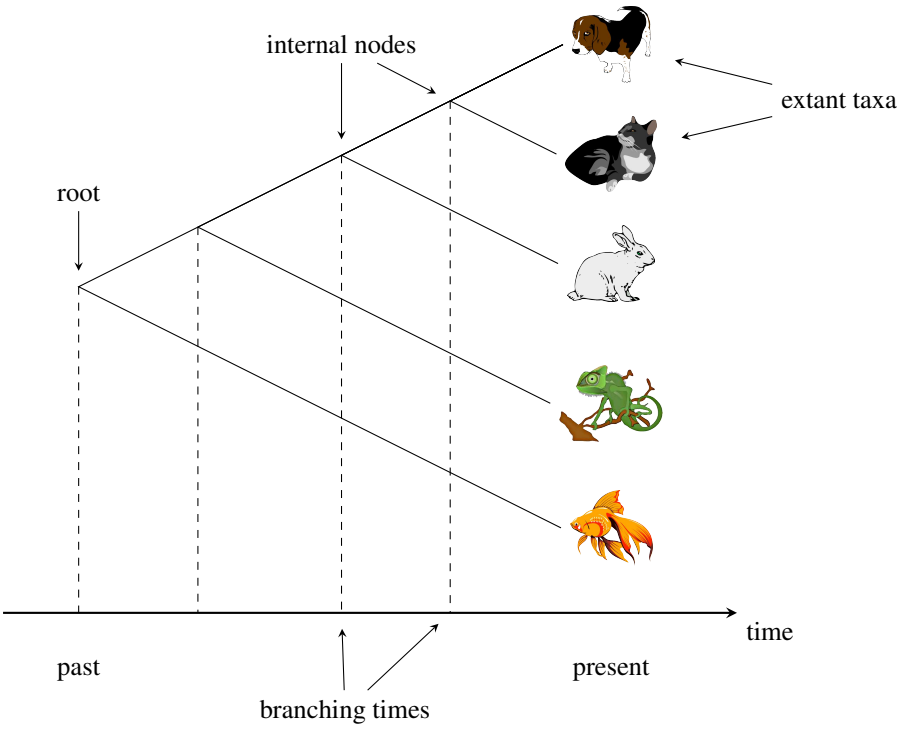

Figure 1.1: Illustration of a rooted, ultrametric, time-scaled phylogeny. The tips of the tree, which represent extant taxa, are placed at the present day on the time axis. Internal nodes, representing extinct common ancestors to the extant taxa, fall in the past. The topology of the tree indicates that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

1.2.2 Transmission trees

In epidemiology, a *transmission tree* is a graphical representation of an epidemic's progress through a population [37]. Like phylogenies, transmission trees have tips, nodes, edges, and branch lengths. However, rather than recording an evolutionary process (speciation), they record an epidemiological process (transmission). The tips of a transmission tree represent the removal by sampling of infected hosts, while internal nodes correspond to transmissions from one host to another. Transmission trees generally have branch lengths in units of calendar time, with branching times indicating times of transmission. The root of a transmission tree corresponds to the initially infected patient who introduced the epidemic into the network, also known as the *index case*. The internal nodes may be labelled with the donor of the transmission pair, if this is known. The tips of the tree, rather than being fixed at the present day, are placed at the time at which the individual was removed from the epidemic, such as by death, recovery, isolation, behaviour change, or migration [38]. Consequently, the transmission tree may not be ultrametric, but may have tips located at varying distances from the root. Such trees are said to have *heterochronous* taxa [39], in contrast to the *isochronous* taxa found in most phylogenies of macro-organisms. A transmission tree is illustrated in fig. 1.2 (right). The object on the right of the figure is called a *contact network*, which depicts the entire susceptible population along with all possible routes of disease transmission. Contact networks, and their relationships to transmission trees, will be discussed further in section 1.3.

Each infected individual in an epidemic may appear at nodes of the transmission tree more than once. This is different from the transmission *network*, in which each infected individual appears exactly once, and edges are in one-to-one correspondence with transmissions [8, 40]. The distinction between the two objects is illustrated in fig. 1.2. However, since transmission networks generally have no cycles (unless re-infection occurs), they are trees in the graph theoretical sense, and hence are sometimes also referred to as transmission trees [*e.g.* 41]. In this work, we reserve the term "transmission tree" for the objects depicted on the right side of fig. 1.2, following *e.g.* [38]. The term "transmission network" is taken to mean the subgraph of the contact network along which transmissions occurred, following *e.g.* [8, 40].

Since transmission trees are essentially a detailed record of an epidemic's progress, they contain substantial epidemiological information. As a basic example, the lineages-through-time (LTT) plot [35], which plots the number of lineages in a phylogeny against time, can be used to quantify the incidence of new infections over the course of an epidemic [42]. However, in all but the most well-studied of epidemics, transmission trees are not possible to assemble through traditional epidemiological methods [40]. The time and effort to conduct detailed interviews and contact tracing of a sufficient number of infected individuals is usually prohibitive, and may additionally be confounded by misreporting and other challenges [43]. However, it turns out that for viral epidemics, some of the epidemiological information contained in the transmission tree leaves a mark on the viral genetic material circulating in the population. A family of methods called *phylodynamics* [44] addresses the challenge of estimating epidemiological parameters from viral sequence data [12].

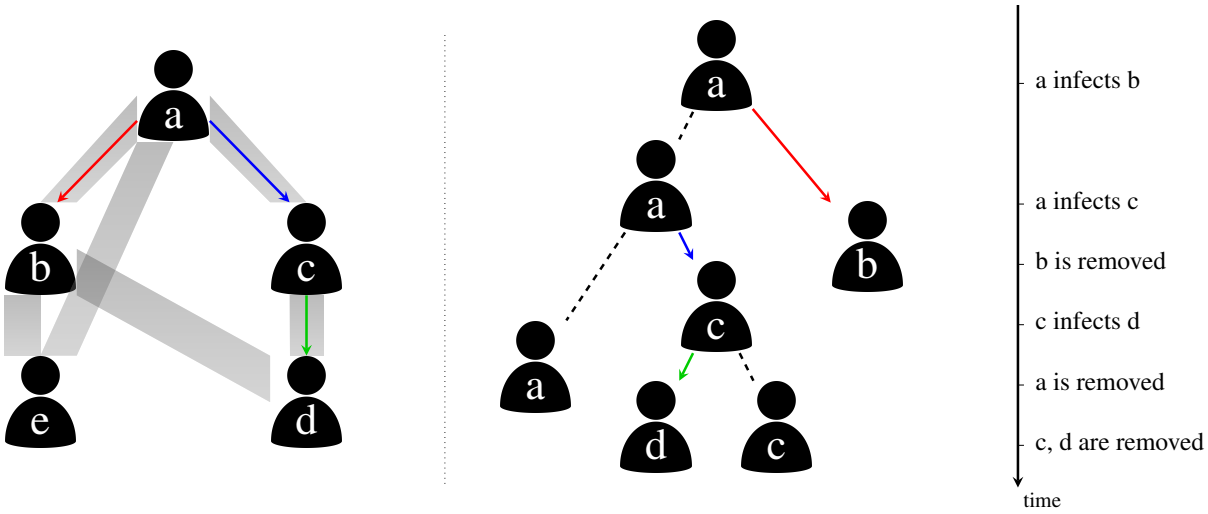

Figure 1.2: Illustration of epidemic spread over a contact network, and the corresponding transmission tree. (Left) A contact network with five hosts, labelled *a* through *e*. Thick shaded edges indicate symmetric contacts among the hosts. The transmission network is indicated by coloured arrows. The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

1.2.3 Phylodynamics: linking evolution and epidemiology

The basis of phylodynamics is the fact that, for RNA viruses, epidemiological and evolutionary processes occur on similar time scales [39]. In fact, these two processes interact, such that it is possible to detect the influence of host epidemiology on the evolutionary history of the virus as recorded in an *inter-host viral phylogeny*. Phylodynamic methods aim to detect and quantify the signatures of epidemiological processes in these phylogenies [11, 12], which relate one representative viral genotype from each host in an infected population. These methods have been used to investigate parameters such as transmission rate, recovery rate, and basic reproductive number [11, 12]. The majority of phylodynamic studies attempt to infer the parameters of an epidemiological model for which the likelihood of an observed phylogeny can be calculated. Most often, this is some variation of the birth-death [45, 46] or coalescent [47, 48] models. These methods either assume the viral phylogeny is known, as we do in this work, or (more commonly) integrate over phylogenetic uncertainty in a Bayesian framework. Phylogenetic inference is a complex topic which we shall not discuss here; see *e.g.* [49] for a full review.

Due to the relationship between the aforementioned processes, there is a degree of correspondence between viral phylogenies and transmission trees [37, 41, 50, 51]. In particular, the transmission process is quite similar to *allopatric speciation* [52], where genetic divergence follows the geographic isolation of a sub-population of organisms. Thus, transmission, which is represented as branching in the transmission tree, causes branching in the viral phylogeny as well [53]. Similarly, the removal of an individual from the transmission tree causes the extinction of their viral lineage in the phylogeny. Consequently, the topology of the viral phylogeny is sometimes used as a proxy for the topology of the transmission tree [54]. Modern likelihood-based methods of phylogenetic reconstruction [*e.g.* 55, 56] produce

unrooted trees whose branch lengths measure genetic distance in units of expected substitutions per site. On the other hand, transmission trees are rooted, and have branches measuring calendar time [11]. Therefore, estimating a transmission tree from a viral phylogeny requires the phylogeny to be rooted and time-scaled. Methods for performing this process include root-to-tip regression [57–59], which we apply in this work, and least-square dating [60]. Alternatively, the tree may be rooted separately with an outgroup [61] before time-scaling.

A caveat of estimating transmission trees in this manner is that the correspondence between the topologies of the viral phylogeny and transmission tree is far from exact [37, 62]. Due to intra-host diversity, the viral strain which is transmitted may have split from another lineage within the donor long before the transmission event occurred. Hence, the branching point in the viral phylogeny may be much earlier than that in the transmission tree. Another possibility is that one host transmitted to two or more recipients in one order, but the transmitted lineages originated within the donor in a different order. In this case, the topology of the transmission tree and the viral phylogeny will be mismatched. In practice, this discordance has not proven an insurmountable problem: for example, Leitner et al. [63] and Paraskevis et al. [64] were able to accurately recover known transmission trees using viral phylogenies. The problem of accurately estimating transmission trees is an ongoing area of research [32, 54, 65–68]. For example, Hall, Woolhouse, and Rambaut [54] developed a Bayesian method to jointly estimate a transmission tree and viral phylogeny by combining models of agent-based transmission, within-host population dynamics, and sequence evolution.

1.2.4 Tree shapes

To perform phylodynamic inference, we must be able to extract quantitative information from viral phylogenies. What is informative about a phylogeny, beyond the demographic characteristics of the individuals it relates, is its *shape*. The shape of a phylogeny has two components: the topology, and the distribution of branch lengths [69]. Methods of quantifying tree shape fall into two categories: summary statistics, and pairwise measures. Summary statistics assign a numeric value to each individual tree, while pairwise measures quantify the similarity between pairs of trees.

One of the most widely used tree summary statistics is Sackin's index [70], which measures the imbalance or asymmetry in a rooted tree. For the ith tip of the tree, we define N_i to be the number of branches between that tip and the root. The unnormalized Sackin's index is defined as the sum of all N_i. It is called unnormalized because it does not account for the number of tips in the tree. Among two trees having the same number of tips, the least-balanced tree will have the highest Sackin's index. However, among two equally balanced trees, the larger tree will have a higher Sackin's index. This makes it challenging to compare balances among trees of different sizes. To correct this, Kirkpatrick and Slatkin [71] derive the expected value of Sackin' index under the Yule model [72]. Dividing by this expected value normalizes Sackin's index, so that it can be used to compare trees of different sizes. An example of a pairwise measure is the normalized lineages-through-time (nLTT) [73], which compares the LTT [35] plots of two trees. Specifically, the two LTT plots are normalized so that they begin at $(0,0)$ and end at $(1,1)$, and the absolute difference between the two plots is integrated between 0 and 1. In the context of

infectious diseases, the LTT is related to the prevalence [42], so large values may indicate that the trees being compared were produced by different epidemic trajectories [73].

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

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

that is, the inner product of the objects' representations in the feature space. Computing $\varphi(x)$ may be computationally prohibitive due to the dimension of $\mathscr{F}$. The utility of a kernel function K is that it is constructed in such a way that it can compute the inner product without explicitly computing $\varphi(x)$. The kernel function developed in [18] will henceforth be referred to as the *tree kernel*. This kernel maps trees into the space of all possible possible *subset trees*, which are subtrees that do not necessarily extend all the way to the tips. The subset-tree kernel was originally developed for comparing parse trees in natural language processing [75] and did not incorporate branch length information. The version developed by Poon et al. [18] includes a radial basis function to compare the differences in branch lengths, thus incorporating both the trees' topologies and their branch lengths in a single similarity score.

The kernel score of a pair of trees, denoted $K(T_1,T_2)$, is defined as a sum over all pairs of nodes (n_1,n_2), where n_1 is a node in T_1 and n_2 is a node in T_2. Following Poon et al. [18], let $N(T)$ denote the set of all nodes in T, $\mathrm{nc}(n)$ be the number of children of node n, c_n^j be the jth child of node n, and l_n be the vector of branch lengths connecting node n to its $\mathrm{nc}(n)$ children. Furthermore, let $\mathrm{nl}(n)$ be the number of children of n which are leaves (we always have $\mathrm{nl}(n) \leq \mathrm{nc}(n)$). The *production rule* of n is the pair $(\mathrm{nc}(n), \mathrm{nl}(n))$. That is, if two nodes have the same number of children and among these, the same number of leaves, then they have the same production rule. Let $k_G(x,y)$ be a Gaussian radial basis function of the vectors x and y,

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

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

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

where

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

The parameter λ in the above expression, is called the *decay factor* [76], and takes a value between 0

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

The tree kernel was later shown to be highly effective in differentiating trees simulated under a compartmental model with two risk groups of varying contact rates [22]. In that paper, Poon used the tree kernel as the distance function in approximate Bayesian computation (ABC) (see section 1.5), to fit epidemiological models to observed trees.

1.3 Contact networks

1.3.1 Overview

Epidemics spread through populations of hosts through *contacts* between those hosts. The definition of contact depends on the mode of transmission of the pathogen in question. For an airborne pathogen like influenza, a contact may be simple physical proximity, while for human immunodeficiency virus (HIV), contact could be via unprotected sexual relations or blood-to-blood contact (such as through needle sharing). A *contact network* is a graphical representation of a host population and the contacts among its members [8, 77, 78]. The *nodes* in the network represent hosts, and *edges* or *links* represent contacts between them. A contact network is shown in fig. 1.2 (left). Contact networks are a particular type of *social network* [79, 80], which is a network in which edges may represent any kind of social or economic relationship. Social networks are frequently used in the social sciences to study phenomena where relationships between people or entities are important [for a review see 81].

Edges in a contact networks may be *directed*, representing one-way transmission risk, or *undirected*, representing symmetric transmission risk. For example, a network for an airborne epidemic would use undirected edges, because the same physical proximity is required for a host to infect or to become infected. However, an infection which may be spread through blood-to-blood contact through transfusions would use directed edges, since the recipient has no chance of transmitting to the donor. Directed edges are also useful when the transmission risk is not equal between the hosts, such as with HIV transmission among men who have sex with men (MSM), where the receptive partner carries a higher risk of infection than the insertive partner [82]. In this case, a contact could be represented by two directed edges, one in each direction between the two hosts, with the edges annotated by what kind of risk they imply [81]. An undirected contact network is equivalent to a directed network where each contact is represented by two

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

Epidemiological models most often assume some form of contact homogeneity. The simplest models, such as the susceptible-infected-recovered (SIR) model [5], assume a completely homogeneously mixed population, where every pair of contacts is equally likely. More sophisticated models partition the population into groups with different contact rates between and among each group [83]. However, these models still assume that every possible contact between a member of group i and a member of group j is equally likely. This assumption is clearly unrealistic for the majority of human communities and can lead to errors in predicted epidemic trajectories when there is substantial heterogeneity present [6, 84, 85]. Contact networks provide a way to relax this assumption by representing individuals and their contacts explicitly. It is important to note that, although panmixia is an unrealistic modelling assumption, it has not proven a substantial hurdle to epidemic modelling in practice [5]. Using this assumption, researchers have been able to derive estimates of the transmission rate and the basic reproductive number of various outbreaks, which have agreed with values obtained by on-the-ground data collection [86]. Therefore, if one is interested only in these population-level variables, the additional complexity of contact network models may not be warranted. Rather, these models are most useful when we are interested in properties of the network itself, such as centrality, structural balance, and transitivity [81].

From a public health perspective, knowledge of contact networks has the potential to be extremely useful. On a population level, network structure can dramatically affect the speed and pattern of epidemic spread [*e.g.* 7, 87]. For example, epidemics are expected to spread more rapidly in networks having the "small world" property, where the average path length between two nodes in the network is relatively low [88]. Some sexually transmitted infections would not be expected to survive in a homogeneously mixed population, but their long-term persistence can be explained by contact heterogeneity [5, 89]. Hence, the contact network can provide an idea of what to expect as an epidemic unfolds. In terms of actionable information, the efficacy of different vaccination strategies may depend on the topology of the network [8–10, 90]. On a local level, contact networks can be informative about the groups or individuals who are at highest risk of acquiring or transmitting infection who would therefore benefit most from public health interventions [29, 30].

Contact networks are a challenging type of data to collect, requiring extensive epidemiological investigation in the form of contact tracing [8, 40, 43, 78]. Therefore, it has been necessary to explore less resource-intensive alternatives which still contain information about population structure. For instance, it is possible to obtain limited information about the contact network by individual interviews without contact tracing. Variables which can be estimated in this fashion are referred to as *node-level* measures [81]. One of the most well-studied of these is the degree distribution mentioned above, which can theoretically be estimated by simply asking each person how many contacts they had in some interval of time. However, the degree distributions often observed in real-world sexual networks are heavy-

tailed [23–25], so dense or respondent-driven sampling [91] would be needed to capture the high-degree nodes characterizing the tail of the distribution.

An alternative approach has been the analysis of other types of network, which can be directly estimated with phylogenetic methods from viral sequence data. Some work focuses on the *phylogenetic network*, in which two nodes are connected if the genetic distance between their viral sequences is below some threshold. Primarily, this work has focused on the detection of *phylogenetic clusters*, which are groups of individuals whose viral sequences are significantly more similar to each other's than to the general population's. The phylogenetic network is informative about "hotspots" of transmission and can be used to identify demographic groups to whom targeted interventions are likely to have the greatest effect [92]. However, this network may show little to no agreement with contact data obtained through epidemiological methods [93–95] and therefore may be a poor proxy for the contact network. Other studies [96] have investigated the *transmission network*, which is the subgraph of the contact network consisting of infected nodes and the edges that led to their infections [40] (fig. 1.2, left). It is possible to estimate the transmission network phylogenetically, although the methods required for doing so are more sophisticated than for estimating the phylogenetic network [96]. These studies again mostly focus on clustering and also on degree distributions.

Other statistical methods have been developed to infer contact network parameters strictly from the timeline of an epidemic, using neither genetic data nor reported contacts. Britton and O'Neill [97] developed a Bayesian method to infer the p parameter of an Erdős-Rényi (ER) network, along with the transmission and removal rate parameters of the susceptible-infected (SI) model, using observed infection and optionally removal times. However, it was designed for only a small number of observations, and was unable to estimate p independently from the transmission rate. Groendyke, Welch, and Hunter [98] significantly updated and extended the methodology of Britton and O'Neill and applied it to a measles outbreak affecting 188 individuals. They were able to obtain a much more informative estimate of p, although this data set included both symptom onset and recovery times for all individuals and was unusual in that the entire contact network was presumed to be infected. Volz [87] developed differential equations describing the dynamics of the SIR model on a wide variety of random networks defined by their degree distributions. Although the topic of estimation was not addressed in the original paper, Volz's method could in principle be used to fit such models to observed epidemic trajectories, similar to what is done with the ordinary SIR model. Volz and Meyers [84] later extended the method to dynamic contact networks and applied it to a sexual network relating 99 individuals investigated during a syphilis outbreak.

1.3.2 Scale-free networks and preferential attachment

A *scale-free* network is one whose degree distribution follows a power law, meaning that the number of nodes in the network with degree k is proportional to $k^{-\gamma}$ for some constant γ [28]. Scale-free networks are characterized by a large number of nodes of low degree, with relatively few "hub" nodes of very high degree. Epidemiological surveys have indicated that human sexual networks tend to be scale-free [23–26]. Interestingly, many other types of network, including computer networks [89], biological

metabolic networks [99], and academic co-author networks [100], also have the scale-free property.

Several properties of scale-free networks are relevant in epidemiology. The high-degree hub nodes are known as *superspreaders* [101], which have been postulated to contribute in varying degree to the spread of diseases such as HIV [38] and severe acute respiratory syndrome (SARS) [102]. Scale-free networks have no epidemic threshold [89], meaning that diseases with arbitrarily low transmissibility ~~can persist~~ have a chance, however small, of persisting at low levels indefinitely. This is in contrast with homogeneously mixed populations, in which transmissibility below the epidemic threshold would result in exponential decay in the number of infected individuals and eventual extinction of the pathogen [5].

One mechanism which has been shown to lead to scale-free networks is *preferential attachment* [28, 103]. The simplest preferential attachment model is known as the Barabási-Albert (BA) model after its inventors [28]. Under this model, networks are formed by starting with a small number m_0 of nodes. New nodes are added one at a time until there are a total of N in the network. Each time a new node is added, $m \geq 1$ edges are added from it to other nodes in the graph. In the original formulation, the partners of the new node are chosen with probability linearly proportional to their degree plus one.

There has been some contention over the idea that contact networks are scale-free. Handcock and Jones [27] fit several stochastic models of partner formation to empirical degree distributions derived from population surveys of sexual behaviour. They found that a negative binomial distribution, rather than a power law, was the best fit to five out of six datasets, although the difference in goodness of fit was extremely small in four out of these five. Bansal, Grenfell, and Meyers [6] found that an exponential distribution, rather than a power law, was the best fit to degree distributions of six social and sexual networks. **dombrowski2013topological** contend that sexual networks are shaped more by homophily ("like attract like") than by preferential attachment, but find that injection drug users (IDU) network do demonstrate a scale-free structure.

In the paper describing the BA model, Barabási and Albert suggest an extension where the probability of choosing a partner of degree d is proportional to $d^\alpha + 1$ for some constant α. When $\alpha \neq 1$, the degree distribution no longer follows a power law [**krapivsky2000connectivity**]. For $\alpha < 1$, the distribution is a stretched exponential, meaning that the number of nodes of degree k is proportional to $\exp(-k^\beta)$ for some constant β. For $\alpha > 1$, the distribution takes on a characteristic called *gelation*, where a one or a few high-degree hub nodes are connected to nearly every other node in the graph. We do not believe these departures from the power law affect the applicability of the model to real world networks. In fact, **de2007preferential** were able to estimate the preferential attachment power from partner count data collected from the same individuals for consecutive time intervals, and found a value less than one in all cases. It is also worth noting that, in addition to the BA model, other investigations of the interaction between contact networks and transmission trees have studied the Erdős-Rényi and Watts-Strogatz models [104], whose degree distributions do not generally follow a power law under any parameter settings.

When $m = 1$, the network takes on the distinctive shape of a tree, that is, it does not contain any cycles. Cycles are present in the network for all other m values. Examples of BA networks with three different values of the preferential attachment power α are shown in fig. 1.3.

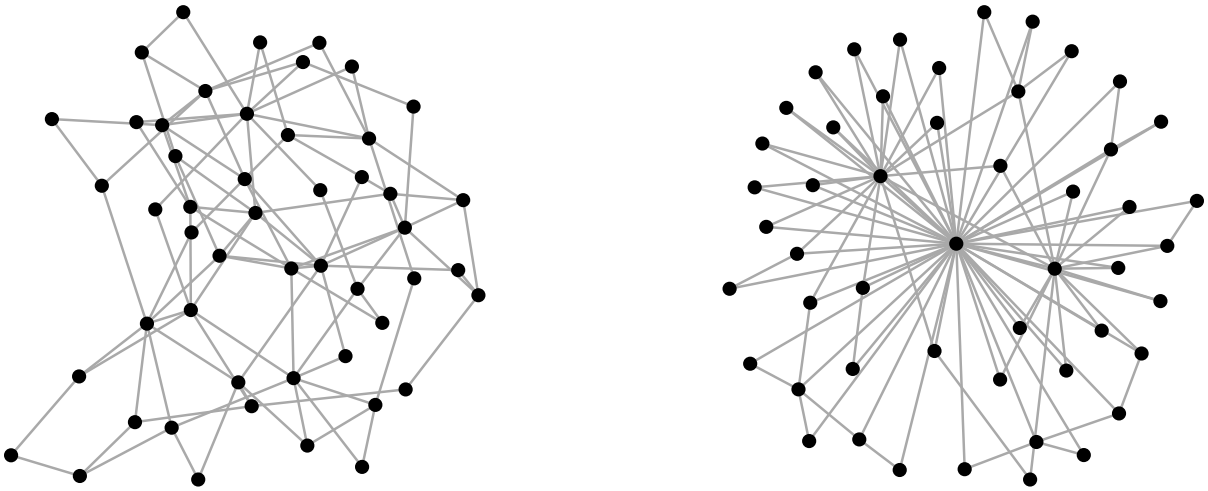

Figure 1.3: Examples of Barabási-Albert networks with preferential attachment power $\alpha = 0$ (left), 1 (centre), and 2 (right). All networks have $N = 50$ nodes and were constructed with $m = 2$ edges per vertex. When $\alpha = 0$, attachments are formed at random and most nodes have low degree.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

1.3.3 Relationship between network structure and transmission trees

The contact network underlying an epidemic constrains the shape of the transmission network, which in turn determines the topology of the transmission tree relating the infected hosts (fig. 1.2). The index case who introduces the epidemic into the network becomes the root of the tree. Each time a transmission occurs, the lineage corresponding to the donor host in the tree splits into two, representing the recipient lineage and the continuation of the donor lineage. Figure 1.2 illustrates this correspondence. It must be emphasized that, although the order and timing of transmissions determines the tree topology uniquely, the converse does not hold. That is, for any given topology, there are in general many transmission networks which would lead to that topology. In other words, it impossible to distinguish who transmitted to whom from a transmission tree alone [105].

A number of studies have made progress in quantifying the relationship between contact networks and transmission trees. O'Dea and Wilke [106] simulated epidemics over networks with four types of degree distribution. They then estimated the Bayesian skyride [107] population size trajectory in two ways: from the phylogeny, using MCMC; and from the incidence and prevalence trajectories, using the method developed by Volz et al. [53]. The concordance between the two skyrides, as well as the relationship between the skyride and prevalence curve, was qualitatively different for each degree distribution. Leventhal et al. [104] investigated the relationship between transmission tree imbalance and several epidemic parameters under four contact network models and found that these relationships varied considerably depending on which model was being considered. The authors also investigated a real-world HIV phylogeny and found a level of imbalance inconsistent with a randomly mixing population. Welch [108] 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. [109] investigated the correspondence between contact network clusters and transmission tree clusters and did find a moderate correspondence between the two in some cases. Goodreau [110] combined a dynamic contact network model with a model of within-host viral evolution to simulate viral phylogenies over eight types of contact network. Estimates of prevalence and effective population size were calculated for each simulated phylogeny under three models of epidemic growth. The author found that estimates for networks with a small high-risk subgroup and networks involving commercial sex workers were substantially different than estimates for random networks or networks with segregated equal-risk groups.

1.4 Sequential Monte Carlo

1.4.1 Overview and notation

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

$$\pi(\theta \mid T) = \frac{f(T \mid \theta)\pi(\theta)}{\int_\Theta f(T \mid \theta)\pi(\theta)}. \tag{1.3}$$

As alluded to in section 1.1, both the likelihood $f(T \mid \theta)$ and the normalizing constant are likely computationally intractable (this will be discussed further in **??**). Hence, rather than computing the posterior distribution analytically, we will approximate it using a *Monte Carlo* approach. The fundamental idea behind Monte Carlo methods is succinctly expressed by Liu, Chen, and Logvinenko [111]:

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

In other words, if we are able to sample enough points from a distribution of interest, we will be able to make reasonably accurate statements about the distribution itself. For example, the expected value of the distribution can be estimated by the sample's population mean. The reason Monte Carlo methods will be useful in this work is that algorithms exist for obtaining samples from distributions that are analytically intractable and from which direct sampling is not possible (for a review see [112]). Sequential Monte Carlo (SMC) [113–115] is one such algorithm.

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

approximated by

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

where n is the number of particles. Initially, the particles do not represent the target distribution but rather a more tractable distribution from which direct sampling is straightforward. The word "sequential" is used to describe the iterative process of perturbation, resampling, and reweighting applied to the particles in such a way that they converge, collectively, to a representation of the target. SMC is also known as the *particle filter*.

In this work, the distribution of interest is the posterior distribution 1.3. The particles are particular values of the parameters θ of the contact network model being studied. If we were taking a typical Bayesian Monte Carlo approach to this problem, the particles would end up weighted by their posterior probability and distributed in such a way that the weighted population was a reasonable representation of $\pi(\theta \mid T)$. In our case, due to the intractable likelihood, we will need to consider an approximation to the posterior (see section 1.5). However, for now, nothing is lost by assuming that our target distribution is the posterior itself.

Here we describe an algorithm called the SMC sampler, developed by Del Moral, Doucet, and Jasra [116], which forms the basis of the adaptive ABC-SMC algorithm we apply toward the main objective of this work. We begin by describing sequential importance sampling (SIS), which is a precursor to SMC that samples from a sequence of distributions defined on spaces of increasing dimension. We then describe SMC itself, which extends SIS with a resampling step to fight particle degeneracy. Finally, we outline the SMC sampler, which allows SMC to be applied to sequences of distributions all defined on the same space. This terminology will become clear as the methods are described.

~~SMC is the name for a family of statistical inference methods that rely on approximating probability distributions of interest with large collections of *particles*, here denoted $\{x^{(k)}\}$ [113–115]. These collections or *populations* are constructed to form a *Monte Carlo approximation* to some distribution of interest π, meaning that the empirical distribution of the particles converges in distribution to π as the population size gets large [111]. The word *sequential* is used because the particle populations are modified in an iterative fashion over time, for example, to incorporate new evidence.~~

To fully describe SMC, we will introduce some notation and terminology. The definitions of these terms will become clearer as they are used. For a sequence $x_1, \ldots, x_d$, we will write $\mathbf{x_i}$ to mean the partial sequence $x_1, \ldots, x_i$. The subscript $^{(k)}$ will be used to indicate the kth particle in a population. To ease the notational burden we will omit the superscripts and subscripts on the weight functions w. We define a *Markov kernel* as the continuous analogue of the transition matrix in a finite-state Markov model. For some spaces X and Y, $K : X \times Y \to [0,1]$ such that

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

for all $x \in X$. This is an "operational" definition of Markov kernel which will be suitable for our purposes. A more rigorous definition can be found in *e.g.* [117]. Note that Markov kernels have nothing to do with the kernel functions defined in section 1.2.4, other than sharing a name (the word "kernel" is

ubiquitous in mathematics). Also, the variable π is used here for a generic target distribution to describe the algorithm, and does not refer to the prior or posterior distributions we will eventually want to work with.

1.4.2 Sequential importance sampling

Sequential importance sampling (SIS) [118] is a particle-based method whose aim is to sample from a distribution π on an high-dimensional space, say $\pi(\mathbf{x}) = \pi(x_1, \ldots, x_d)$. The basis of SIS is importance sampling (IS), which is a method of estimating summary statistics of distributions which are known only up to a normalizing constant, and therefore cannot be sampled from directly. That is, if π is such a distribution and f is any real-valued function, IS is concerned with estimating

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

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

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

Since η can be sampled from exactly, and γ and f can both be evaluated pointwise, the integral $\int w(x)\eta(x)f(x)\mathrm{d}x$ can be approximated by a Monte Carlo estimate. We simply take a sample from η, multiply each point in the sample by f and γ evaluated at that point, and sum the results. Moreover, the normalizing constant Z can be expressed in terms of the importance weight and distribution, $Z = \int w(x)\eta(x)\mathrm{d}x$. Therefore, we have all the ingredients we need to obtain an estimate of $\pi(f)$ using eq. (1.5). Although this is a simple and elegant approach, the drawback is that the variance of the estimate is proportional to the variance of the importance weights [115], which may be quite large if η and γ are very different. Therefore, the practical use of IS on its own is limited, since it depends on finding an importance distribution similar to π, which we usually know very little about *a priori*.

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

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

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

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

The importance weights for η can be written recursively as

$$\begin{aligned} w(\mathbf{x_i}) &= \frac{\pi(\mathbf{x_i})}{\eta(\mathbf{x_i})} && \text{definition of importance weight} \\ &= \frac{\pi(x_i \mid \mathbf{x_{i-1}})\pi(\mathbf{x_{i-1}})}{\eta(x_i \mid \mathbf{x_{i-1}})\eta(\mathbf{x_{i-1}})} && \text{definition of conditional probability} \\ &= \frac{\pi(x_i \mid \mathbf{x_{i-1}})}{\eta(x_i \mid \mathbf{x_{i-1}})} \cdot w(\mathbf{x_{i-1}}) && \text{definition of importance weight.} \end{aligned} \tag{1.6}$$

Thus, we can choose $\eta(x_i \mid \mathbf{x_{i-1}})$ such that the variance of the importance weights is as small as possible at every step, eventually arriving at a full importance distribution. This choice is made on a problem-specific basis, taking any available information about $\pi(x_i \mid \mathbf{x_{i-1}})$ into account (see *e.g.* [115, 119] for many examples). One potential choice for $\eta(x_i \mid \mathbf{x_{i-1}})$ is simply $\pi(x_i \mid \mathbf{x_{i-1}})$, if it is possible to compute. In a Bayesian setting, the prior distribution may be used. The exact form of $\eta(x_i \mid \mathbf{x_{i-1}})$ which minimizes the variance of the weights is called the *optimal kernel* [120], the name deriving from the fact that $k(x_i, \mathbf{x_{i-1}}) = \eta(x_i \mid \mathbf{x_{i-1}})$ is a Markov kernel. In some applications, it is possible to approximate the optimal kernel or even compute it explicitly.

The recursive definition 1.6 suggests an algorithm for obtaining a sample from η and using it to obtain an approximate sample from π by IS (algorithm 1). We begin with n particles, which have been sampled from the importance distribution $\eta(x_1)$ for $\pi(x_1)$. ~~The particles are updated and reweighted d times, corresponding to the d elements of the decomposition of π. At the ith step, each particle is extended to include x_i drawn according to the chosen $\eta(x_i \mid \mathbf{x_{i-1}})$, and the importance weights are recalculated and normalized.~~ Each particle is extended by drawing x_2 from $\eta(x_2 \mid x_1)$, and the importance weights are updated by eq. (1.6). This procedure is repeated d times until the particles have dimension equal to that of π.

Algorithm 1 Sequential importance sampling.

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

1.4.3 Sequential Monte Carlo

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

weights low, the cumulative errors at each sequential step tend to push many of the weights to very low values [113]. This results in a poor approximation to π, since only a few particles retain high importance weights after all d sequential steps, a problem known as *particle degeneracy*. To mitigate this problem, Gordon, Salmond, and Smith [118] introduced technique they called the *bootstrap filter*, which involves a resampling of the population of particles after each sequential step in accordance with their importance weights. A similar idea, termed *particle rejuvination*, was proposed by Liu and Chen [121]. These approaches cause particles with high importance weights to be replicated in the population, while particles with low weights may be removed. After each resampling step, the importance weights for all particles are set equal.

The resampling step was formally integrated with SIS by Doucet, Godsill, and Andrieu [113] to form the first SMC algorithm (algorithm 2). Rather than resample at every step as the bootstrap filter proposed, the authors use a criterion based on the expected sample size (ESS) the particle population to determine when resampling is necessary. The ESS of the population of particles is defined as

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

where n is the number of particles in the population. Resampling is triggered when the ESS drops below the threshold (conventionally $n/2$ [115]) . ~~This results in the removal of low-weight particles from the population, and also equalizes all the weights.~~ Various resampling strategies beyond basic sampling with replacement have been proposed [122], but we will not discuss those here.

Algorithm 2 Sequential Monte Carlo [113].

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

1.4.4 The sequential Monte Carlo sampler

The SIS and SMC algorithms described above aim to sample from a high-dimensional distribution $\pi(\mathbf{x})$, by sequentially sampling from d distributions of lower but increasing dimension. Del Moral, Doucet, and Jasra [116] developed the *SMC sampler* with an alternative objective: to sample sequentially from d distributions $\pi_1, \ldots, \pi_d$, all of the *same* dimension and defined on the same space. The π_i are assumed to form a related sequence, such as posterior distributions attained by sequentially considering new evidence. As with SIS, we assume that $\pi_i(x) = \gamma_i(x)/Z_i$, where each γ_i is known pointwise and the normalizing constants Z_i are unknown.

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

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

where K_i is a Markov kernel and the integral is over the space on which the π_i are defined. The choice of K_i should be based on the perceived relationship between π_{i-1} and π_i. Del Moral, Doucet, and Jasra [116] propose the use of a MCMC kernel with equilibrium distribution π_i. That is,

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

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

Although this method of building up η appears straightforward, the drawback is that the importance distribution itself becomes intractable. In particular, evaluating $\eta_i(x)$ involves a i-dimensional integral of the type in eq. (1.7). As it is necessary to evaluate $\eta(x)$ pointwise to perform IS, this construction appears to have defeated the purpose of providing an importance distribution for each π_i. Del Moral, Doucet, and Jasra [116] overcome this problem with two "artificial" objects. First, they propose the existence of *backward* Markov kernels $L_{i-1}(x_i, x_{i-1})$. For now, these kernels are arbitrary; they will later be precisely defined on a problem-specific basis. Second, the authors define an alternative sequence of target distributions

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

of increasing dimension. That is, $\tilde{\pi}_i$ has dimension equal to the original dimension of π_i raised to the power of i. This brings us back to the SIS setting described above (section 1.4.2), namely of building up an importance distribution sequentially through lower-dimensional distributions. The dimension of the importance distributions η is similarly augmented with the forward kernels,

$$\eta_i(\mathbf{x_i}) = \eta_1(x_1)\prod_{k=1}^{i-1} K_k(x_k, x_{k+1}) \tag{1.8}$$

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

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

and hence

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

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

The importance distributions can also be expressed recursively using the forward kernels, which follows directly from eq. (1.8),

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

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

$$\begin{aligned}
w(\mathbf{x_i}) &= \frac{\tilde{\pi}_i(\mathbf{x_i})}{\eta_i(\mathbf{x_i})} && \text{definition of importance weight} \\
&= \frac{\tilde{\pi}_i(\mathbf{x_i})}{\eta_{i-1}(\mathbf{x_{i-1}})K_i(x_{i-1}, x_i)} && \text{recursive definition of } \eta_i \\
&= \frac{\tilde{\pi}_{i-1}(\mathbf{x_{i-1}})\pi_i(x_i)L_{i-1}(x_i, x_{i-1})}{\eta_{i-1}(\mathbf{x_{i-1}})\pi_{i-1}(x_{i-1})K_i(x_{i-1}, x_i)} && \text{recursive definition of } \pi_i \\
&= w(\mathbf{x_{i-1}}) \cdot \frac{\pi_i(x_i)L_{i-1}(x_i, x_{i-1})}{\pi_{i-1}(x_{i-1})K_i(x_{i-1}, x_i)} && \text{definition of importance weight} \\
&\propto w(\mathbf{x_{i-1}}) \cdot \frac{\gamma_i(x_i)L_{i-1}(x_i, x_{i-1})}{\gamma_{i-1}(x_{i-1})K_i(x_{i-1}, x_i)} && \text{remove normalizing constant } Z_i/Z_{i-1}
\end{aligned} \tag{1.9}$$

The final key piece of information is to notice that, because the L_i are Markov kernels, π_i is simply the marginal in $\mathbf{x_{i-1}}$ of $\tilde{\pi}$. Therefore, a sample from $\tilde{\pi}_i$ automatically gets us a sample from π_i, by considering only the ith component of $\mathbf{x_i}$. In fact, since the weight update eq. (1.9) depends only on the ith and $i-1$st components of each particle, we do not even need to keep track of the complete particles if we are only interested in the final distribution. ~~These are all the ingredients we need to apply SIS.~~ The sequences of kernels L and K should be chosen based on the problem at hand to minimize the variance in the importance weights as well as possible. For a fixed choice of K, the backward kernels which minimize this variance are called the *optimal* backward kernels. The full SMC sampler algorithm is presented as algorithm 3. A resampling step is applied whenever the ESS of the population drops too low, as discussed in the previous section.

Algorithm 3 Sequential Monte Carlo sampler of Del Moral, Doucet, and Jasra [116].

for $k = 1$ to n **do**
  Sample $x_1^{(k)}$ from $\eta_1(x_1)$ ▷ Initialize the kth particle
  $w^{(k)} \leftarrow \gamma_1(x_1^{(k)})/\eta_1(x_1^{(k)})$ ▷ Initialize the importance weights
  Normalize the weights so that $\sum w = 1$
end for
for $i = 2$ to d **do**
  for $k = 1$ to n **do**
    Sample $x_i^{(k)}$ from $K(x_{i-1}^{(k)}, x_i)$ ▷ Extend the kth particle
    $w^{(k)} \leftarrow w^{(k)} \cdot \dfrac{\gamma_i(x_i)L_{i-1}(x_i, x_{i-1})}{\gamma_{i-1}(x_{i-1})K_i(x_{i-1}, x_i)}$ ▷ Update the importance weights
  end for
  Normalize the weights so that $\sum w = 1$
  if $\mathrm{ESS}(w) < T$ **then** ▷ T is a user-defined threshold
    Resample the particles according to w
    for $k = 1$ to n **do**
      $w^{(k)} \leftarrow 1/n$
    end for
  end if
end for

1.5 Approximate Bayesian computation

1.5.1 Overview and motivation

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

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

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

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

Approximate Bayesian computation (ABC) [14–16] was developed to estimate posterior distributions with intractable likelihoods, which have arisen frequently in the domain of population genetics [17, 123]. ABC navigates around the intractable likelihood by replacing the posterior as the target of inference by an *approximate* posterior. This distribution is constructed in such a way that the ratios required for MCMC and the SMC sampler can be computed, conveniently allowing us to apply those algorithms with minimal changes. In the next section, we shall demonstrate how this is done, but first we give the definition of the approximate posterior.

~~Most mathematical models are amenable to fitting via one or both of the approaches, ML or Bayesian inference, discussed above. However, there are some, particularly in the domain of population genetics [17, 123], for which calculation of either the likelihood or the product of the likelihood and the prior may be infeasible. For example, one or both of these quantities may be expressible only as an intractable integral. ABC is designed for such cases, where standard likelihood-based techniques for model fitting cannot be applied.~~

~~Ordinarily, Bayesian inference targets the posterior distribution $\pi(\theta \mid y)$. That is, in the Bayesian framework,~~ By targeting the posterior distribution, Bayesian inference makes the assertion that model parameters with higher posterior density are "better" in the sense that they offer a more credible explanation for the observed data. The approximate posterior targeted by ABC uses an alternative metric for parameter credibility: the similarity of simulated datasets to the observed data. If datasets simulated under the model closely resemble the real data, it follows that the model is a reasonable approximation to the real-world process generating the observed data. More formally, let y be the observed data to which we are trying to fit a model with parameters θ. In the case of this work, the data is a transmission tree T, but we shall stick with the generic variable y for now. Suppose we have a distance measure ρ defined on the space of all possible data our model could generate. ABC aims to sample from the joint posterior distribution of model parameters and simulated datasets z which are within some small distance ε of the observed data y,

$$\pi_\varepsilon(\theta, z \mid y) = \frac{\pi(\theta) f(z \mid \theta) \mathbb{I}_{A_{\varepsilon,y}}(z)}{\int_{A_{\varepsilon,y} \times \Theta} \pi(\theta) f(z \mid \theta) \mathrm{d}\,\theta}. \tag{1.10}$$

Here, $A_{\varepsilon,y}$ is an ε-ball around y with respect to ρ, Θ is the space of all possible model parameters, and $\mathbb{I}$ is the indicator function [124]. The distribution $\pi_\varepsilon(\theta, z \mid y)$ will be referred to as the *ABC target distribution*. The term $f(z \mid \theta)$ appears to be the bothersome likelihood again, but this will turn out not

to be a problem because we are simulating z ourselves.

To return to the context of this thesis, the observed data y is an estimated transmission tree for a viral epidemic under investigation. The model in question is a contact network model with parameters θ. The simulated dataset z is a transmission tree, obtained by first generating a contact network under the model, and then simulating the spread of an epidemic over that network. A transmission tree can be constructed by keeping track of who infected whom during the simulated epidemic (further details will be given in **??**). The distance function ρ must compare two transmission trees - the observed tree, which takes the place of y, and the simulated tree z. We will define this distance function using the tree kernel discussed in section 1.2.4. In words, the approximate posterior we consider here is a distribution which assigns a joint probability density to model parameters and simulated transmission trees under those parameters. The probability density is proportional to the product of the prior on the parameters, and the likelihood of the simulated transmission tree under those parameters, but only if the simulated transmission tree is sufficiently close to the true tree. Otherwise, the probability density is zero.

~~As we shall see in the next section, this distribution can be sampled from exactly. The word "approximate" derives from assumption that, for a suitably chosen distance ρ and a small enough ε, the marginal in z of this distribution approximates the posterior of interest [124]. That is,~~ In fact, it is not the ABC target distribution itself, but rather its marginal in z, which approximates the posterior distribution. In other words, we claim that

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

The intuition for why this approximation might be reasonable comes from considering the case when $\varepsilon = 0$ and ρ has the property that $\rho(z, y)$ if and only if $x = y$. In that case, ~~the fact that, when $\varepsilon = 0$,~~ the ε-ball around y should contain only y itself, hence the integral on the left is exactly equal to the posterior. Thus, by taking ε small, we should attain something close to the posterior if ρ captures the similarity between datasets reasonably well. However, the accuracy of the ABC approximation depends heavily on the choice of distance function [125, 126].

Distance functions and summary statistics in ABC

In many applications (eg. [16, 127]) , ρ is defined as $\rho(S(\cdot), S(\cdot))$ where S is a function which maps data points into a vector of summary statistics. In the context of ABC, a summary statistic S is called *sufficient* if

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

That is, sufficiency implies that the data can be replaced with the summary statistic without losing any information about the posterior distribution [128]. For most problems, it is not possible to find sufficient summary statistics [128]. A number of sophisticated methods have been developed for selecting and weighting summary statistics based on various optimality criteria [125, 126, and references therein]. We do not apply these methods in this work, instead focusing on a distance function which is not based on summary statistics.

~~Summary statistics can be useful if the data are high-dimensional or of a complex type, but~~ Although summary statistics are often presented as a fundamental part of ABC, they are not strictly necessary. For instance, if the data are numeric and of low dimension, the distance function may simply be the Euclidean distance [129]. Park et al. [19] proposed the use of a kernel function (as defined in section 1.2.4) in place of a distance function. The authors referred to their approach as "double-kernel ABC" due to the use of a second (unrelated) kernel function to compute the weights of the particles. The work by Poon [22], upon which ours is based, employed a similar approach, replacing the likelihood ratio in Bayesian MCMC with a ratio of kernel scores.

1.5.2 Algorithms for ABC

Algorithms for performing ABC can be grouped into three categories: rejection, MCMC, and SMC [124]. To simplify the notation, we shall restrict the descriptions of these algorithms to the case of one simulated dataset per parameter particle (the meaning of this will become clear shortly). The extension to multiple datasets per particle is straightforward and will be given at the end of the section. We use the variable x to refer to the pair (θ, z), so that the ABC target distribution can be written $\pi_\varepsilon(x \mid y)$. This makes our notation consistent with section 1.4.

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

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

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

Algorithm 4 Rejection ABC.

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

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

of dimensionality - the measure of the ε-ball around y decreases exponentially with the number of dimensions. ABC-MCMC (algorithm 5) was designed to overcome these hurdles [130]. The approach is similar to ordinary Bayesian MCMC (**??**), except that a distance cutoff replaces the likelihood ratio. That is, the transition probability between states x and x' is defined as

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

Algorithm 5 ABC-MCMC.

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

Some of the same computational inefficiencies arise with ABC-MCMC as with rejection. For example, in regions of low posterior density, the probability to simulate a dataset proximal to the observed data is low. Various strategies have been developed to mitigate this, including reducing the tolerance level ε as the chain progresses [131].

The most recently developed class of algorithm for ABC is ABC-SMC [129, 132]. As with ABC-MCMC, the algorithm is a straightforward modification of an existing Bayesian inference method, in this case the SMC sampler (section 1.4.4). The sequence of target distributions is defined as $\pi_i(x) = \pi_{\varepsilon_i}(x \mid y)$ for a decreasing sequence of tolerances ε_i. The intention is for the algorithm to progress smoothly through a sequence of target distributions which ends at the ABC approximation to the posterior. The initial value ε_1 is set to ∞, which makes the first distribution in the sequence

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

This initial distribution does not depend on the observed data y. In the terminology of the SMC sampler (algorithm 2), the numerator is the first of the γ's, that is, $\gamma_1 = \pi(\theta)f(z \mid \theta)$. Sampling in proportion to γ_1 is straightforward and was already demonstrated for rejection ABC above. Because the sampling is exact, the initial importance weights are all set equal to 1 and normalized to $1/n$ where n is the number of particles.

As discussed in section 1.4.4, the choices of the kernels K and L is problem-specific, and so appropriate kernels must be chosen for ABC. Several options have been proposed [21, 129, 132]. With an appropriately chosen kernel, the weight update (eq. (1.9)) will simplify into a computable expression. Sisson, Fan, and Tanaka [129] and Beaumont et al. [132] both suggest random walk kernels for K, where each particle is perturbed according to a Gaussian distribution. The backwards kernels L are chosen to approximately minimize the variance in importance weights. In this thesis, we use an MCMC kernel, as

proposed by Del Moral, Doucet, and Jasra [21]. The associated backwards kernels and weight updates are discussed in the next chapter (**??**). With generic kernels K and L, the ABC-SMC algorithm is almost identical to the SMC sampler (algorithm 3), with γ_i replaced with π_i. Therefore we will not repeat the full algorithm here.

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

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

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

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

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

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

1.6 Summary

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

The desired output of our method is a posterior distribution of the parameters of a contact network model. Rather than assuming a homogeneously mixed population, as most epidemiological models do, network models take the more realistic view that human populations are structured. That is, contacts which allow for transmission may occur in a nonrandom way, rather than between every pair of individuals in the population. A network model parameterizes this structure. In particular, our second research objective is to characterize our ability to fit the Barabási-Albert (BA) model, which incorporates preferential attachment to generate networks with realistic degree distributions.

To fit these models, our method will apply approximate Bayesian computation (ABC), a simulation-based approach. Simulating a transmission tree according to a network model is straightforward: a network can be generated according to the model, and the spread of an epidemic can be simulated over the network and recorded in a transmission tree. ABC uses the concordance between these simulated

transmission trees and the "true" estimated tree as an indicator of parameter credibility. The closer the simulated transmission trees appear to the true tree, the more weight is assigned to the associated network parameters.

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

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

Chapter 2

Reconstructing contact network parameters from viral phylogenies

2.1 Discussion

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

Contact networks can have a strong influence on epidemic progression, and are potentially useful as a public health tool [29, 30]. Despite this, few methods exist for investigating contact network parameters in a phylodynamic framework [although see 87, 96, 98, 104, 133, for related work]. Kernel-ABC is a model-agnostic method which can be used to investigate any quantity that affects tree shape [22]. In this work, we developed *netabc*, a method based on kernel-ABC to infer the parameters of a contact network model. The method ~~is general, meaning that it~~ can be used to infer parameters of any network model, as long as it allows simulated networks to be easily generated. We have included generators for the BA model discussed here, as well as the Erdős-Rényi (ER) [134] and Watts-Strogatz (WS) [88] network models. Instructions for adding additional models are available in the project's online documentation. We have made *netabc* publicly available at github.com/rmcclosk/netabc under a permissive open source license, to encourage other researchers to apply and extend our method.

Several alternative network models and modelling frameworks have been developed which may provide useful future targets for ABC. Waring models [27, 135] are a more flexible type of preferential attachment model which permit a subset of attachments to be formed non-preferentially. These models were used by Brown et al. [96] to characterize the transmission network in the United Kingdom. Exponential random graph models (ERGMs) [136] are a flexible and expressive parameterization of contact networks in terms of statistics of network features such as pairs and triads. Goodreau [110] evaluated the effect of several different ERGM parameterizations on transmission tree shape and effective population size, and suggested the use of ERGMs as a general framework for estimation of epidemiological quantities related to HIV transmission. Except for a few special cases, simulating a network according to an ERGM generally requires MCMC, which would be too computationally intensive to integrate into *netabc* as it currently stands. To fit ERGM with ABC, one possibility would be to consider the network itself as a parameter to be modified by the MCMC kernel. Other network modelling frameworks include the partnership-centric formulation developed by Eames and Keeling [137] and the log-linear adjacency matrix parameterization applied by Morris [78].

The two-step process of simulating a contact network and subsequently allowing an epidemic to spread over that network carries with it the assumption that the contact network is static over the duration of the epidemic. Clearly this assumption is false, as people make and break partnerships on a regular basis. Addressing the impact of this simplifying assumption is outside the scope of this work. However, the same assumption is made by most studies using contact network models in an epidemiological context [6, 40]. In principle, ABC could be adapted to dynamic contact networks by using a method such as that developed by Robinson, Cohen, and Colijn [138] to simulate such a network, while concurrently simulating the spread of an epidemic.

It is important to note that *netabc* takes a transmission tree as input, rather than a viral phylogeny. In reality, true transmission trees are not available and must be estimated; these estimates are often based on the viral phylogeny. Although this has been demonstrated to be a fair approximation [e.g. 63], and is

frequently used in practice [e.g. 38], the topologies of a viral phylogeny and transmission tree can differ significantly [37, 54] due to within-host evolution and the sampling process. We have left the estimation of a transmission tree up to the user. In theory, it is possible to incorporate the process by which a viral phylogeny is generated along with a transmission tree into our method, for example by simulating within-host dynamics. Indeed, progress has very recently been demonstrated on this front in a talk by Giardina [139], who have independently developed in a method similar to ours to fit contact network models to phylogenetic data that additionally incorporates a within-host evolutionary model. Although this may be an avenue for future extension, we felt that it would obscure the primary purpose of this work, which is to study contact network parameters. In addition, there are a number of different methods available for inferring transmission trees [32, 54, 65, 66, 68], some of which incorporate geographic and/or epidemiological data not accommodated by our method. We therefore felt it would be best to allow researchers to use their own preferred method of constructing a transmission tree.

Our implementation of SMC uses a simple multinomial scheme to sample particles from the population according to their weights. Several other sampling strategies have been developed [122], and it is possible that the use of a more sophisticated technique might increase the algorithm's accuracy. Finally, the ABC-SMC algorithm is computationally intensive, taking about a day when run on 20 cores in parallel with the settings described in the methods section. Implementing parallelization using message passing interface (MPI), rather than Portable Operating System Interface (POSIX) threads as we have done here, would allow the program to be run over a larger number of cores on multiple CPUs in parallel.

2.1.2 Analysis of Barabási-Albert model with synthetic data

The preferential attachment power α had a very strong influence on tree shape in the range of values we considered (**????**). Although the tree kernel was the most effective classifier for α, a Sackin's index of tree imbalance performed nearly as well (**??**). This result was intuitive: high α values produce networks with few well-connected "superspreader" nodes which are involved in a large number of transmissions, resulting in a highly unbalanced ladder-like tree structure (**??**). ~~There was no observable bias in the estimates of α obtained with *netabc*, however the variation in these estimates was higher for $\alpha < 1$ than for $\alpha \geq 1$~~ There appeared to be weaker identifiability for $\alpha < 1$ than for $\alpha \geq 1$ (**????**), which may be partially explained by the relationship between α and the power law exponent γ (**??**). Although the degree distributions do not truly follow a power law for $\alpha \neq 1$, the fitted exponent still captures the slope of the degree distribution reasonably well (**??**), and distributions with similar fitted γ values appear qualitatively similar in other respects. The γ values fitted to $\alpha = 0$ and $\alpha = 0.5$ are nearly identical (about 2.28 for $\alpha = 0$ and 2.33 for $\alpha = 0.5$ with $N = 5000$ and $m = 2$). In other words, the degree distributions of networks with $\alpha < 1$ are similar to each other, which may result in similarity of corresponding transmission trees as well.

The I parameter, representing the prevalence at the time of sampling, was also generally identifiable, although it was slightly over-estimated for both cases we considered with ABC. Sackin's index was better able to capture the effect of I on tree shape than the nLTT (**????**), suggesting that

this parameter impacts the distribution of branching times in the tree more than the topology. In a homogeneously-mixed population, branching times can be modelled by the coalescent process [47], in our case under the SI model [53]. Although our populations are not homogeneously mixed, the forces which affect the distributions of branching times still apply. ~~The dynamics of the SI model, and the coalescent process, offer a potential explanation for the identifiability of I.~~ In our simulations, we assumed that all discordant edges shared the same transmission rate, so that the waiting time until the next transmission in the entire network was always inversely proportional to the number of discordant edges. In the initial phase of the epidemic, when I is small, each new transmission results in many new discordant edges. Hence, there is an early exponential growth phase, producing many short branches near the root of the tree. As the epidemic gets closer to saturating the network, the number of discordant edges decays, causing longer waiting times. ~~The distribution of coalescence times in the tree should therefore be informative about I. This information is captured by the tree kernel, and also by the nLTT statistic, which both performed quite well in classifying I (**??**).~~

The number of nodes in the network, N, exhibited the most variation in terms of its effect on tree shape. There was almost no difference between trees simulated under different N values when the number of infected nodes I was small (**????**). There is an intuitive explanation for this result, namely that adding additional nodes does not change the edge density or overall shape of a BA network. In fact, this is why the degree distributions of these networks are called "scale-free." This can be illustrated by imagining that we add a small number of nodes to a network after the epidemic simulation has already been completed. It is possible that none of these new nodes attains a connection to any infected node. Thus, running the simulation again on the new, larger network could produce the exact same transmission tree as before. On the other hand, when I is large relative to N, the coalescent dynamics discussed above also apply. The waiting times until the next infection increase, resulting in longer coalescence times toward the tips. The relative accuracy of the nLTT in these situations (**????**) corroborates this hypothesis, as the nLTT uses only information about the coalescence times. When all BA parameters were simultaneously estimated with ABC, N was nearly always over-estimated by approximately a factor of two (**????**). One factor which may have contributed to this bias was our choice of prior distribution. Since the prior for I and N was jointly uniform on a region where $I \leq N$, more prior weight was assigned to higher N values. Another contributing factor relates to the dynamics of the SI model and the coalescent process and is discussed below.

I and N were both systematically over-estimated by *netabc*, although the bias was more severe for N than for I. The number of infected individuals follows a logistic growth curve under the SI model. This kind of growth curve has three qualitative phases: a slow ramp-up, an exponential growth phase, and a slow final phase when the susceptible population is almost depleted. The waiting times until the next transmission, which determine the coalescence times in the tree, are dependent on the growth phase of the epidemic. Therefore, we hypothesize that it is the growth phase at the time of sampling which most affects tree shape, rather than the specific values of I or N. To investigate this hypothesis, we simulated transmission trees over networks on a grid of I and N values in the region of uniform prior density. We fit logistic growth curves to the proportion of infected individuals over time, and calculated the first and

second derivatives of these curves (with respect to time) at the time of transmission tree sampling. These derivatives give us an indication of the growth rates of the epidemics at the time of sampling. As shown in fig. 2.1, there are bands along which both derivatives are similar which contain the values we tested. These bands span mostly higher values of N and I than the true values. Therefore, if N and I are free to vary (as is the case in ABC), and our hypothesis is true, both parameters will tend to be overestimated due to being less identifiable within their own band. However, when N is fixed at 5000, the derivatives vary substantially along the I-axis, which explains why the grid search estimates of I were accurate and unbiased (**??**).

Figure 2.1: First and second time derivatives of epidemic growth curves at time of sampling for various values of I and N. Networks were simulated under the BA model with $\alpha = 1.0$, $m = 2$, and N varied along the values shown on the x-axis. Transmission trees were sampled at the time when I nodes were infected (y-axis). Logistic growth curves were fit to epidemic trajectories derived from the transmission trees, and their first and second derivatives were calculated at the time of sampling. Contours show first derivatives, while colours indicate second derivatives. Values of I and N used in simulation experiments with ABC are indicated by diamonds.

The m parameter, which controls the number of connections added to the network per vertex, did not have a measurable impact on tree shape and was not identifiable with ABC, nor with grid search when the other parameters were held fixed. The exception to this was the value $m = 1$, which produces networks without cycles whose associated trees were more easily distinguished (**????**). However, all the analyses presented here did not take the absolute size of the transmission trees into account, as the branch lengths were rescaled by their mean. Because higher m values imply higher edge density, an epidemic should spread more quickly for higher m than lower m with the same per-edge transmission probability. Hence, considering the absolute height of the trees may improve our method's ability to reconstruct m. It was also pointed out to us by an anonymous reviewer that for a fixed I, an infected node may only end up transmitting along a fraction of its outgoing edges, which could mask the presence of the extra edges associated with higher m .