The Transmission Process: A Combinatorial Stochastic Process on Binary Trees over the Contact Network of Hosts in an Epidemic

UCDMS Research Report No. UCDMS2015/4,

School of Mathematics and Statistics, University of Canterbury, Christchurch, NZ

Raazesh Sainudiin · **David Welch**

Abstract We derive a combinatorial stochastic process for the evolution of the transmission tree over the infected nodes of a host contact network in a susceptible-infected (SI) model of an epidemic. This is an explicit description of the transmission process on the product state space of (rooted planar ranked labelled) binary transmission trees and labelled host contact networks with SI-tags as a discrete-state continuous-time Markov chain. We give the exact probability of any transmission tree under various equivalence classes when the host contact network is a complete, star or path network – three illustrative examples. We then develop a biparametric Beta-splitting model that directly generates transmission trees without explicitly modeling the underlying contact network and show that for specific values of the parameters we can get the exact probabilities for our three example networks. We use the maximum likelihood estimator to consistently infer the two parameters driving the transmission process based on observations of the transmission trees. Finally, we show that as the underlying contact networks are interpolated smoothly across the three example networks, the maximum likelihood estimator of the parameters obtained from the corresponding transmission trees are consistent and change smoothly. This suggests that the biparametric Beta-splitting family of transmission trees can be thought of as being generated by contact networks that smoothly span over a large and rich class of networks beyond the three illustrative examples.

Raazesh Sainudiin
Biomathematics Research Centre and School of Mathematics and Statistics
University of Canterbury
Private Bag 4800
Christchurch 8041, New Zealand
E-mail: raazesh.sainudiin@gmail.com

David Welch
Computational Evolution Group and Department of Computer Science
University of Auckland
Private Bag 92019
Auckland 1142, New Zealand
E-mail: david.welch@auckland.ac.nz

Keywords Transmission tree, host contact graph, susceptible-infected epidemic model, non-parametric combinatorial stochastic process, parametric beta-splitting model

Mathematics Subject Classification (2000) 92D30 · 05C05 · 05C20 · 60J10 · 60J27 · 05C85

1 Introduction

The detailed picture of the path an epidemic takes through a population over its course is encapsulated in the *transmission tree*. To understand the process by which a transmission tree grows, we need to consider (i) the *structure of the population* in which the epidemic spreads and (ii) the *state of the individuals* in the population as the epidemic spreads. Network models are a natural candidate for describing population structure where the population is identified with a network in which each vertex represents an individual and an arc (a directed weighted edge) from vertex $\imath_i$ to $\imath_j$, given by a non-negative $w_{i,j} \in [0,\infty)$, represents the propensity with which the infection can be transmitted from $\imath_i$ to $\imath_j$. This propensity can be given meaning in terms of *frequency of contacts* by taking each $w_{i,j} > 0$ to specify independent rate-$w_{i,j}$ Poisson process for the contact times between $\imath_i$ and $\imath_j$, for instance (this is the "meeting process" of Aldous (2013)). We call these networks *contact networks* and assume that they are fixed or static through time. Thus, the contact network of a population summarizes 'who can contact whom and how frequently' and is depicted in Fig. 1(a) for a small population with nodes labeled by individuals $\imath_1, \imath_2, \ldots, \imath_9$ (the edges are undirected).

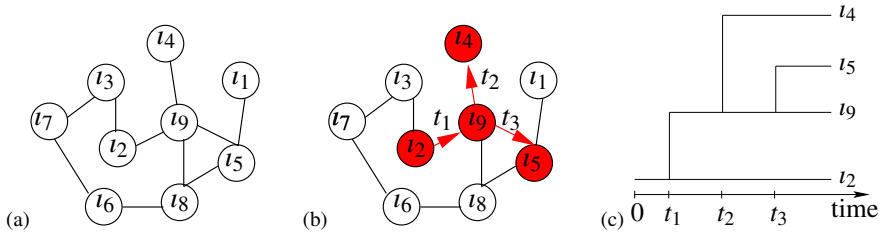

Fig. 1 Spread of an epidemic over (a) the contact network of a population as shown by (b) a sub-network where edges representing transmission events are labelled by the time of event and the infected nodes are colored red and (c) the corresponding transmission tree.

The epidemic state of each individual at a given time can be in one of several possible states, depending on the particularities of the epidemic model. The simplest case, known as the SI model, involves only two states that indicate whether an individual at a given time is susceptible (S) to or infected (I) by a pathogen. Under this model, the only possible state transition is from S to I as specified by the contact network. In other words, a susceptible individual can be infected by any individual in its in-neighbourhood who is already infected. The contact network whose individual

vertices are "tagged" by their epidemic states (S or I) is called the *tagged contact network*. The epidemic states of the individuals in the population after some time are shown by tagging (coloring) the infected or susceptible individuals with I or S tags (red or white colors) in Fig. 1(b).

The *transmission digraph*, a directed edge-labeled subgraph of the contact network containing all infected nodes and directed edges labelled by the time of transmission is a basic object of interest. It is depicted in Fig. 1(b). The transmission digraph can also be represented by the more convenient *transmission tree* shown in Fig. 1(c). The internal vertices of the transmission tree correspond to times of transmission events, the below (or left) and above (or right) planar sub-trees encode who infected whom, and the leaf nodes correspond to the set of infected individuals. Since the tagged contact network co-evolves with the transmission tree, the transmission process is naturally seen as a Markov chain on the product space of tagged contact networks and transmission trees. We consider a stochastic model, as opposed to a deterministic one, to be natural because the spread of an epidemic is inherently probabilistic (Andersson and Britton, 2000).

The transmission tree captures several details about how an infection spreads through the population, including combinatorial structural information such as: who infected whom, order and timing of infection events, the time it takes for a specified set of individuals to be infected, tree shape statistics such as indices of Sackin (1975) and Colless (1982), number of cherries or sub-terminal nodes (McKenzie and Steel, 2000), etc., various isomorphism classes, such as, (un)ranked/(non)planar unlabelled trees and so on, but also classical epidemiological univariate statistics, such as prevalence and incidence through time, reproduction numbers, total time of epidemic and so on.

While various analytical results are available for the univariate epidemiological statistics and can often be obtained without explicitly modelling the tree, most insights about the structural information in the tree are difficult to derive analytically and so are based on simulation studies over parametric families of specific models.

Empirical efforts to understand the transmission process have historically focused on time series and individual event times (such as infection or recovery times) as the main data source. These relatively sparse forms of data have been difficult to collect and not particularly informative, providing limited information about the transmission tree (but see Haydon et al, 2003; Wallinga and Teunis, 2004) or the underlying contact network.

Recently, there has been an increasing attention paid to using the large amounts of viral and bacterial genomic data now available to study outbreaks. The key observation suggesting this data will be informative about the transmission tree is that, if there is little within-host viral genetic diversity, the phylogenetic tree of pathogenic genomes will match the transmission tree (though, in many cases, this assumption does not hold (Romero-Severson et al, 2014; Ypma et al, 2013)). This insight has seen the rise of a new area of research, known as phylodynamics (Grenfell et al, 2004), that specifically treats genomic data in the context of infectious diseases.

The ultimate goal of phylodynamic methods would be to reconstruct the transmission tree (or some sampled subtree) and therefore any interesting properties of the epidemic process. To approach this goal, we need to have good models of how

transmission trees grow which, in turn, requires a thorough understanding of how the structure of the network influences the topology of the transmission tree (Frost et al, 2015).

Previous work on how network structure influences tree topology used computer simulations to vary some property of the network while attempting to hold others constant and observing their influence on simulated transmission trees. For example, Leventhal et al (2012) investigated a number of standard random network models (Erdos-Renyi, Barabasi-Albert and Watts-Strogatz) with a range of parameter values to show that gross changes in the network structure can cause significant and detectable changes in the resulting transmission tree, as measured using the Sackin index of tree imbalance. Frost and Volz (2013) suggest that while this effect is real, it may be swamped by other effects such as sampling strategy. O'Dea and Wilke (2010) concentrate on varying degree heterogeneity in the contact network while holding mean degree constant and also find that heterogeneity is detectable in the transmission tree using standard phylogenetic methods. Welch (2011) employs a simulation approach to study the effect of clustering on transmission trees using exponential family random graph models (ERGMs). Clustering is the most basic of pure network properties, reflecting transitivity (or anti-transitivity) in relationships: if edges (i, j) and (i,k) are present, then high (low) clustering in the network implies that (j,k) is more (less) likely present than when (i, j) and (i,k) are not present. While some changes in various measures of the transmission tree are observed as clustering changes over a wide range of values with degree distribution held constant, a strong effect is not observed suggesting that inference of the clustering property would be difficult. More recent work (Colijn and Gardy, 2014) describes a method that roughly classifies epidemics into host population structures such as homogeneous, super-spreading (Lloyd-Smith et al, 2005) or having a path-like contact network using machine-learning classifiers trained on simulated data.

There is no work that we know of that explicitly estimates a contact network as we have described it here based on transmission trees or genetic data, though some early, ad-hoc attempts exist (Leigh Brown et al, 2011). There is a series of papers (Britton and O'Neill, 2002; Groendyke et al, 2011, 2012) that uses time-series data from epidemics to infer the parameters of an ERGM but the transmission tree here is incidental and poorly inferred. Groendyke et al (2012) suggest that inference within this framework would be greatly improved by having more informative data.

Thus, insights in the literature about the structural or topological information in the tree are primarily based on simulation-intensive programs over parametric families of specific models of the epidemic and the contact network. Formalizing a large class of such simulation programs as a discrete-time Markov chain with transition probabilities in Eq. (2.1) that is embedded in the continuous-time Markov chain with generator in Eq. (2.2) is our first contribution. Such a formalization along with the SageMath/Python code in Sect. A.1 concretizes the meaning of the transmission process, which currently does not seem to be defined unambiguously in the literature.

Models for population structure have increased in complexity over the years; from simple homogeneous models over a static complete network in which each individual has an equal propensity to infect any other individual, to ones which incorporate varying degrees of heterogeneity across the population (who can infect whom) and

through time (time-varying contact networks). Recent reviews by Pastor-Satorras et al (2015) and Holme (2015) summarize this literature.

Basic phylogenetic models such a Kingman's coalescent (Kingman, 1982) that are used for phylodynamic inference assume a fully mixing population of genomes, an assumption that is typically violated in host populations when observed on the epidemic time-scale. Moving to a more complex model such as the structured coalescent (Hudson, 1990; Notohara, 1990) or multi-type branching process (Stadler and Bonhoeffer, 2013) allows incorporation of a few large population features such as country of sampling, but struggles to deal with more than four or five homogeneously mixing population groups at a time (Vaughan et al, 2014; Stadler and Bonhoeffer, 2013; Rasmussen et al, 2014) and is therefore far from the fine scale heterogeneity of a *given static contact network* — our main focus in this paper.

Although static networks are epidemiologically reasonable approximations when the speed of epidemic spread is much faster than the speed of change in the population's structure or vice versa in the case of annealed networks (Pastor-Satorras et al, 2015, III.E), our restriction to static networks in this paper is motivated by finding the simplest and yet interesting mathematical setting to formulate the transmission process. We restrict our attention to the simplest epidemic model on a given static contact network in order to focus on explicitly modeling the *random* transmission tree itself, as the epidemic spreads through the population. To the best of our knowledge, Markov models of transmission trees, over a fixed contact network, and their probabilities are not available explicitly as a function of both branch-lengths and tree topologies even for well-known networks. A straightforward derivation of the probabilities of transmission trees in Sect. 2.1 for some simple static contact networks from the general Markov chains of Eqs. (2.1) and (2.2) is the second contribution of this paper. These examples are meant to illustrate that the general formulae hold for some special cases of contact networks.

We also restrict our attention in this paper to the most basic transmission process which we describe as an SI epidemic model in which hosts are either susceptible (S) to or infected (I) by a pathogen. Our restriction to the simplest model is due to the following reasons. First, this model can be seen as the two-state *Finite Markov Information Exchange* (FMIE) process (Aldous, 2013, Sec. 2.2) called the *Pandemic Process* (Aldous, 2013, Sec. 7) that is shown to be a fundamental building-block (Aldous, 2013, Sec. 3.2,7) for a large class of FMIE processes which includes various classical epidemic models (see Aldous, 2013, Sec. 8,9 and references therein). For instance, the SI model exhibits the fastest possible spread of information in any FMIE model (Aldous, 2013, Sec. 3.2) and it approximates the initial time evolution of the SIS (where infectious hosts return to susceptibility) and SIR (where infectious hosts are removed from the population) models (Pastor-Satorras et al, 2015, II.A). Second, we are mainly interested in allowing the underlying contact network to be essentially 'arbitrary', but fixed. Specifically, we develop a biparametric Beta-splitting family of models for the growth of transmission trees via pure birth events in Sect. 3 that has the following properties:

- gives the exact probability of any transmission tree as a function of α and β,

- avoids having to explicitly model the underlying contact network that is typically unobservable,
- can be interpreted in terms of a Beta-splitting construction for the "infection potential" of the infector and the infectee in a transmission event,
- contains the models generated by the complete, star and path networks when (α,β) equals $(0,0)$, approaches $(\infty,-1)$ and $(-1,\infty)$, respectively,
- smoothly interpolates between the complete, star and path networks,
- has explicit expressions for its maximum likelihood estimators from independent observations of the transmission trees and
- gives the exact probability of various equivalence classes of transmission (unlabelled) trees that are (un)ranked/(non)planar with or without continuous branch-lengths.

This is the most important contribution of this paper and is to be contrasted with what is typically done in the literature since 2000 according to Aldous (2013, Sec. 2.4), whereby various quantitative statements (not of the structural properties of the transmission tree itself but of its univariate summary statistics such as the time for a random individual to be infected) are made on more complex models with increasingly elaborate update rules while considering only a standard number of fixed network "geometries" (or structures) as specific contact networks or as specific random contact networks.

The remainder of the paper is organised as follows. In Sect. 2 we introduce the model for the random growth of a transmission tree over an arbitrary contact network as a discrete-state continuous-time Markov chain and give examples of transmission trees on three specific networks. In Sect. 3 we introduce a parametric Beta-splitting model for the transmission tree, derive the likelihood for a given tree and explore the relationship between this beta-splitting model and the coupled transmission tree-contact network model described in Sect. 2. In Sect. 4 we discuss future directions that this work may take.

2 Model

Consider a population of n individuals with labels in $\mathbb{I}_n = \{\iota_1, \iota_2, \ldots, \iota_n\}$. Let $i(z) : \mathbb{Z}_+ \to [n]$ be a map from the set of non-negative integers $\mathbb{Z}_+ := \{0,1,2,\ldots\}$ to the set of natural numbers no greater than n, $[n] := \{1,2,\ldots,n\}$, so that, $\iota_{i(z)} \in \mathbb{I}_n$ denotes the z-th infected individual as the epidemic evolves in the population. Thus, $\iota_{i(0)}$ is the initially infected individual in the population. In the example of Fig. 2, $\iota_{i(0)} = \iota_2$.

Augment each vertex ι_j in $\mathbb{I}_n$ with a binary status tag:

$$s_j = \begin{cases} 1 & \text{if } \iota_j \text{ is susceptible} \\ 0 & \text{if } \iota_j \text{ is infected} \end{cases}$$

Thus the status of each vertex $\iota_j \in \mathbb{I}_n$ is:

$$s := \{s_j : \iota_j \in \mathbb{I}_n\} \in \{0,1\}^{\mathbb{I}_n}$$

Let k_n be the complete *weighted directed graph* or *network* over the vertex set $\mathbb{I}_n$ with weighted directed edge set $w_n := \{w(\imath_i,\imath_j) \in [0,\infty) : \imath_i \neq \imath_j, (\imath_i,\imath_j) \in \mathbb{I}_n^2\}$. Let 2^{w_n} be the power set of w_n, i.e., the set of all subsets of w_n. For the given set of labelled individuals in the population $\mathbb{I}_n$, let the *susceptible-infected contact network* or SICN be the double

$$c = (w,s) \in \mathscr{C}_n := 2^{w_n} \times \{0,1\}^{\mathbb{I}_n}$$

that is comprised of a weighted directed edge set $w \in 2^{w_n}$ and status tags of the individuals $s \in \{0,1\}^{\mathbb{I}_n}$. Now, for each $z \in \mathbb{Z}_+$, let $c(z) : \mathbb{Z}_+ \to \mathscr{C}_n$ give the SICN at discrete time z standing for the z-th infection event.

We can view the discrete-time discrete-space Markov chain with state space $\mathscr{T}_n \times \mathscr{C}_n$, the product space of $\mathscr{T}_n$, *rooted planar ranked leaf-labelled binary transmission trees*, and $\mathscr{C}_n$, the set of SICNs on $\mathbb{I}_n$. A sample path of this Markov chain for a population of size 3 is shown in Fig. 2. We give the one-step transition probabilities for this Markov chain next.

Let $L(m;\tau(z))$ or $R(m;\tau(z))$ denote the label of the left or right node, respectively, subtending from the internal node labelled by m in $\tau(z)$, the transmission tree at time z. Let $\mathbb{L}(\tau(z))$ denote the set of leaf nodes, i.e., the set of potential infectors, of $\tau(z)$ and let $w(\imath_i,\imath_j;c(z))$ denote the weight of the edge between nodes labelled by $\imath_i$ and $\imath_j$ in $c(z)$, the SICN at time z. Then, the one-step transition probabilities for the discrete-time discrete-space transmission Markov chain is:

$$\begin{aligned}
&\Pr((\tau(z+1),c(z+1))|(\tau(z),c(z))) \\
&= \begin{cases}
\frac{w(L(z+1;\tau(z+1)),R(z+1;\tau(z+1));c(z))}{\sum\limits_{\forall \imath_\ell \in \mathbb{L}(\tau(z))} \sum\limits_{\substack{\forall \imath_j \in \mathbb{I}_n: \\ s_j(z)=1}} w(\imath_\ell,\imath_j;c(z))} & \text{if } (\tau(z),c(z)) \prec (\tau(z+1),c(z+1)) \\
0 & \text{otherwise.}
\end{cases}
\end{aligned} \tag{2.1}$$

By the immediate precedence relation: $(\tau(z),c(z)) \prec (\tau(z+1),c(z+1))$, we mean that $(\tau(z+1),c(z+1))$ can be obtained from $(\tau(z),c(z))$ by a single transmission event. Note that $L(z+1;\tau(z+1))$ and $R(z+1;\tau(z+1))$ are the latest or $(z+1)$-th infector and infectee labels in $\mathbb{I}_n$.

Thus, in words, the transition probability of reaching state $(\tau(z+1),c(z+1))$ from state $(\tau(z),c(z))$ is $w(L(z+1;\tau(z+1)),R(z+1;\tau(z+1));c(z))$, the weight of the edge from the $(z+1)$-th infector to the $(z+1)$-th infectee, that is normalized by the sum of the edge-weights $w(\imath_\ell,\imath_j;c(z))$ from every potential infector, i.e., $\forall \imath_\ell \in \mathbb{L}(\tau(z))$, to every potential infectee within its susceptible out-neighborhood of the SICN at time z, i.e., $\forall \imath_j \in \mathbb{I}_n$ such that $s_j(z) = 1$.

Independent samples of transmission trees from the Markov chain with transition probabilities in Eq. (2.1) over a given SICN `C` and an initial infected individual `initialI` can be generated using `transmissionProcessTC(C,initialI)`, an algorithmic implementation using SageMath/Python (Developers, 2015) in Sect. A.1.

By allowing the time for each infection event to be exponentially distributed with rate $\lambda > 0$, we obtain a continuous-time discrete-space Markov chain from the jump

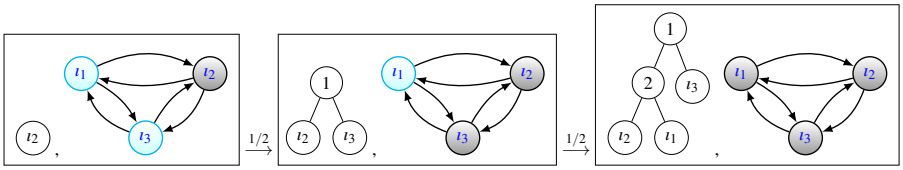

Fig. 2 A sequence of states from the product state space of transmission trees and contact networks in the discrete-time discrete-space jump Markov chain embedded in the transmission process. Initially (left panel) the transmission tree has the root node labelled by the first infected individual $\iota_{i(0)} = \iota_2$ with the corresponding complete contact network k_3 with nodes colored by their susceptible (lightly shaded) or infected status (darkly shaded) over a population of 3 individuals labelled by $\mathbb{I}_3 = \{\iota_1, \iota_2, \iota_3\}$. After the first transmission event from ι_2 to ι_3 with probability $1/2$, the transmission tree splits with the internal node labelling the first infection event by 1 and the first infector ι_2 labelling its left leaf node and the first infectee $\iota_{i(1)} = \iota_3$ labelling its right leaf node (middle panel). In the final absorbing state (right panel), with probability $1/2$, the transmission tree has a new internal node labelled by 2 for the second infection event with its left leaf node labelled by the second infector ι_3 and its right leaf node labelled by the second infectee $\iota_{i(2)} = \iota_1$.

chain in Eq. (2.1) with the following generator:

$$
\begin{aligned}
& q\left((\tau(z), c(z)), (\tau(z+1), c(z+1))\right) \\
& = \begin{cases}
\lambda\, w(L(z+1;\tau(z+1)), & \text{if } (\tau(z), c(z)) \\
\qquad R(z+1;\tau(z+1)); c(z)) & \qquad \prec (\tau(z+1), c(z+1)) \\
-\lambda \sum_{\forall \ell \in \mathbb{L}(\tau(z))} \sum_{\substack{\forall \iota_j \in \mathbb{I}_n: \\ s_j(z)=1}} w(\iota_\ell, \iota_j; c(z)) & \text{if } (\tau(z), c(z)) = (\tau(z+1), c(z+1)) \\
0 & \text{otherwise.}
\end{cases}
\end{aligned}
\tag{2.2}
$$

Note that the parameter λ is usually called β in the epidemiology literature; we use λ to avoid confusion with notation introduced later in the article.

Remark 1 This continuous-time transmission Markov chain and its embedded jump chain is nonparametric since the underlying state space allows for transmission trees to encode an SI epidemic evolving on arbitrary contact networks, i.e., any element of 2^{w_n}. We mainly formulate the model to be concrete about what is typically simulated by computational epidemiologists. We will often, as done in epidemiology, assume that the edges are bi-directional or "undirected". We also focus on connected contact graphs under the assumption that the ideas can be applied to each connected component of a disconnected contact network (but see Sect. 4 for generalization to generic digraphs that may contain a strongly connected giant component).

To gain concrete insights, let us consider the generator of Eq. (2.2) for three specific cases of the contact network.

2.1 Examples

Let us look at Eq. (2.2) for specific initial SICN and initial distributions for the 0-th infected individual. We focus on three of the simplest contact networks to concretely study the effect on the transmission tree distributions they induce.

2.1.1 Transmission on complete network

If the contact network is initially the complete network, i.e., complete weighted directed graph, k_n on $\mathbb{I}_n$ with weights $w(\iota_i, \iota_j) = 1$ for each $\iota_i \neq \iota_j$, then since there are z infected individuals and $n-z$ individuals in each of their susceptible out-neighborhoods after the z-th infection event, the one-step transition probability in Eq. (2.1) simplifies to the following:

$$\Pr((\tau(z+1), c(z+1))|(\tau(z), c(z))) = \begin{cases} \frac{1}{z(n-z)} & \text{if } (\tau(z), c(z)) \prec (\tau(z+1), c(z+1)) \\ 0 & \text{otherwise,} \end{cases} \tag{2.3}$$

and the generator Eq. (2.2) simplifies to the following:

$$\begin{aligned} &q((\tau(z), c(z)), (\tau(z+1), c(z+1))) \\ &\quad = \begin{cases} \lambda & \text{if } (\tau(z), c(z)) \prec (\tau(z+1), c(z+1)) \\ -\lambda\, z(n-z) & \text{if } (\tau(z), c(z)) = (\tau(z+1), c(z+1)), |\mathbb{L}(\tau(z))| = z \\ 0 & \text{otherwise.} \end{cases} \end{aligned} \tag{2.4}$$

If we assume that the first infected individual $\iota_{i(0)}$ is uniformly distributed in $\mathbb{I}_n$, then the probability of a discrete transmission tree $\tau(k)$ with k infection events, where $1 \leq k < n$ is

$$\begin{aligned} \Pr(\tau(k), c(k)) &= \Pr(\tau(0), c(0)) \times \prod_{z=1}^{k} \Pr((\tau(z), c(z))|(\tau(z-1), c(z-1))) \\ &= \frac{1}{n} \times \prod_{z=1}^{k} \left(\frac{1}{z} \times \frac{1}{n-z} \right) = \frac{(n-k-1)!}{n!\, k!} \end{aligned} \tag{2.5}$$

Due to independent exponential waiting times at rate λ, the probability of a transmission tree with branch-lengths $t_{1:k} := (t_1, t_2, \ldots, t_k)$ belonging to $t_{1:k} + dt_{1:k}$, after k infection events, is:

$$\begin{aligned} \Pr(\tau(k), c(k), t_{1:k} + dt_{1:k}) &= \Pr(\tau(k), c(k)) \times \Pr\{t_{1:k} + dt_{1:k}\} \\ &= \Pr(\tau(k), c(k)) \times \prod_{z=1}^{k} z(n-z)\lambda \exp(-\lambda z(n-z)t_z) dt_z \\ &= \frac{1}{n} \prod_{z=1}^{k} (\lambda \exp(-\lambda z(n-z)t_z))\, dt_z \end{aligned} \tag{2.6}$$

Note that when $z = n-1$ and the entire population is infected, then each of the discrete transmission trees (ignoring the branch-lengths) with n leaves labelled by $\mathbb{I}_n$ is equally likely:

$$\Pr(\tau(n-1), c(n-1)) = \frac{1}{n} \times \prod_{j=1}^{n-1} \left(\frac{1}{j} \times \frac{1}{n-j} \right) = \frac{1}{n!(n-1)!}$$

Thus, the number of discrete transmission trees over the complete SI contact network, initialized uniformly at random from any individual in $\mathbb{I}_n$, for different values of $n \in \{1,2,3,4,5,6,7,8,9,10,\ldots\}$ is given respectively by:

$$\{1, 2, 12, 144, 2880, 86400, 3628800, 203212800, 14631321600, 1316818944000, \ldots\} \ .$$

2.1.2 Transmission on star network

If the only initially infected individual is $\iota_{i(0)} = \iota_\star \in \mathbb{I}_n$ and the initial SI contact network is the star network, $\star_n$, centered at $\iota_\star$ with directed edge weights $\{w(\iota_\star, \iota_j) = 1 : \iota_j \in \mathbb{I}_n \setminus \iota_\star\}$, then since there are $n-z$ individuals in the non-empty susceptible out-neighborhood of the only possible infector $\iota_\star$ after the z-th infection event, the one-step transition probability in Eq. (2.1) simplifies to the following:

$$\Pr\left((\tau(z+1), c(z+1)) | (\tau(z), c(z))\right) = \begin{cases} \frac{1}{(n-z)} & \text{if } (\tau(z), c(z)) \prec (\tau(z+1), c(z+1)) \\ 0 & \text{otherwise,} \end{cases} \tag{2.7}$$

and the generator in Eq. (2.2) simplifies to the following:

$$q\left((\tau(z), c(z)), (\tau(z+1), c(z+1))\right) = \begin{cases} \lambda & \text{if } (\tau(z), c(z)) \prec (\tau(z+1), c(z+1)) \\ -\lambda\,(n-z) & \text{if } (\tau(z), c(z)) = (\tau(z+1), c(z+1)), |\mathbb{L}(\tau(z))| = z \\ 0 & \text{otherwise.} \end{cases} \tag{2.8}$$

Let $\mathbb{1}_{\iota_\star}(\iota_{i(0)}) = 1$ if the only initially infected individual is $\iota_\star$ on the star SICN with source node $\iota_\star$, and 0 otherwise. Then the probability of a discrete transmission tree $\tau(k)$ with k infection events, where $1 \leq k < n$ is

$$\begin{aligned} \Pr(\tau(k), c(k)) &= \Pr(\tau(0), c(0)) \times \prod_{z=1}^{k} \Pr\left((\tau(z), c(z)) | (\tau(z-1), c(z-1))\right) \\ &= \mathbb{1}_{\iota_\star}(\iota_{i(0)}) \times \prod_{z=1}^{k} \left(\frac{1}{n-z} \right) = \mathbb{1}_{\iota_\star}(\iota_{i(0)}) \frac{(n-k-1)!}{(n-1)!} \end{aligned} \tag{2.9}$$

Due to independent exponential waiting times at rate λ, the probability of a transmission tree with branch-lengths $t_{1:k} := (t_1, t_2, \ldots, t_k)$ belonging to $t_{1:k} + dt_{1:k}$, after k

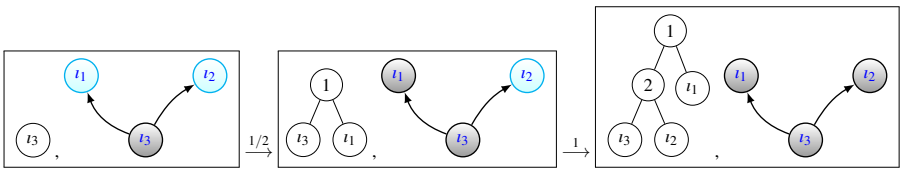

Fig. 3 A sequence of states from the product state space of transmission trees and contact networks in the discrete-time discrete-space jump Markov chain embedded in the transmission process. Initially (left panel) the transmission tree has the root node labelled by the first infected individual $\iota_{i(0)} = \iota_\star = \iota_3$ with the corresponding star network $\star_3$ with nodes colored by their susceptible (lightly shaded) or infected status (darkly shaded) over a population of 3 individuals labelled by $\mathbb{I}_3 = \{\iota_1, \iota_2, \iota_3\}$. After the first transmission event from ι_3 to ι_1 with probability $1/2$, the transmission tree splits with the internal node labelling the first infection event by 1 and the first infector ι_3 labelling its left leaf node and the first infectee $\iota_{i(1)} = \iota_1$ labelling its right leaf node (middle panel). In the final absorbing state (right panel), with probability 1, the transmission tree has a new internal node labelled by 2 for the second infection event with its left leaf node labelled by the second infector ι_3 and its right leaf node labelled by the second infectee $\iota_{i(2)} = \iota_2$.

infection events, is:

$$\begin{aligned}
\Pr(\tau(k), c(k), t_{1:k} + dt_{1:k}) &= \Pr(\tau(k), c(k)) \times \Pr\{t_{1:k} + dt_{1:k}\} \\
&= \Pr(\tau(k), c(k)) \times \prod_{z=1}^{k} (n-z)\lambda \exp(-\lambda(n-z)t_z) dt_z \\
&= \mathbb{1}_{\iota_\star}(\iota_{i(0)}) \prod_{z=1}^{k} (\lambda \exp(-\lambda(n-z)t_z))\, dt_z. \qquad (2.10)
\end{aligned}$$

Note that when $z = n-1$ and the entire population is infected, then each of the discrete transmission trees with the "left-branching comb" topology (ignoring the branch-lengths) with the left-most leaf labelled by the the first infected individual $\iota_{i(0)} = \iota_\star$ and whose remaining $n-1$ leaves is labelled from $\mathbb{I}_n \setminus \iota_\star$ with uniform probability:

$$\Pr(\tau(n-1), c(n-1)) = \mathbb{1}_{\iota_\star}(\iota_{i(0)}) \times \prod_{j=1}^{n-1} \frac{1}{n-j} = \mathbb{1}_{\iota_\star}(\iota_{i(0)}) \frac{1}{(n-1)!}$$

Thus, the number of discrete transmission trees over a star contact network on $\mathbb{I}_n$ with the initially infected individual having degree $n-1$ is $(n-1)!$.

2.1.3 Transmission on path network

If the contact network is the path network on $\mathbb{I}_n$ with directed edge weights equalling 1 along a linear path, and the initial infected individual $\iota_{i(0)}$ is at the beginning of the path, then since there is exactly 1 individual in the non-empty susceptible out-neighborhood of the only possible infector after the z-th infection event, the one-step

transition probability in Eq. (2.1) simplifies to the following:

$$\Pr((\tau(z+1),c(z+1))|(\tau(z),c(z))) = \begin{cases} 1 & \text{if } (\tau(z),c(z)) \prec (\tau(z+1),c(z+1)) \\ \\ 0 & \text{otherwise,} \end{cases} \tag{2.11}$$

and the generator in Eq. (2.2) simplifies to the following:

$$q((\tau(z),c(z)),(\tau(z+1),c(z+1))) = \begin{cases} \lambda & \text{if } (\tau(z),c(z)) \prec (\tau(z+1),c(z+1)) \\ -\lambda & \text{if } (\tau(z),c(z)) = (\tau(z+1),c(z+1)), \\ 0 & \text{otherwise,} \end{cases} \tag{2.12}$$

Let $\mathbb{1}_{\iota\hookrightarrow}(\iota_{i(0)}) = 1$ if the only initially infected individual is $\iota\hookrightarrow$ at the beginning of the path and 0 otherwise. Then the probability of a discrete transmission tree $\tau(k)$ with k infection events, where $1 \leq k < n$ is

$$\Pr(\tau(k),c(k)) = \Pr(\tau(0),c(0)) \times \prod_{z=1}^{k} \Pr((\tau(z),c(z))|(\tau(z-1),c(z-1))) = \mathbb{1}_{\iota\hookrightarrow}(\iota_{i(0)}) \tag{2.13}$$

Due to independent exponential waiting times at rate λ, the probability of a transmission tree with branch-lengths $t_{1:k} := (t_1, t_2, \ldots, t_k)$ belonging to $t_{1:k} + dt_{1:k}$, after k infection events, is:

$$\begin{aligned} \Pr(\tau(k),c(k),t_{1:k}+dt_{1:k}) &= \Pr(\tau(k),c(k)) \times \Pr\{t_{1:k}+dt_{1:k}\} \\ &= \mathbb{1}_{\iota\hookrightarrow}(\iota_{i(0)}) \times \prod_{z=1}^{k} \lambda \exp(-\lambda t_z) dt_z \end{aligned}$$

Thus when $z = n-1$ and the entire population is infected, the discrete transmission tree with the "right-branching comb" topology (ignoring the branch-lengths) with the right-most leaf labelled by the latest infectee is the only possible one.

2.2 Branch-lengths

We can obtain the expected branch-length of the transmission tree between the $(z-1)$-th and z-th infection event or equivalently when there are z infected individuals by simply taking the mean of the exponentially distributed holding-time random variable in the generators given by Eqs. (2.4), (2.8) and (2.12) as shown in Fig. 5. Here we take the 0-th infection event as the initial infection.

Thus, if the underlying SI contact network is k_n then initially at the start of the transmission, the transition rate is $\lambda 1 \times (n-1)$ with expected branch-length $E(T_1) = 1/(\lambda(n-1))$, where T_1 is the duration of the epoch when there is only one

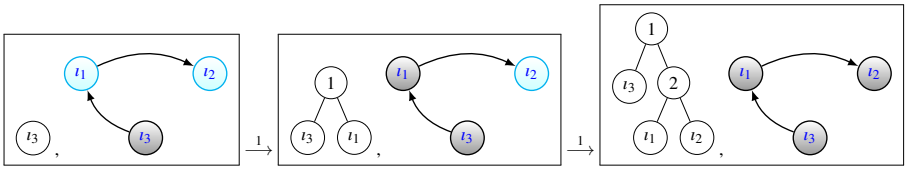

Fig. 4 A sequence of states from the product state space of transmission trees and contact networks in the discrete-time discrete-space jump Markov chain embedded in the transmission process. Initially (left panel) the transmission tree has the root node labelled by the first infected individual $\iota_{i(0)} = \iota_3$ with the corresponding path network with directed edge set $\{(\iota_3,\iota_1),(\iota_1,\iota_2)\}$ and nodes colored by their susceptible (lightly shaded) or infected status (darkly shaded) over a population of 3 individuals labelled by $\mathbb{I}_3 = \{\iota_1,\iota_2,\iota_3\}$. After the first transmission event from ι_3 to ι_1 with probability 1, the transmission tree splits with the internal node labelling the first infection event by 1 and the first infector ι_3 labelling its left leaf node and the first infectee $\iota_{i(1)} = \iota_1$ labelling its right leaf node (middle panel). In the final absorbing state (right panel), with probability 1, the transmission tree has a new internal node labelled by 2 for the second infection event with its left leaf node labelled by the second infector ι_1 and its right leaf node labelled by the second infectee $\iota_{i(2)} = \iota_2$.

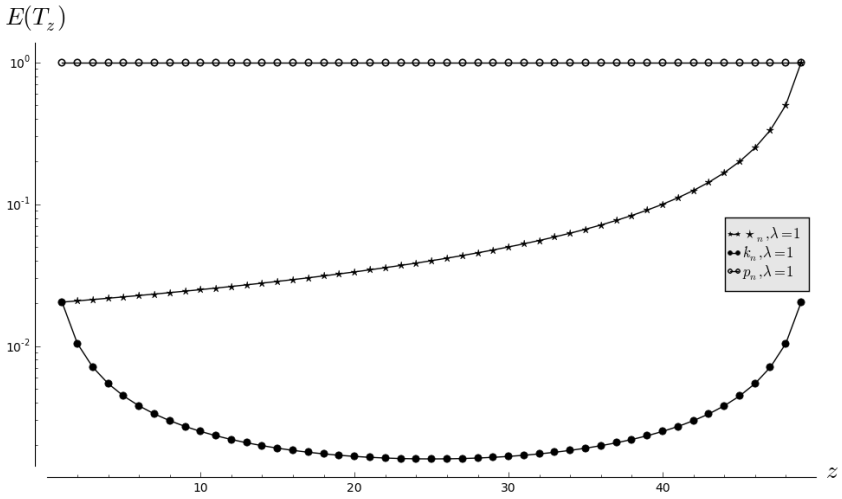

Fig. 5 Expected branch-lengths when there are z infection events or $z+1$ infected individuals, $E(T_z)$, for the three cases. Here $n = 50$ and $\lambda = 1$. $E(T_z) = 1/\lambda = 1$ with the path network p_n of Sect. 2.1.3, $E(T_z) = 1/(\lambda(n-z)) = 1/(50-z)$ with the star network $\star_n$ of Sect. 2.1.2 and $E(T_z) = 1/(\lambda z(n-z)) = 1/(z(50-z))$ with the complete network k_n of Sect. 2.1.1 as z ranges in $\{1,2,\ldots,n-1=49\}$.

infected individual. In general, T_z is the duration of time when there are z infected individuals and is the length of the transmission tree when there are z branches, where $z \in [n-1]$. The transition rate $\lambda z \times (n-z)$ increases and the expected branch-length $E(T_z) = 1/(\lambda z \times (n-z))$ decreases at the z-th infection event as z increases to $n/2$. The expected branch-length is smallest at $4/(\lambda n^2)$ when $z = n/2$ and then starts increasing to $1/(\lambda(n-1))$ as $z \to n-1$ when all n individuals are infected. This is shown as a "bath-tub" curve in Fig. 5. This means that the branch length of the tree at the z-th transmission step, which gives the duration of continuous time taken for

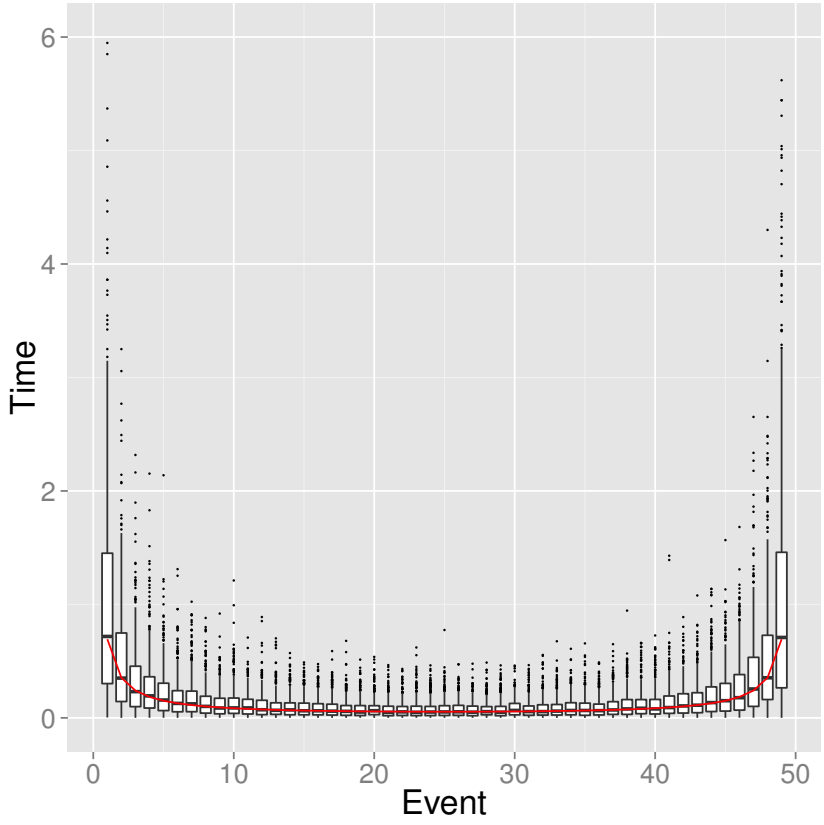

Fig. 6 The sampling distribution of T_z, branch-lengths (times in y-axis) of the transmission tree when there are exactly z infected individuals or between the $(z-1)$-th and z-th infection event (x-axis), where $z \in \{1,2,\ldots,n-1\}$, from 500 independent simulations of the transmission tree over the complete SI contact network for a population of size $n = 50$ (as box plots) and the median branch-lengths given by $E(T_z)\log 2 = (\lambda z(n-z))^{-1}\log 2$, with $\lambda = 1/(n-1)$ (as red solid line).

the z-th infection event, will have mean length $1/(\lambda z \times (n-z))$, such that any one of the k infected leaf nodes can branch with uniform probability $1/z$ at equal rate $\lambda(n-z)$ to infect one of the $(n-z)$ susceptible (and yet uninfected) individuals with uniform probability $1/(n-z)$. The sampling distribution of branch-lengths between consecutive infection events from 500 independent simulations of the transmission tree is shown in Fig. 6 and two typical transmission trees with branch-lengths and topologies over the complete SI contact network for a population of size $n = 50$ is shown in Fig. 7.

Furthermore, by rescaling time in units of population size with $\lambda = 1/(n-1)$, the time of the z-th infection event, T_z, is independent exponential random variable with rate $z(n-z)/(n-1)$ and satisfies the following *randomly-shifted-logistic-limit* (see

for eg. (Aldous, 2013, Eq. 7.13)):

$$T_{\lfloor un \rfloor} - \log n \xrightarrow{d} F^{-1}(u) + G, \quad 0 < u < 1,$$

where, F is the logistic function:

$$F(t) = \frac{\exp(t)}{1+\exp(t)}, \quad -\infty < t < \infty$$

and G has Gumbel distribution with $\Pr(G < x) = \exp(e^{-x})$.

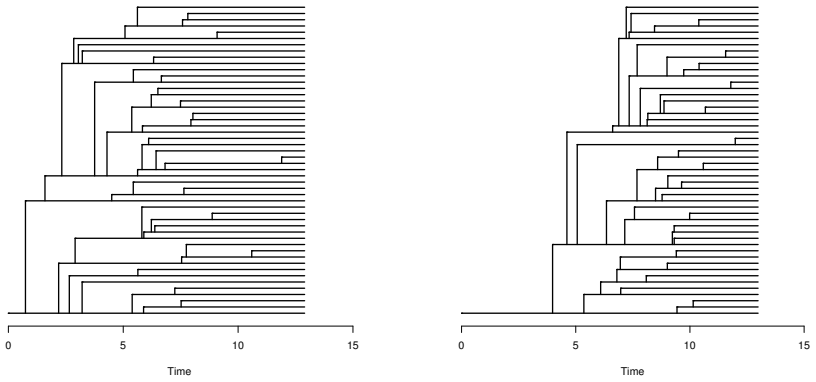

Fig. 7 Two of the 500 independent simulations of the (unlabelled) transmission tree with branch-lengths over the complete SI contact network for a population of size 50 from Fig. 6. Notice the variation in branch-lengths (times between infection events) at the start and end of the epidemic when the variance is largest.

The expected branch-length $E(T_z)$, as a function of $z \in \{1,2,\ldots,n-1\}$, when the SI contact network is the star network ($\star_n$) or the path network (p_n), is inversely proportional to $(n-z)$ or independent of z and n with $E(T_z)$ equalling $1/(\lambda(n-z))$ or $1/\lambda$, respectively, as depicted in Fig. 5.

3 A biparametric Beta-splitting transmission process

We gave a non-parametric description of the transmission process for arbitrary contact networks in the previous section. This Markov construction over the state space of SI contact networks and transmission trees can be too detailed. Often, one does not have knowledge of the state space at this detailed resolution so it is useful to construct transmission processes without explicitly tracking the underlying SI contact network. Here, we give a parametric construction for such a process, by integrating over a Beta-splitting family of transmission trees with interval-labelled leaves, that captures the three Examples in Sects. 2.1.1, 2.1.2 and 2.1.3 as special cases.

The biparametric Beta-splitting model is described in Sainudiin and Veber (2015) for evolutionary trees. We adapt that construction here for transmission trees. To match the standard definition of the Beta distribution, for any $\alpha, \beta > 0$ we call $\mathscr{B}(\alpha, \beta)$ the distribution on $[0,1]$ with density $B(\alpha,\beta)^{-1} x^{\alpha-1}(1-x)^{\beta-1}$, where

$$B(\alpha,\beta) := \int_0^1 x^{\alpha-1}(1-x)^{\beta-1} dx. \tag{3.1}$$

If $\alpha = \beta$, this distribution is symmetric: if $X \sim \mathscr{B}(\beta,\beta)$, then $1-X \sim \mathscr{B}(\beta,\beta)$. We call $\mathscr{B}(\alpha+1,\beta+1)$ as the Beta-splitting density (for $\alpha,\beta > -1$), with density proportional to $x^{\alpha}(1-x)^{\beta}$. This parametric choice corresponds to that used by Aldous (2001) for the symmetric case with $\alpha = \beta$.

We fix $\alpha, \beta > -1$. Let $(B_1, B_2, \ldots)$ be a sequence of independent and identically distributed (i.i.d.) random variables, with the $\mathscr{B}(\alpha+1,\beta+1)$ distribution. Let $(U_1, U_2, \ldots)$ be a sequence of i.i.d. random variables with the uniform distribution on $[0,1]$ that is independent of $(B_1, B_2, \ldots)$. Thus, each of these variables takes its values in $[0,1]$. We call $(G_z = (U_z, B_z))_{z \in \mathbb{N}}$ the *generating sequence* for the Beta-splitting trees. It will be the basis of an incremental construction of transmission tree as a labelled ranked planar binary tree with k leaves and $k-1$ internal nodes.

Our core idea relies on decomposing the transmission tree construction into two stages: (1) constructing a random transmission tree without infector-infectee leaf labels such that it biparametrically captures an essential aspect of the underlying SI contact network's structure and (2) labelling the leaf nodes with infected individuals from $\mathbb{I}_n$ for each transmission or splitting event from stage (1). Stage (2) is optional and the construction of transmission trees without leaf labels from $\mathbb{I}_n$ can be obtained just from stage (1) — such leaf-unlabelled transmission trees can still provide useful prior distributions for integration during inference with partial observations.

Stage (1) of the transmission tree construction involves a deterministic mapping followed by an integration. We first describe the deterministic mapping that takes a realization of the generating sequence $(G_z)_{z \in \mathbb{N}}$ and turns it into a Beta-splitting tree, i.e. a planar binary tree in which the internal nodes are ranked with integer labels and the leaves are labelled by subintervals that partition $[0,1]$. We then describe an integration over (α,β)-specific random partitions by such sub-intervals.

As we shall see below, the integer labels of the internal nodes will give the order in which these nodes have been split during the construction, i.e., the order of infections or successful transmissions. The interval labels of the leaves will form a partition of the interval $[0,1]$ and will be used to decide which leaf is split and becomes an internal node in the next step. The left and right leaf nodes resulting from a split stand for the infector and infectee in the underlying (unobserved) SI contact network after the infection event.

Let $(g_z = (u_z, b_z))_{z \in \mathbb{N}}$ be a realization of the generating sequence. The *organizing map* $O(g)$ proceeds incrementally as follows, until the tree created has k internal nodes and $k+1$ leaves. We start with a single root node, labelled by the interval $[0,1]$.

- Step 1: Split the root into a left leaf with interval label $[0, b_1]$ and a right leaf labelled by $[b_1, 1]$. Change the label of the root to the integer 1.

- Step 2: If $u_2 \in [0, b_1]$, split the left child node of the root into a left leaf and a right leaf labelled by $[0, b_1 b_2]$ and $[b_1 b_2, b_1]$, respectively. If $u_2 \in [b_1, 1]$, then instead split the right child node of the root into left and right leaves with respective labels $[b_1, b_1 + (1-b_1)b_2]$ and $[b_1 + (1-b_1)b_2, 1]$. Label the former leaf that is split during this step by 2.
- Step z: Find the leaf whose label $[a,b]$ contains u_z. Change its label to the integer z and split it into a left leaf with label $[a, a+(b-a)b_z]$ and a right leaf with label $[a+(b-a)b_z, b]$.
- Stop at the end of Step k.

In words, at each step z, the interval labels of the leaves form a partition of the interval $[0,1]$. We find the next leaf node to be split by checking which leaf interval contains the corresponding u_z and then b_z is used to split the interval of that former leaf, say with interval width d, into two intervals of lengths $b_z d$ and $(1-b_z)d$. Thus, the width of the left interval of a current leaf node that is about to be split should be constructed by the Beta-splitting density such that it is proportional to all infection events that will subtend from the current infector and its future infectees after this infection event. Similarly, the width of the right leaf label of this current leaf node should be such that it is proportional to all infection events that will subtend from the current infectee and its future infectees. Intuitively, one can think of the width of the interval label of a leaf node as the *infection potential* of the individual associated with that leaf and the widths of the left and right interval labels upon a split or an infection event as the infection potentials of the infector and the infectee, respectively, after the event. Thus, the beta-splitting trees capture the essence of transmission trees that are co-evolving with underlying SI contact networks, without explicitly requiring complete knowledge of the networks during their construction. The internal node just created is then labelled by z to record the order of the splits. At the end of step z, the tree has $z+1$ leaves, and so we stop the procedure at step k to ensure $k+1$ leaves, where $1 \leq k \leq n-1$. Figure 8 shows an example of such a Beta-splitting tree construction for $k = 3$.

After the Beta-splitting construction, we first integrate over $(G_z)_{z\in\mathbb{N}}$ to 'erase' the interval-valued leaf labels and then assign infected individuals in $\mathbb{I}_n$ as leaf labels to obtain transmission trees from integrated Beta-splitting trees. These trees have k integer-labelled internal nodes or splits and $k+1$ unlabelled leaves. The process of assigning leaf labels from $\mathbb{I}_n$ via a pre-order traversal on the k internal nodes, in increasing order, i.e., Stage (2) of the construction, is described next.

We start with the internal node labelled 1 and assign the initial infected individual $\iota_{i(0)}$ to its left node. Then we assign the first infectee to the right node of 1. In general, as we descend down the internal nodes of the integrated beta-splitting tree in increasing order of its integer labels we slide the individual label ι_ℓ to the left of the split and assign a new label to the right node as the infectee ι_j chosen according to the infectee distribution for ι_ℓ:

$$\iota_j \sim \{\Pr\{\iota_\ell \overset{z}{\rightsquigarrow} \iota_j\} : \iota_j \in \mathbb{I}_n\} \ , \tag{3.2}$$

the probability that ι_ℓ infects ι_j at discrete time-step z. This distribution is defined to be generic on purpose, without necessarily making explicit reference to $c(z)$, the

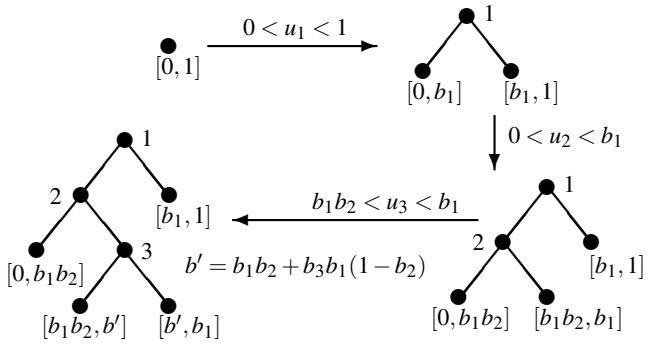

Fig. 8 An example of a Beta-splitting tree construction for $k = 3$.

underlying SI contact network at time z, that is typically unknown or partially known. We can always obtain specific form for Eq. (3.2) by making explicit assumptions on $c(z)$ via the infector-specific infectee distribution within Eq. (2.1):

$$\{\Pr\{\iota_\ell \overset{z+1}{\rightsquigarrow} \iota_j\} : \iota_j \in \mathbb{I}_n\} = \begin{cases} \frac{w(\iota_\ell, \iota_j; c(z))}{\sum_{\substack{\forall \iota_j \in \mathbb{I}_n: \\ s_j(z)=1}} w(\iota_\ell, \iota_j; c(z))} & \text{if } (\tau(z), c(z)) \prec (\tau(z+1), c(z+1)) \\ 0 & \text{otherwise.} \end{cases} \tag{3.3}$$

3.1 Probability of a given Beta-splitting transmission tree

For a given (unlabelled) ranked planar tree, and an internal node labelled by i, let us write n_i^L (resp., n_i^R) for the number of internal nodes in the left (resp., right) subtree below node i. In particular, if node i subtends two leaves, then $n_i^L = 0 = n_i^R$.

Theorem 1 *The probability of any discrete transmission tree $\tau(k)$ with k splits and $k+1$ leaves under the integrated Beta-splitting model is:*

$$\begin{aligned} \Pr\{\tau(k)\} &= \prod_{z=1}^{k} \left\{ \frac{1}{B(\alpha+1, \beta+1)} \int_0^1 b_z^{n_z^L + \alpha} (1 - b_z)^{n_z^R + \beta} db_z \right\} \times \Pr(\textit{leaf labels}) \\ &= \prod_{z=1}^{k} \frac{B(n_z^L + \alpha + 1, n_z^R + \beta + 1)}{B(\alpha+1, \beta+1)} \\ &\qquad \times \Pr\{\iota_{i(0)}\} \prod_{z=1}^{k} \Pr\{(L(z); \tau(z)) \overset{z-1}{\rightsquigarrow} (R(z); \tau(z))\}, \end{aligned} \tag{3.4}$$

where $B(\alpha, \beta)$ was defined in Eq. (3.1).

Proof outline. The second term in the product is due to the independent assignment of infected individual according to Eq. (3.2) as we recursively descend through the infection events encoded by the ranked internal nodes of the tree after the initial infection with $\Pr\{\iota_{i(0)}\}$. We now focus on the first term which results from integrating over the $(U_z, B_z)_{z\in[k]}$, for $1 \le k \le n-1$. Remember that if a leaf is labelled by an interval $[a,b]$, the probability that it is split during the z-th step is $b-a$, the probability that the uniform random variable U_z falls within $[a,b] \subset [0,1]$. If it is chosen to split, it is given label z and the left and right leaves created are labelled by intervals of respective lengths $B_z(b-a)$ and $(1-B_z)(b-a)$. Then these intervals may split later, but into intervals of lengths that are always proportional to B_z or $1-B_z$ (respectively). Now the probability of the tree τ is the product of the k probabilities of choosing a given leaf to split at each step, each of which is equal to the length of the interval labeling that leaf. As a consequence, each split occurring in the left subtree below node z brings in another B_z in the product, or another $1-B_z$ if the split occurs in the right subtree below node z. Averaging over the possible values of the B_z's, which are independent $\mathscr{B}(\alpha+1,\beta+1)$ random variables, yields the result. □

3.2 Examples

Now we reconsider the three specific SI contact networks and show that they arise for specific values of α and β.

Recall that $B(\alpha,\beta)$ is related to the Gamma function Γ by the equality

$$B(\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}, \qquad \alpha,\beta > 0, \tag{3.5}$$

and that $\Gamma(\beta) = (\beta-1)! = (\beta-1)(\beta-2)\cdots 2\cdot 1$ if $\beta \in \mathbb{N}$.

3.2.1 Complete network underlies Beta-splitting transmission trees with $\alpha = \beta = 0$

Let us assume that the initial infection is uniformly distributed in $\mathbb{I}_n$ and that the SICN is the complete contact network k_n with unit weights as in Sect. 2.1.1 and show that the probability of the discrete transmission tree after k infections has the same probability as Eq. (2.5).

The first term in Eq. (3.4) with $\alpha = \beta = 0$, simplifies as follows:

$$\prod_{z=1}^{k} \frac{B(n_z^L+1, n_z^R+1)}{B(1,1)} = \prod_{z=1}^{k} \frac{n_z^L! n_z^R!}{(n_z^L+n_z^R+1)!} = \frac{1}{k!}, \tag{3.6}$$

where the second equality is obtained by observing that $n_z^L + n_z^R + 1$ is the number of internal nodes of the tree rooted at node z, which is the left or the right subtree below the internal node z. Hence, each term $n_z^L!$ in the numerator of the product cancels with the term in the denominator that corresponds to the left child node of z, except if $n_z^L = 0$ and the left child node of z is a leaf. But in this case, $0! = 1$ by convention. The same holds true for each of the $n_z^R!$. Likewise, the terms in the denominator which are not compensated by some term in the numerator are those corresponding to internal

nodes having no ancestral nodes. But the only such node is the root ($z = 1$) with $n_1^L + n_1^R + 1 = k$. This gives us the result.

From Eq. (3.3), the infectee probability is uniformly distributed over $n - z$ infectees for each infector at time-step z and thus the second term in Eq. (3.4) simplifies to:

$$\Pr\{\iota_{i(0)}\} \prod_{z=1}^{k} \Pr\{(L(z);\tau(z)) \overset{z-1}{\rightsquigarrow} (R(z);\tau(z))\} = \frac{1}{n} \prod_{z=1}^{k} \frac{1}{n-z} = \frac{(n-k-1)!}{n!} \tag{3.7}$$

Finally, putting Equations (3.6) and (3.7) into Eq. (3.4), we get the desired identity with Eq. (2.5). Since the probabilities of the discrete transmission trees are identical between the integrated Beta-splitting trees with $\alpha = \beta = 0$ and the construction of Sect. 2.1.1 with an explicit complete SI contact network, the continuous-time process will also be identical to Eq. (2.4) due to independent Exponential rates for the infection events.

Remark 2 The transmission tree thus constructed with $\alpha = \beta = 0$ corresponds to Yule (1924) model for evolutionary trees (ignoring planarity and leaf labels). This Beta-splitting construction is very different from the standard evolutionary construction of the Yule tree, in which the next leaf to split is chosen uniformly at random from among the current set of leaves. Here the choice of the next split is dictated by the lengths of the intervals labeling the current leaves, which will all be distinct will probability one. However, by averaging over the law of the generating sequence (when $\alpha = \beta = 0$) yields the same uniform distribution on rooted ranked planar binary trees with k splits and $k+1$ unlabelled leaves. These $k!$ many trees are in bijective correspondence with permutations of $\{1,\ldots,k\}$ through the *increasing binary tree-lifting* operation (Flajolet and Sedgewick, 2009, Ex 17, p. 132).

3.2.2 Star network underlies Beta-splitting transmission trees with $\alpha \to \infty, \beta \to -1$

To obtain a left-branching comb we let (α,β) approach the limiting bottom-right corner $(\infty,-1)$ of the parameter space. As $\alpha \to \infty$ from the left and $\beta \to -1$ from above, the $\mathscr{B}(\alpha+1,\beta+1)$ distribution concentrates on the boundary of the support at 1. In the limit, each random variable B_z in the generating sequence takes the value 1, with probability 1. Thus, the root is first split into a left leaf with label $[0,1]$ and a right leaf with label $\{1\}$ (i.e., an interval reduced to a single point 1). Next, the uniform random variable U_2 belongs to the interval $[0,1]$ with probability one, so that the left leaf labelled by $[0,1]$ is necessarily that chosen to split next. Again, it is split into two leaves with left leaf label $[0,1]$ and right leaf label $\{1\}$, implying that the next leaf to split is again the left one which inherited the full interval $[0,1]$ with probability one. This recursive reasoning can be carried on until step k with $k+1$ leaves. Hence, morally the tree corresponding to $\alpha \to \infty$ and $\beta \to -1$ is a fully unbalanced tree, with a left-branching comb with $k+1$ leaves. See Fig. 9 for an example with $k = 3$. Recall that this is exactly the transmission tree obtained when the underlying SICN is the star network of Sect. 2.1.2.

For Stage (2) of the construction where we assign leaf labels to the integrated Beta-splitting tree we assume that the underlying SICN is the star network initialized

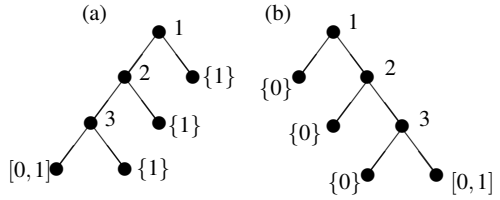

Fig. 9 (a) The discrete transmission tree corresponding to the limiting case $\alpha \to \infty$ and $\beta \to -1$ is a left-branching comb, and (b) the discrete transmission tree corresponding to the limiting case $\beta \to \infty$ and $\alpha \to -1$ is a right-branching comb.

at the source node. Since there is only one discrete transmission tree topology, i.e., the left-branching comb, we can label the leaves of the integrated Beta-splitting tree in $\prod_{z=1}^{k} \left(\frac{1}{n-z}\right)$ many ways to obtain the same probability in Eq. (2.9) for the discrete transmission tree with individuals leaf labels from $\mathbb{I}_n$.

3.2.3 Path network underlies Beta-splitting transmission trees with $\alpha \to -1, \beta \to \infty$

By an analogous argument to that in Sect. 3.2.2 with $\beta \to \infty$ and $\alpha \to -1$, the $\mathscr{B}(\alpha+1, \beta+1)$ distribution concentrates on the boundary of the support at 0 and each random variable B_z in the generating sequence takes the value 0, with probability 1. Thus, the only discrete transmission tree topology for the Beta-splitting tree with $(\alpha, \beta) \to (-1, \infty)$, the limiting top-left corner of the parameter space, is the right-branching comb shown in Fig. 9 (b), the same one obtained by assuming that the underlying SICN is the path network in Sect. 2.1.3. By further assuming that the underlying SICN is the path network for the leaf-labelling Stage (2) with the initial infection spreading from the individual $\iota_{\hookrightarrow}$ at the beginning of the path as in Sect. 2.1.3, we obtain exactly one possible labelling and obtain the same probability in Eq. (2.13).

3.3 A family of contact networks interpolating the star, complete and path networks

In the previous three sections we saw that the distribution on discrete transmission trees generated by the the beta-splitting model with (α, β) taking (limiting) values $(\infty, -1)$, $(0,0)$, and $(-1, \infty)$ corresponds to that under $\star_n$ (the star SICN), k_n (the complete SICN) and p_n (the path SICN), respectively. Since these three specific SICNs seem to be isolated instances of all possible SICNs, we next show that other SICNs that sequentially interpolate between $\star_n$, k_n and p_n can be constructed such that their transmission tree distributions correspond to that under the Beta-splitting model with (α, β) values that also sequentially interpolate between $(\infty, -1)$, $(0,0)$ and $(-1, \infty)$.

In order to find the (α, β) values that correspond to an arbitrary SICN, we use the following inferential procedure. First we generate a sample of r independent transmission trees $(\tau_1, \tau_2, \ldots, \tau_r)$ from the given SICN `C` and initial infected individual `initialI` by calling `transmissionProcessTC(C,initialI)` in Sect. A.1 r times. Then we compute $(\widehat{\alpha}, \widehat{\beta})$, the maximum likelihood estimate (MLE) of the

parameters by maximizing the likelihood function as follows:

$$(\widehat{\alpha},\widehat{\beta}) = \underset{(\alpha,\beta)\in(-1,\infty)\times(-1,\infty)}{\arg\max} \prod_{i=1}^{r} \Pr(\tau_i;\alpha,\beta) \ .$$

The probability of the tree τ_i for a given (α,β), $\Pr(\tau_i;\alpha,\beta)$, is obtained from a post-order traversal of τ_i to compute the first term in Eq. (3.4). To focus on the jump chain's discrete structural information in the transmission trees, our likelihood of the transmission tree ignores leaf labels and the waiting times between events as implemented in Sect. A.2 (such additional information can be included in more elaborate likelihood functions). The demonstration at the end of Sect. A.2 shows two independent MLE computations based on $r = 10$ independent transmission trees (without leaf labels) that were sampled from the complete SICN on $n = 50$ nodes. As expected the MLE $(\widehat{\alpha},\widehat{\beta})$ takes the following values: $(0.0121, -0.0841)$ and $(-0.0406, -0.0379)$. As expected, these are close to $(\alpha,\beta) = (0,0)$, the parameters of the Beta-splitting model corresponding to the transmission tree distribution generated from the complete SICN. The variability in MLE is expected due to natural sampling variability. Now that we have an inferential procedure to consistently estimate the (α,β) parameters of the best-fitting (most likely) beta-splitting transmission process from a set of transmission trees generated from the transmission process on any given SICN, we are ready to present a family of SICNs that interpolate our three extreme SICNs.

A circulant network or digraph on n vertices labelled by $\mathbb{V} = \{0,1,\ldots,n-1\}$ is specified by a set $\mathbb{A} \subset \mathbb{V}$, such that there is an directed edge from vertex i to vertex j if and only if $(j-i) \bmod n$ is an element of $\mathbb{A}$. We denote a circulant digraph on n vertices with edge-specifying set $\mathbb{A}$ by $\mathscr{C}(n,\mathbb{A})$. First note that $\mathscr{C}(n,\{1,2,\ldots,n-1\})$ is the complete network k_n and $\mathscr{C}(n,\{d\})$ has constant degree sequence with degree d since each node i is connected to d neighbours in $\{j : (j-i) \bmod n \in \{k\}\}$.

The transmission process on the linear path network p_n is identical to that on the circular path network $\mathscr{C}(n,\{1\})$ when the infection starts at vertex 0 (at the head of p_n) since the extra directed edge $(n-1,0)$ in $\mathscr{C}(n,\{1\})$ plays no role in the SI model due to vertex 0 already being infected. Thus, we do not distinguish between the circular path and linear path in the sequel. By letting $\mathbb{A}_i = \{1,2,\ldots,i\}$ we get the sequence of circulant graphs to interpolate from the path network to the complete network:

$$(\mathscr{C}(n,\mathbb{A}_i))_{i=1}^{n-1} = (\mathscr{C}(n,\{1\}),\mathscr{C}(n,\{1,2\}),\mathscr{C}(n,\{1,2,\ldots,n-1\}))$$

This sequence is shown for $n = 5$ in the bottom row of Fig. 10 (going from right to left). To achieve an interpolating sequence from the star network to the complete graph we note that $\mathscr{C}(n,\emptyset)$ has no edges. By letting $\mathbb{A}_0 = \emptyset$, we can obtain the desired sequence by simply adding the edges of the star network, $\{(0,i) : i \in \{1,2,\ldots,n\}$, to the edge set of each $\mathscr{C}(n,\mathbb{A}_i)$ in $(\mathscr{C}(n,\mathbb{A}_i))_{i=0}^{n-2}$, as shown in the top row of Fig. 10 for $n = 5$. Putting this sequence of networks between $\star_n$ and k_n and the other between k_n and p_n we get a total of $2n-2$ networks (including $\star_n$, k_n and p_n). This sequence can be generated for any n using the function `star2Complete2Path(n)` in Sect. A.3.

We can finally see in Fig. 11 how the probability density function of the MLEs, $(\widehat{\alpha},\widehat{\beta})$, change as we sequentially vary the SICN in the family that interpolates from

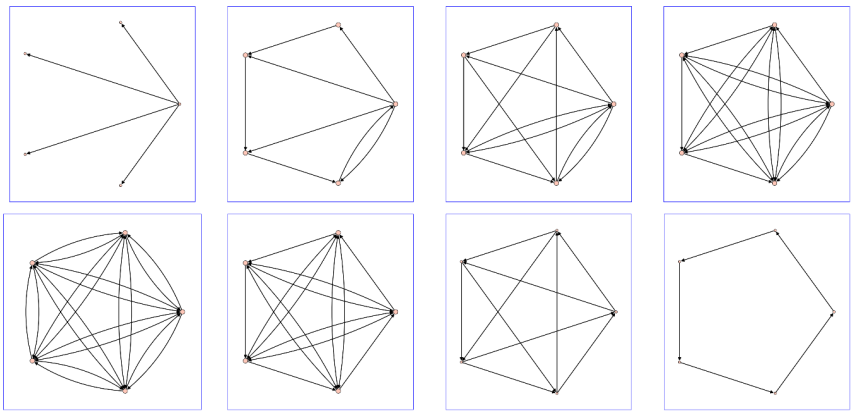

Fig. 10 A path from star network to circular path network through the complete network with 5 vertices.

the star network (red hue) to the circular path network (pink hue) via the complete network (blue hue). The MLEs is based on 10 independent transmission trees simulated from each SICN in the sequence of 98 SICNs over a population of size $n = 50$. In Fig. 11, the hue of the PDFs sequentially change from red which is concentrated entirely on the boundary at 1 (star network), to orange and yellow which are decreasing their concentration at 1 due to disappearance of the star's signal from the larger neighbourhoods of the circulant graphs $\mathscr{C}(n, \mathbb{A}_i)$. As i approaches $n-1$ the green and azure hues of the PDFs become increasingly uniform around blue when the SICN is the complete network. The hue of the PDFs become purple and start concentrating at 0 as the SICN approaches the path network that is fully concentrated at 0 (pink hue). The pattern of the PDFs is stochastic since it is based on MLEs from just 10 samples. However, it clearly demonstrates that the interpolating sequence in the space of SICNs does convey continuity in the parameter space of (α, β). In other words, this suggests that there is an (α, β) under the beta-splitting model (recall that the beta-splitting model need not explicitly refer to the contact network), that best fits the distribution of transmission trees generated from any specific contact network.

4 Discussion

We give a probabilistic description of the transmission process in Sect. 2 as a Markov chain on the product space of SI-tagged contact networks and transmission trees in discrete and continuous time. The Markov chain is also constructed as a randomized algorithm in the SageMath/Python code in Sect. A.1. This formalizes a large class of simulation programs in the computational epidemiology literature as a transmission process. The probabilities of transmission trees as an explicit function of both branch-lengths and tree topologies are derived in Sect. 2.1 from the general Markov chains of Eqs. (2.1) and (2.2) for some simple static contact networks.

Although the Markov chain model is general and only needs a directed weighted graph, our examples were limited to simple connected networks. It is straightforward

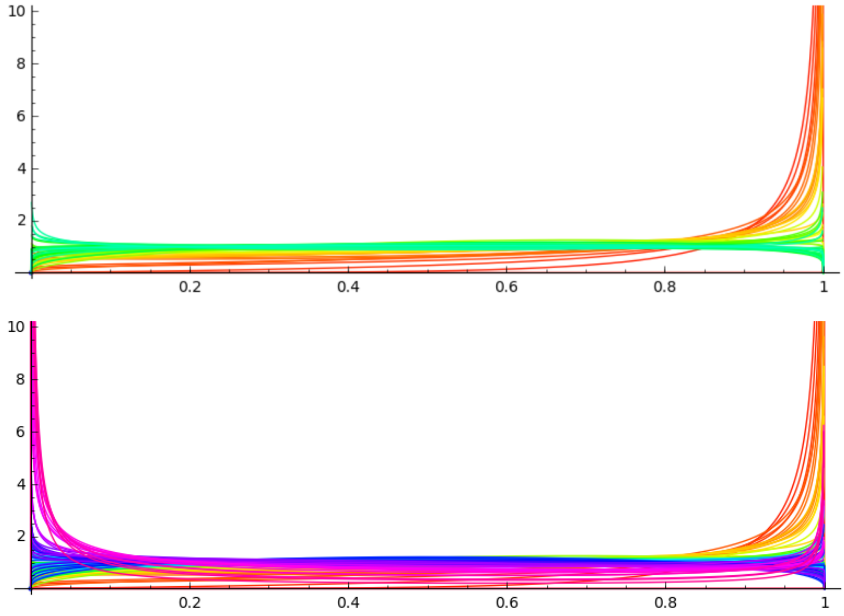

Fig. 11 Probability density function (PDF) of the $B(\alpha+1,\beta+1)$ distribution at the maximum likelihood estimates of α and β) based on 10 sampled transmission trees from each SICN in the sequential family that interpolates from the star network (red hue) to the circular path network (pink hue) via the complete network (blue hue) with $n = 50$ nodes. The hue of the PDFs sequentially change from red (star network), orange, yellow and green (complete network) as shown on the top plot and continue on with azure, blue, purple, to pink (path network) as shown in the bottom plot.

to consider the dynamics on more general networks using the richer language for digraphs (Pastor-Satorras et al, 2015, Fig. 4). For example, the epidemic will spread to the strongly connected giant component (if it exists) and the giant out-component, provided the infection starts from one of the nodes in either the strongly connected giant component or in a giant in-component.

We develop a biparametric Beta-splitting family of models for the growth of transmission trees via pure birth events in Sect. 3 that gives the exact probability of any transmission tree as a function of $\alpha > -1$ and $\beta > -1$. The approach avoids the explicit modeling of the underlying contact network (that is typically unobserved) in order to grow transmission trees, unlike the general Markov chain models of Eqs. (2.1) and (2.2). The Beta-splitting family of models is show analytically to contain the models generated by the complete network (k_n) when (α,β) equals $(0,0)$, star network ($\star_n$) when $(\alpha,\beta) \to (\infty,-1)$ and path network (p_n) when $(\alpha,\beta) \to (-1,\infty)$. We also derive explicit expressions for maximum likelihood estimators for the Beta-splitting model from independent observations of the transmission trees (which can be extended to the case of samples of several smaller transmission trees). Finally, the model can be interpreted in terms of a Beta-splitting construction for the "infection potential" of the infector and the infectee. Thus, the model captures aspects of the un-

derlying contact network up to how its contact structure affects the infection potential of the infector and infectee after the infection event.

We have also shown by simulations coupled with an inferential maximum likelihood procedure that the best-fitting parameters for a sequential family of SI-tagged contact networks from $\star_n$ to k_n to p_n do indeed follow a path in $(-1,\infty)^2$, the parameter space, from $(\infty,-1)$ to $(0,0)$ to $(-1,\infty)$. We conjecture that there is an equivalence class of SI-tagged contact networks that are indistinguishable by their transmission tree distributions for some given $(\alpha,\beta) \in (-1,\infty)^2$. As a trivial example we already saw that the circular and linear path networks have identical transmission tree distributions. Let $\widehat{L}(w) = (\alpha,\beta) : \mathscr{S}_n \to (-1,\infty)^2$ map each (connected) contact network w in $\mathscr{S}_n$ to the exact maximum likelihood estimate (α,β), i.e., $\widehat{L}(w)$ transforms the distribution on transmission trees induced by w to the MLE (α,β) in the parameter space over the quarter plane $(-1,\infty)^2$. $\mathscr{S}_n$ is the poset under subset ordering of the connected elements of 2^{w_n}, the power set of the edge set w_n of k_n, the complete network with unit edge-weights. We can use $\widehat{L}(w)$ to map each contact network $w \in \mathscr{S}_n$ to its MLE in $(-1,\infty)^2$ while maintaining the partial ordering between contact networks. Such a planar geometric embedding of the contact networks into the quarter-plane can help one gain a more systematic understanding of the connection between the transmission tree distributions specified by the beta-splitting model and that specified directly by the contact network.

Although random graph models of contact networks add another level of randomness, we can informally think of a static contact network as a typical realization of a random graph model (Aldous, 2013, Sec. 2.5). Thus the transmission process on any given static contact network can be used to provide insights into the sampling distribution of transmission trees for a large class of random graph models already available in SageMath's graph libraries. For example, the following code:

```
ts=[transmissionProcessTC(graphs.RandomRegular(k,n).to_directed(),0)
                                                    for _ in range(1000)]
```

can produce 1000 independent samples of transmission trees from 1000 independent realizations of the random k-regular graph over n nodes.

The jump Markov chain of the transmission process on static SI-tagged contact networks is a prerequisite for contemplating appropriate partial orders on the set of all SI-tagged contact networks in order to define natural transitions in the state space that can allow for contact networks to vary in time by possibly depending on the current state of the tagged contact network as well as the transmission tree – a natural state space for formalizing epidemics over adaptive or coevolving contact networks. Such adaptive contact networks are known in simulation studies to be highly sensitive to the structure of the initial contact network (see Pastor-Satorras et al (2015, VII.B.7) and the references therein). Future research on Markov chains with transitions over partially ordered contact networks (that could be geometrically embedded in the quarter-plane by their Beta-splitting MLEs as described above) as well as transmission trees could build upon insights from this simpler setting of transitions over static SI-tagged contact networks and partially ordered transmission trees.

We hope that tractable extensions of the transmission process from this most basic and fundamental setting of the SI model will be pursued in the future. By considering

birth and death processes, as opposed to a pure birth process, we can make progress on developing transmission processes for the more complex SIS epidemic model that not only allows susceptible individuals to become infected by any infected individual at a given 'birth' rate but also allows infected individuals to become susceptible again according to a given 'death' rate. Extensions to SIR model which allows for the 'removal' of infected individuals from the population at a given rate is conceivable via mapping to percolation on semi-directed networks (see Pastor-Satorras et al (2015, V.B.4) and the references therein).

We only looked at the resolution of leaf-labeled and leaf-unlabeled transmission trees with and without branch-lengths in this work. Transmission trees are rooted, binary, ranked, and planar. Fortunately, it is straightforward to carry over these probabilities to planar unranked trees, nonplanar ranked trees and nonplanar unranked trees using the explicit formulae and code in Sainudiin and Veber (2015). These formulae can be used to conduct simulation intensive inference based on projections of the transmission trees onto coarser tree shape statistics or used as prior distributions to constrain the micro-structure of the continuum of contacting hosts in space-time within which the pathogens can evolve through transmission events.

Acknowledgements RS thanks Robert C. Griffiths for combinatorial guidance on planar binary trees, Bhalchandra Thatte and Charles Semple for pointers on partially ordered sets of networks. This work grew out of a lecture in a postgraduate biometrics course at Cornell University (2014) and was refined by feedback from Tom Britton, Mia Deijfen, Federica Giardina and Pieter Trapman at an Epidemics Group meeting in Stockholm University (2015). RS was partly supported by a Sabbatical Grant from College of Engineering, University of Canterbury, a Visiting Scholarship at Department of Mathematics, Cornell University, Ithaca, NY, USA, consulting revenues from Wynyard Group and by the chaire Modélisation Mathématique et Biodiversité of Veolia Environnement-École Polytechnique-Museum National d'Histoire Naturelle-Fondation X. DW was partially supported by Marsden grant UOA1324 from the Royal Society of New Zealand.

References

Aldous D (2013) Interacting particle systems as stochastic social dynamics. Bernoulli 19(4):1122–1149, DOI 10.3150/12-BEJSP04

Aldous DJ (2001) Stochastic models and descriptive statistics for phylogenetic trees, from Yule to today. Statist Sci 16(1):23–34

Andersson H, Britton T (2000) Stochastic Epidemic Models and Their Statistical Analysis. Lecture Notes in Statistics, Springer New York

Britton T, O'Neill P (2002) Bayesian Inference for Stochastic Epidemics in Populations with Random Social Structure. Scandinavian Journal of Statistics 29(3):375–390

Colijn C, Gardy J (2014) Phylogenetic tree shapes resolve disease transmission patterns. Evolution, Medicine, and Public Health

Colless DH (1982) Review of phylogenetics: the theory and practice of phylogenetic systematics. Systematic Zoology 31:100–104

Developers TS (2015) Sage Mathematics Software (Version 6.8). URL http://www.sagemath.org

Flajolet P, Sedgewick R (2009) Analytic Combinatorics, 1st edn. Cambridge University Press, New York, NY, USA

Frost SD, Pybus OG, Gog JR, Viboud C, Bonhoeffer S, Bedford T (2015) Eight challenges in phylodynamic inference. Epidemics 10:88–92, challenges in Modelling Infectious {DIsease} Dynamics

Frost SDW, Volz EM (2013) Modelling tree shape and structure in viral phylodynamics. Philosophical Transactions of the Royal Society of London B: Biological Sciences 368(1614), DOI 10.1098/rstb.2012.0208

Grenfell BT, Pybus OG, Gog JR, Wood JL, Daly JM, Mumford JA, Holmes EC (2004) Unifying the epidemiological and evolutionary dynamics of pathogens. science 303(5656):327–332

Groendyke C, Welch D, Hunter DR (2011) Bayesian inference for contact networks given epidemic data. Scandinavian Journal of Statistics 38(3):600–616, DOI 10.1111/j.1467-9469.2010.00721.x

Groendyke C, Welch D, Hunter DR (2012) A network-based analysis of the 1861 hagelloch measles data. Biometrics 68(3):755–765, DOI 10.1111/j.1541-0420.2012.01748.x

Haydon DT, Chase–Topping M, Shaw DJ, Matthews L, Friar JK, Wilesmith J, Woolhouse MEJ (2003) The construction and analysis of epidemic trees with reference to the 2001 uk foot–and–mouth outbreak. Proceedings of the Royal Society of London B: Biological Sciences 270(1511):121–127

Holme P (2015) Modern temporal network theory: a colloquium*. Eur Phys J B 88(9):234, DOI 10.1140/epjb/e2015-60657-4

Hudson R (1990) Gene genealogies and the coalescent process. Oxford Surv Evol Biol 7:1–44

Kingman JFC (1982) The coalescent. Stochastic Processes and their Applications 13:235–248

Leigh Brown AJ, Lycett SJ, Weinert L, Hughes GJ, Fearnhill E, Dunn DT (2011) Transmission network parameters estimated from hiv sequences for a nationwide epidemic. Journal of Infectious Diseases 204(9):1463–1469

Leventhal GE, Kouyos R, Stadler T, von Wyl V, Yerly S, Böni J, Cellerai C, Klimkait T, Günthard HF, Bonhoeffer S (2012) Inferring epidemic contact structure from phylogenetic trees. PLoS Computational Biology 8(3):e1002,413

Lloyd-Smith JO, Schreiber SJ, Kopp PE, Getz W (2005) Superspreading and the effect of individual variation on disease emergence. Nature 438(7066):355–359

McKenzie A, Steel M (2000) Distribution of cherries for two models of trees. Math Biosci 164:81–92

Notohara M (1990) The coalescent and the genealogical process in geographically structured population. Journal of Mathematical Biology 29(1):59–75

O'Dea EB, Wilke CO (2010) Contact heterogeneity and phylodynamics: how contact networks shape parasite evolutionary trees. Interdisciplinary perspectives on infectious diseases 2011

Pastor-Satorras R, Castellano C, Van Mieghem P, Vespignani A (2015) Epidemic processes in complex networks. Rev Mod Phys 87:925–979, DOI 10.1103/RevModPhys.87.925

Rasmussen DA, Volz EM, Koelle K (2014) Phylodynamic inference for structured epidemiological models. PLoS Comput Biol 10(4):e1003,570

Romero-Severson E, Skar H, Bulla I, Albert J, Leitner T (2014) Timing and order of transmission events is not directly reflected in a pathogen phylogeny. Molecular Biology and Evolution 31(9):2472–2482

Sackin MJ (1975) "Good" and "bad" phenograms. Systematic Zoology 21:225–226

Sainudiin R, Veber A (2015) A beta-splitting model for evolutionary trees. UCDMS Research Report 2015/3 pp 1–20, URL http://www.math.canterbury.ac.nz/~r.sainudiin/preprints/20151129_betaSplitEvolTree.pdf

Stadler T, Bonhoeffer S (2013) Uncovering epidemiological dynamics in heterogeneous host populations using phylogenetic methods. Philosophical Transactions of the Royal Society of London B: Biological Sciences 368(1614), DOI 10.1098/rstb.2012.0198

Vaughan TG, Khnert D, Popinga A, Welch D, Drummond AJ (2014) Efficient bayesian inference under the structured coalescent. Bioinformatics 30(16):2272–2279

Wallinga J, Teunis P (2004) Different epidemic curves for severe acute respiratory syndrome reveal similar impacts of control measures. American Journal of Epidemiology 160:509–516

Welch D (2011) Is network clustering detectable in transmission trees? Viruses 3(6):659–676, DOI 10.3390/v3060659

Ypma RJF, van Ballegooijen WM, Wallinga J (2013) Relating phylogenetic trees to transmission trees of infectious disease outbreaks. Genetics 195(3):1055–1062

Yule GU (1924) A mathematical theory of evolution: based on the conclusions of Dr. J.C. Willis. Philos Trans Roy Soc London Ser B 213:21–87

A Code

This code is publicly shared in sagemathcloud at https://cloud.sagemath.com/projects/58dfa924-55ae-4b6c-9fd4-1cd0ef49eb7c/files/2015-10-25-165503.sagews. The code was mainly used to aid intuition during this study and is not written to be efficient for large scale simulation studies. The code is presented here instead of pseudo-code in order to communicate the Algorithms used in this study in a more concrete and reproducible manner. This also allows the reader to perform computational experiments in sage/Python immediately to further extend this work.

A.1 Simulating the Transmission Process

```
LBT = LabelledBinaryTree

def markAsInfected(C,v,m):
    '''mark node v as infected with marker m on each of the incoming edges of v in SICN C'''
    for e in C.incoming_edge_iterator([v]):
        C.set_edge_label(e[0],e[1],m)

def susceptibleOutEdges(C,vs):
    '''return the the susceptible outedges of node v in vs in SICN C'''
    SOE = [e for e in C.outgoing_edge_iterator(vs) if e[2]==None]
    return SOE

def growTransmissionTree(Ttree, pDict, z, infector, infectee):
    '''grow the transmission tree Ttree and update pathsDict pDict by adding the
        z-th infection event with infector -> infectee '''
    newSubTree = LBT([LBT([None,None], label=infector),
                      LBT([None, None], label=infectee)], label=z).clone()
    path2Infector = pDict[infector]
    if z==1:
        Ttree = newSubTree
    else:
        Ttree[tuple(path2Infector)] =newSubTree
    pDict[infector]=path2Infector+[0]
    pDict[infectee]=path2Infector+[1]
    pDict[z]=path2Infector
    return Ttree

def forgetLeafLabels(T):
    '''return the transmission tree T with all leaf labels set to 0'''
    leafLabelSet=set(T.leaf_labels())
    leafUnlabelledT=T.map_labels(lambda z:0 if (z in leafLabelSet) else z)
    return leafUnlabelledT

def forgetAllLabels(T):
    '''return the transmission tree T with all node labels removed'''
    return T.shape()

def justTree(T):
    '''return the transmission tree T as nonplanar unranked unlabelled tree'''
    return  Graph(T.shape().to_undirected_graph(),immutable=True)

def transmissionProcessTC(C,initialI):
    '''return transmission tree outcome of the DTDS transmission MC on SICN C with
       initial infection at node initialI'''
    #initialisation of SICN
    z=0 # infection event count
    ExpectedTimes=[] # vector of expected waiting times
    markAsInfected(C,initialI,'infected')
    infectedIs = [initialI]
    popSize=C.order()
    # initialisation of Transmission Tree
    pathsDict={} # dictionary of nodes -> paths from root in tree
    # individuals in tree are labelled by "i"+str(integer_label)
```

```
    T = LBT([None,None],label="i"+str(initialI)).clone()
    pathsDict["i"+str(initialI)]=[]
    while (len(infectedIs) < popSize):
        z=z+1 # increment infection event count
        currentSOE = susceptibleOutEdges(C,infectedIs)
        ExpectedTimes.append(1/len(currentSOE))
        nextEdge = currentSOE[randrange(0,len(currentSOE))]
        C.set_edge_label(nextEdge[0],nextEdge[1],z)
        infectedIs.append(nextEdge[1])
        markAsInfected(C,nextEdge[1],'inf')
        T=growTransmissionTree(T, pathsDict, z, "i"+str(nextEdge[0]),"i"+str(nextEdge[1]))
        print "step z = ",z; print ascii_art(T); print "--------------------"
    return [T.as_ordered_tree(with_leaves=False), ExpectedTimes]

# demo
sage: transmissionProcessTC(graphs.CompleteGraph(4).to_directed(),0)
# output
step z =  1
  1_
 /  \
i0   i3
--------------------
step z =  2
    __1__
   /     \
  2_      i3
 /  \
i0   i1
--------------------
step z =  3
    ___1___
   /       \
  2_        3_
 /  \      /  \
i0   i1   i3   i2
--------------------
[1[2[i0[], i1[]], 3[i3[], i2[]]], [1/3, 1/4, 1/3]]
```

A.2 Likelihood of Beta-splitting Transmission Trees

```
def splitsSequence(T):
    '''return a list of tuples (left,right) split sizes at each split node'''
    l = []
    LabelledBinaryTree(T).post_order_traversal(lambda node:
       l.append((node[0].node_number(),node[1].node_number())))
    return l

def prob_RPT(T,a,b):
    '''probability of ranked planar tree T under beta-splitting model
       a,b>-1, where (a+1,b+1) are the parameters of the beta distribution'''
    return prod(map(lambda x: beta(x[0]+a+1,x[1]+b+1)/beta(a+1,b+1),
                              splitsSequence(T)))

# demo of the mle for a complete graph with 50 nodes and 10 sampled trees
c_1 = lambda p: p[0]+0.9999999 # constraint for alpha > -1
c_2 = lambda p: p[1]+0.9999999 # constraint for beta > -1
n=50
reps=10
ts=[transmissionProcessTC(graphs.CompleteGraph(n).to_directed(),0)  for _ in range(reps)]
def negLkl(AB):
        return sum([-log(1.0*prob_RPT(ts[j],AB[0],AB[1])) for j in range(reps)])
mle=minimize_constrained(negLkl,[c_1,c_2],[0.0,0.0],disp=0)
[n,reps,mle]

[50, 10, (0.012116930598591959, -0.08408627968100374)]
```

```
# another execution of the above demo block to show variability in MLE
[50, 10, (-0.04056211875882902, -0.03788115149950636)]
```

A.3 A family of contact networks interpolating the star, complete and path networks

```
def star2Complete2Path(n):
    '''list of digraphs from star to complete to circular path with n vertices'''
    connects=[]
    for i in range(1,n+1):
        connects.append(range(1,i))
    L=[]
    for i in range(0,n):
        g=digraphs.Circulant(n,connects[i])
        g.add_edges([(0,i) for i in range(n)])
        L.append(g)
    for i in range(n-2,0,-1):
        g=digraphs.Circulant(n,connects[i])
        L.append(g)
    return L
```

A.4 Transmission Tree Distributions

See (Sainudiin and Veber, 2015, Appendix: Algorithms) for simulating transmission trees and obtaining the probability for various equivalence classes of trees under the Beta-splitting model.