The Transmission Process: A Combinatorial Stochastic Process for the Evolution of Transmission Trees over Networks

Raazesh Sainudiin[a,b,*], David Welch[c]

[a] Laboratory for Mathematical Statistical Experiments, Christchurch Centre and Biomathematics Research Centre, School of Mathematics and Statistics, University of Canterbury, Private Bag 4800, Christchurch 8041, New Zealand
[b] Current address: Department of Mathematics, Stockholm University, SE - 106 91 Stockholm, Sweden.
[c] Computational Evolution Group and Department of Computer Science, University of Auckland, Private Bag 92019, Auckland 1142, New Zealand

Abstract

We derive a combinatorial stochastic process for the evolution of the transmission tree over the infected vertices of a host contact network in a susceptible-infected (SI) model of an epidemic. Models of transmission trees are crucial to understanding the evolution of pathogen populations. We provide 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 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 with exact probabilities as a function of the model parameters, but without explicitly modeling the underlying contact network, and show that for specific values of the parameters we can recover the exact probabilities for our three example networks through the Markov chain construction that explictly models the underlying contact network. We use the maximum likelihood estimator (MLE) to consistently infer the two parameters driving the transmission process based on observations of the transmission trees and use the exact MLE to characterize equivalence classes over the space

[*] Corresponding author
Email addresses: raazesh.sainudiin@gmail.com (Raazesh Sainudiin), david.welch@auckland.ac.nz (David Welch)

of contact networks with a single initial infection. An exploratory simulation study of the MLEs from transmission trees sampled from three other deterministic and four random families of classical contact networks is conducted to shed light on the relation between the MLEs of these families with some implications for statistical inference along with pointers to further extensions of our models. The insights developed here are also applicable to the simplest models of "meme" evolution in online social media networks through transmission events that can be distilled from observable actions such as 'likes', 'mentions', 'retweets' and '+1s' along with any concomitant comments.

Keywords: rooted ranked planar binary tree, contact network, susceptible-infected epidemic model, non-parametric combinatorial stochastic process, parametric Beta-splitting model, random graph models, meme evolution

MSC: 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*. The transmission tree represents the physical continuum of contacting hosts and thus frames the host-level structure within which pathogens are transmitted in a communicable disease. Therefore, models of transmission trees are crucial to understanding the evolution of pathogen populations. Constructing models of transmission trees is the main focus of this paper. Although we limit ourselves here to the epidemiological context of transmissions of a communicable disease over a contact network of hosts for concreteness of language and notions from a field with a longer research history, most of our basic results and insights are naturally applicable, as briefly discussed in Sect. 5.2, to the cultural context of transmissions of "memes" [1, p. 192] over a social network of individuals, such as Twitter [2]. More generally, they can be used to model transmission events in *Finite Markov Information Exchange* processes [3, Sec. 2.2] as described below.

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 ι_i to ι_j, given by a non-negative $w_{i,j} \in [0,\infty)$, represents the propensity with which the infection can be transmitted from ι_i to ι_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 ι_i and ι_j, for instance (this is the "meeting process" of [3]). 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 vertices labeled by individuals $\iota_1, \iota_2, \ldots, \iota_9$ (the edges are undirected). Note that we sometimes label the vertices starting from ι_0 to stay true to the indexing convention in sageMath/python (but this should be clear from the context).

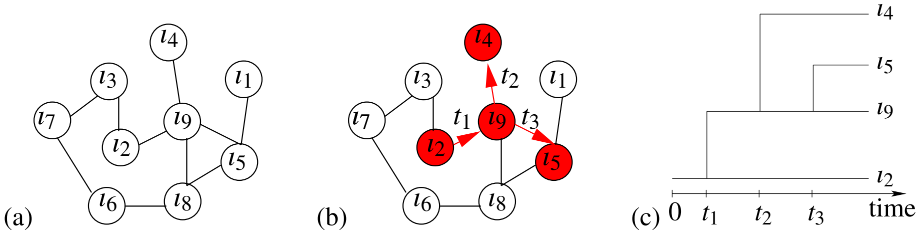

Figure 1: Spread of an epidemic over (a) the contact network of a population as shown by (b) a sub-network where edges representing transmission events are labelled by the time of event and the infected vertices are colored red and (c) the corresponding transmission tree.

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 with its individual vertices further "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* is a directed edge-labeled subgraph of the contact network containing all infected vertices and directed edges labelled by the time of transmission. It is a basic object of interest and 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 vertices 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 [4].

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 [5] and [6], number of cherries or sub-terminal vertices [7], 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 and total time of epidemic.

Furthermore, by a natural extension of the pure-birth process underpinning the SI model to a birth-and-death process that is combinatorially more involved with an additional epidemic state indicating whether the individual is 'removed' (R) from the population, one can extend the transmission process developed here for the SI epidemic model to the more realistic susceptible-infected-recovered (SIR) epidemic model. With such an extension, which we won't pursue in this elementary study of the simplest SI epidemic model (for reasons explained below), the leaves of the SIR transmission trees will not only be tagged by I but also by R and they will naturally capture various univariate statistics of interest to applied epidemiologists including the final-size or total number of infections [8, 9, 10]. We outline a set of combinatorial steps needed towards such a future direction of work in Sect. 5.1.

While various analytical results [eg. 4] and computationally intensive methods [eg. 10] are available for various univariate epidemiological statistics and can often be obtained without explicitly modelling the tree, most insights about the structural information in the tree (even for the simplest SI epidemic model) 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 11, 12] 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 [13, 14]). This insight has seen the rise of a new area of research, known as phylodynamics [15], 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 [16].

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 [17] investigated a number of standard random network models (Erdős-Rényi, Barabási-Albert preferential attachment and Watts-Strogatz small-world – we also analyze these networks via simulations in Sect. 4.2 in addition to seven other families of contact networks in Sect. 4) 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 [18] suggest that while this effect is real, it may be swamped by other effects such as sampling strategy. O'Dea and Wilk [19] 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 [20] 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 [21] describes a method that roughly classifies epidemics into host population structures such as homogeneous, super-spreading [22] 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 [23]. There is a series of papers [24, 25, 26] that uses time-series data from epidemics to infer the parameters of an ERGM but the transmission tree here is incidental and poorly inferred. It is suggested in [26] 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. Appendix 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 in [27] and [28] summarize this literature.

Basic population genetic models such as the coalescent [29] 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 [30, 31] or multi-type branching process [32] 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 [33, 32, 34] 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 [27, 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 given by 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 [3, Sec. 2.2] called the *Pandemic Process* [3, Sec. 7] that is shown to be a fundamental building-block [3, Sec. 3.2,7] for a large class of FMIE processes which includes various classical epidemic models [see 3, Sec. 8,9 and references therein]. For instance, the SI model exhibits the fastest possible spread of information in any FMIE model [3, 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 [27, 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 in Sect. 3 that has the following properties:

- gives the exact probability of any transmission tree as a function of α and β (Theorem 1),
- 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 approaches $(-1,\infty)$, respectively,
- has explicit expressions for its maximum likelihood estimators from independent observations of transmission trees and their sufficient statistics (Theorem 2),
- specifies an equivalence class of contact networks that have the same (α,β)-specified distribution of transmission trees (Theorem 3) and
- is amenable to exact probability calculations for various equivalence classes of transmission trees as rooted (leaf-unlabelled) trees that are (un)ranked/(non)planar

with or without continuous branch-lengths based on the results for the same Beta-splitting model, but studied in the context of species diversification [35].

The Beta-splitting model is the third and perhaps 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 [3, 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.

Finally in Sect. 4 we explore, mostly by simulations, the nature of distributions on transmission trees that are induced by classical families of three other deterministic and four random contact networks through their most likely Beta-splitting model parameters. Furthermore, using a sequential family of contact networks that interpolates in the space of contact networks from the star network to the circular path network via the complete network by means of edge addition and deletion operations, we show that the maximum likelihood estimate of the Beta-splitting model that are obtained from the induced transmission trees over each contact network in the family also smoothly interpolates, in the parameter space $[-1,\infty]^2$, between that for the complete, star and path networks. These insights lead to some implications for statistical inference as described in Sect. 4.4. This is the fourth and final contribution of our paper.

In summary, the rest of the paper is organized 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 deterministic networks. In Sect. 3 we introduce a parametric Beta-splitting model for the transmission tree, derive the likelihood for a given tree, explore the relationship between this Beta-splitting model and the coupled transmission tree-contact network Markov chain model described in Sect. 2, obtain sufficient statistics, derive numerically robust expressions for the maximum likelihood estimator, and characterize the equivalence class of contact networks with the same Beta-splitting model of transmission trees. In Sect. 4 we gather insights through the most likely Beta-splitting models fitted to independent samples drawn from the distributions over transmission trees that are induced by various deterministic and random contact networks carefully chosen from several classical families of networks with implications for statistical inference. In Sect. 5 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 $w_n := \{w(\iota_i, \iota_j) \in [0, \infty) : \iota_i \neq \iota_j, (\iota_i, \iota_j) \in \mathbb{I}_n^2\}$. 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 vertex, respectively, subtending from the internal vertex labelled by m in $\tau(z)$, the transmission tree at time z. Let $\mathbb{L}(\tau(z))$ denote the set of leaf vertices, i.e., the set of potential infectors, of $\tau(z)$ and let $w(\iota_i,\iota_j;c(z))$ denote the weight of the edge between vertices labelled by ι_i and ι_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:

$$\begin{aligned}&\Pr\left((\tau(z+1),c(z+1))|(\tau(z),c(z))\right)\\&=\begin{cases}\dfrac{w(L(z+1;\tau(z+1)),R(z+1;\tau(z+1));c(z))}{\sum\limits_{\forall\iota_\ell\in\mathbb{L}(\tau(z))}\ \sum\limits_{\substack{\forall\iota_j\in\mathbb{I}_n:\\ s_j(z)=1}} w(\iota_\ell,\iota_j;c(z))} & \text{if } (\tau(z),c(z))\prec(\tau(z+1),c(z+1))\\[2em] 0 & \text{otherwise.}\end{cases}\end{aligned} \tag{2.1}$$

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 $w(\iota_\ell,\iota_j;c(z))$ from every potential infector, i.e., $\forall\iota_\ell\in\mathbb{L}(\tau(z))$, to every potential infectee within its susceptible out-neighborhood of the SICN at time z, i.e., $\forall\iota_j\in\mathbb{I}_n$ 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 [36] in Sect. Appendix A.1.

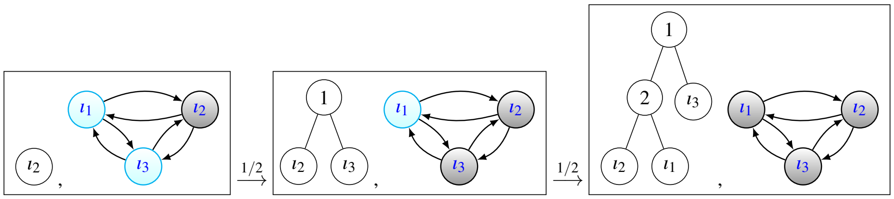

Figure 2: A sequence of states from the product state space of transmission trees and contact networks in the discrete-time discrete-space jump Markov chain embedded in the transmission process. Initially (left panel) the transmission tree has the root vertex labelled by the first infected individual $\iota_{i(0)} = \iota_2$ with the corresponding complete contact network k_3 with vertices 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 vertex labelling the first infection event by 1 and the first infector ι_2 labelling its left leaf vertex and the first infectee $\iota_{i(1)} = \iota_3$ labelling its right leaf vertex (middle panel). In the final absorbing state (right panel), with probability $1/2$, the transmission tree has a new internal vertex labelled by 2 for the second infection event with its left leaf vertex labelled by the second infector ι_3 and its right leaf vertex labelled by the second infectee $\iota_{i(2)} = \iota_1$.

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 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. 5 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 $w(\imath_i, \imath_j) = 1$ for each $\imath_i \neq \imath_j$, 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:

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

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

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

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 $\tau(m)$ with m infection events, where $1 \leq m < n$ is

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

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

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

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 $\imath_{i(0)} = \imath_\star \in \mathbb{I}_n$ and the initial SI contact network is the star network, $\star_n$, centered at $\imath_\star$ with directed edge weights $\{w(\imath_\star, \imath_j) = 1 : \imath_j \in \mathbb{I}_n \setminus \imath_\star\}$, then since there are $n-z$ individuals in the non-empty susceptible out-neighborhood of the only possible infector $\imath_\star$ 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:

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

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

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

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

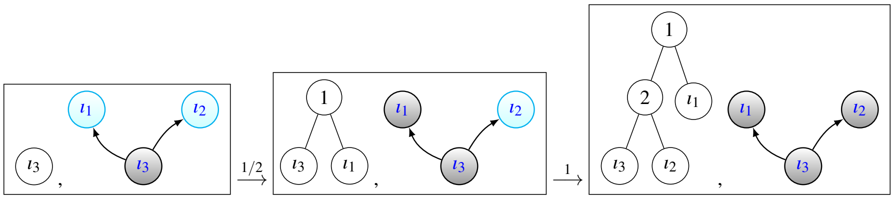

Figure 3: A sequence of states from the product state space of transmission trees and contact networks in the discrete-time discrete-space jump Markov chain embedded in the transmission process. Initially (left panel) the transmission tree has the root vertex labelled by the first infected individual $\imath_{i(0)} = \imath_\star = \imath_3$ with the corresponding star network $\star_3$ with vertices colored by their susceptible (lightly shaded) or infected status (darkly shaded) over a population of 3 individuals labelled by $\mathbb{I}_3 = \{\imath_1, \imath_2, \imath_3\}$. After the first transmission event from $\imath_3$ to $\imath_1$ with probability $1/2$, the transmission tree splits with the internal vertex labelling the first infection event by 1 and the first infector $\imath_3$ labelling its left leaf vertex and the first infectee $\imath_{i(1)} = \imath_1$ labelling its right leaf vertex (middle panel). In the final absorbing state (right panel), with probability 1, the transmission tree has a new internal vertex labelled by 2 for the second infection event with its left leaf vertex labelled by the second infector $\imath_3$ and its right leaf vertex labelled by the second infectee $\imath_{i(2)} = \imath_2$.

infection events, is:

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

Note that if $z = n-1$ and the entire population is infected then each of the discrete transmission trees, with the “left-branching comb” topology (ignoring the branch-lengths) such that the left-most leaf is labelled by the the first infected individual $\imath_{i(0)} = \imath_\star$ and the remaining $n-1$ leaves are labelled uniformly from $\mathbb{I}_n \setminus \imath_\star$, is equally likely as follows:

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

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 $\imath_{i(0)}$ is at the beginning or source vertex 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:

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

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

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

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

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

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

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

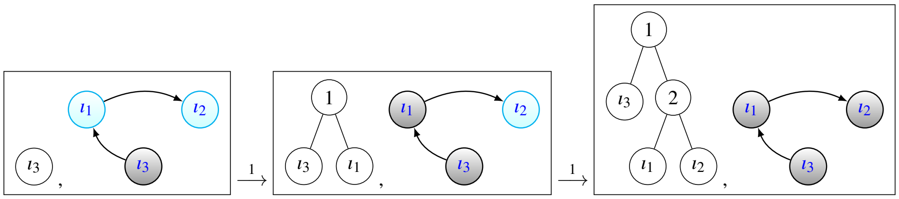

Figure 4: A sequence of states from the product state space of transmission trees and contact networks in the discrete-time discrete-space jump Markov chain embedded in the transmission process. Initially (left panel) the transmission tree has the root vertex labelled by the first infected individual $\iota_{i(0)} = \iota_3$ with the corresponding path network with directed edge set $\{(\iota_3,\iota_1),(\iota_1,\iota_2)\}$ and vertices 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 vertex labelling the first infection event by 1 and the first infector ι_3 labelling its left leaf vertex and the first infectee $\iota_{i(1)} = \iota_1$ labelling its right leaf vertex (middle panel). In the final absorbing state (right panel), with probability 1, the transmission tree has a new internal vertex labelled by 2 for the second infection event with its left leaf vertex labelled by the second infector ι_1 and its right leaf vertex labelled by the second infectee $\iota_{i(2)} = \iota_2$.

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 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 $E(T_z) =$

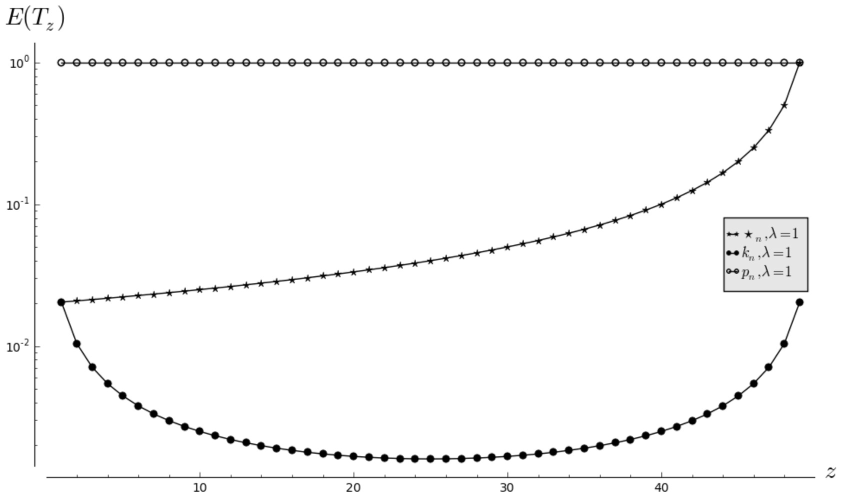

Figure 5: Expected branch-lengths when there are z infection events or $z+1$ infected individuals, $E(T_z)$, for the three cases. Here $n = 50$ and $\lambda = 1$. $E(T_z) = 1/\lambda = 1$ with the path network p_n of Sect. 2.1.3, $E(T_z) = 1/(\lambda(n-z)) = 1/(50-z)$ with the star network $\star_n$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

$1/(\lambda z \times (n-z))$ 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 the z-th infection event, will have mean length $1/(\lambda z \times (n-z))$, such that any one of the z infected leaf vertices 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

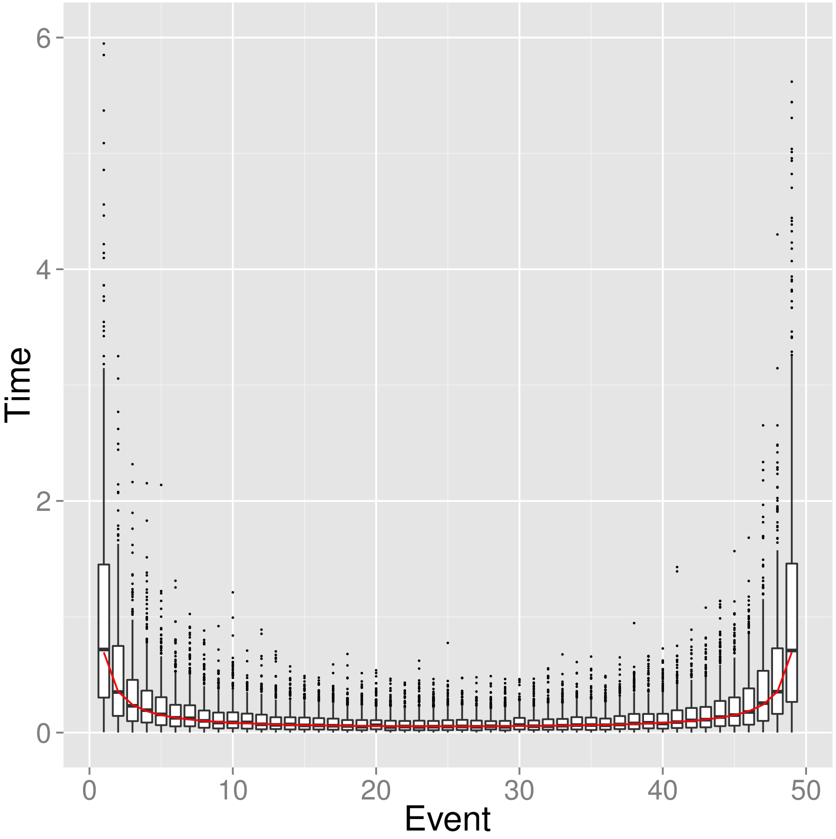

Figure 6: The sampling distribution of T_z, branch-lengths (times in y-axis) of the transmission tree when there are exactly z infected individuals or between the $(z-1)$-th and z-th infection event (x-axis), where $z \in \{1, 2, \ldots, n-1\}$, from 500 independent simulations of the transmission tree 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).

rate $z(n-z)/(n-1)$ and satisfies the following *randomly-shifted-logistic-limit* (see for eg. [3, Eq. 7.13]):

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

where, F is the logistic function:

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

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

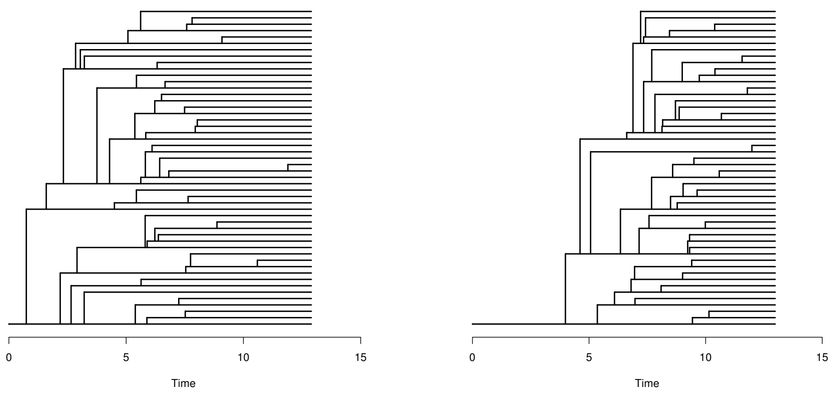

Figure 7: Two of the 500 independent simulations of the (unlabelled) transmission tree with branch-lengths over the complete SI contact network for a population of size 50 from Fig. 6. Notice the variation in branch-lengths (times between infection events) at the start and end of the epidemic when the 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 [35] 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 [37] 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 m leaves and $m-1$ internal vertices.

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 vertices 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 vertices 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 vertices will give the order in which these vertices 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 vertex in the next step. The left and right leaf vertices 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 m internal vertices and $m+1$ leaves. We start with a single root vertex, 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 vertex of the root into a left leaf and a right leaf labelled by $[0,b_1b_2]$ and $[b_1b_2,b_1]$, respectively. If $u_2\in[b_1,1]$, then instead split the right child vertex 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 m.

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 vertex 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 vertex 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 vertex 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 vertex 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 vertex 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 m to ensure $m+1$ leaves, where $1 \leq m \leq n-1$. Figure 8 shows an example of such a Beta-splitting tree construction for $m=3$.

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 m integer-labelled internal vertices or splits and $m+1$ unlabelled leaves. The process of assigning leaf labels from $\mathbb{I}_n$ via a pre-order traversal on the m internal vertices, in increasing order, i.e., Stage (2) of the construction, is described next.

We start with the internal vertex labelled 1 and assign the initial infected individual $\iota_{i(0)}$ to its left child vertex. Then we assign the first infectee to the right child vertex of 1. In general, as we descend down the internal vertices 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 vertex as the infectee ι_j chosen according

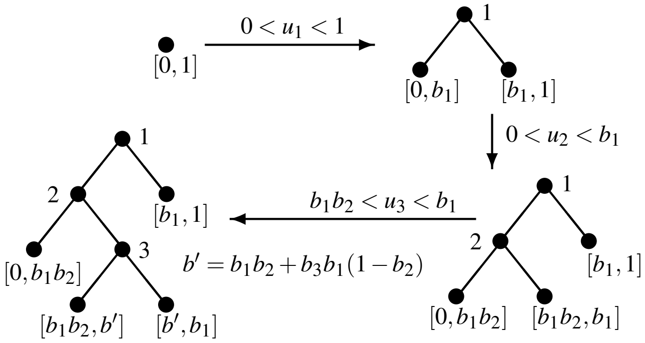

Figure 8: An example of a Beta-splitting tree construction for $m = 3$.

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 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):

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

3.1. Probability of a given Beta-splitting transmission tree

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

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

$$\begin{aligned}
\Pr\{\tau(m)\} &= \prod_{z=1}^{m} \left\{ \frac{1}{B(\alpha+1,\beta+1)} \int_0^1 b_z^{s_z^L+\alpha}(1-b_z)^{s_z^R+\beta} db_z \right\} \times \Pr\left(\textit{leaf labels}\right) \\
&= \prod_{z=1}^{m} \frac{B(s_z^L+\alpha+1, s_z^R+\beta+1)}{B(\alpha+1,\beta+1)} \times \Pr\left(\textit{leaf labels}\right), \qquad (3.4) \\
&= \prod_{z=1}^{m} \left(\frac{\prod_{j=0}^{s_z^R} \frac{\beta+j}{\beta+j+\alpha} \prod_{i=0}^{s_z^L} \frac{\alpha+i}{\alpha+i+\beta+s_z^R+1}}{\frac{\alpha\beta}{(\alpha+\beta)(\alpha+\beta+1)}} \right) \times \Pr\left(\textit{leaf labels}\right), \qquad (3.5)
\end{aligned}$$

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

$$\Pr\left(\textit{leaf labels}\right) = \Pr\{\iota_{i(0)}\} \prod_{z=1}^{m} \Pr\{(L(z);\tau(z)) \overset{z-1}{\rightsquigarrow} (R(z);\tau(z))\} \ . \qquad (3.6)$$

Proof outline. The second term in the product of Eq. (3.4) and Eq. (3.5) given by Eq. (3.6) 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 vertices of the tree after the initial infection with $\Pr\{\iota_{i(0)}\}$.

We now focus on the first term in Eq. (3.4) which results from integrating over the $(U_z,B_z)_{z\in[m]}$, for $1 \leq m \leq n-1$. 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 m 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 vertex z brings in another B_z in the product, or another $1-B_z$ if the split occurs in the right subtree below vertex 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.

Finally, to prove the first term in Eq. (3.5) we exploit the fact that s^L and s^R are non-negative integers and repeatedly apply the following elementary properties of the beta function:

$$B(x+1,y) = B(x,y)\frac{x}{x+y}, \quad B(x,y+1) = B(x,y)\frac{y}{x+y} \quad . \tag{3.7}$$

□

Remark 2. *Note that the expression for the probability of a transmission tree with m infection events given by Eq.* (3.5) *as a function of the parameters* α *and* β, *i.e., the likelihood function, only involves additions, multiplications and divisions. It is therefore numerically more robust during local optimization for maximum likelihood estimation in Sect. 3.3 than the expression in Eq.* (3.4), *which further requires numerical evaluations of the beta function.*

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

$$B(\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}, \qquad \alpha,\beta > 0, \tag{3.8}$$

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 m infections has the same probability as Eq. (2.5).

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

$$\prod_{z=1}^{m} \frac{B(s_z^L+1, s_z^R+1)}{B(1,1)} = \prod_{z=1}^{m} \frac{s_z^L! s_z^R!}{(s_z^L+s_z^R+1)!} = \frac{1}{m!}, \tag{3.9}$$

where the second equality is obtained by observing that $s_z^L + s_z^R + 1$ is the number of internal vertices of the tree rooted at vertex z, which is the left or the right subtree below the internal vertex z. Hence, each term $s_z^L!$ in the numerator of the product cancels with the term in the denominator that corresponds to the left child vertex of z, except if $s^L_z = 0$ and the left child vertex of z is a leaf. But in this case, $0! = 1$ by convention. The same holds true for each of the $s^R_z!$. Likewise, the terms in the denominator which are not compensated by some term in the numerator are those corresponding to internal vertices having no ancestral vertices. But the only such vertex is the root ($z = 1$) with $s^L_1 + s^R_1 + 1 = m$. 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:

$$\Pr\{\iota_{i(0)}\}\prod_{z=1}^{m}\Pr\{(L(z);\tau(z)) \overset{z-1}{\rightsquigarrow} (R(z);\tau(z))\} = \frac{1}{n}\prod_{z=1}^{m}\frac{1}{n-z} = \frac{(n-m-1)!}{n!} \qquad (3.10)$$

Finally, putting Equations (3.9) and (3.10) 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 3. *The transmission tree thus constructed with $\alpha = \beta = 0$ corresponds to [38] 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 m splits and $m+1$ unlabelled leaves. These $m!$ many trees are in bijective correspondence with permutations of $\{1,\dots,m\}$ through the* increasing binary tree-lifting *operation [39, 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

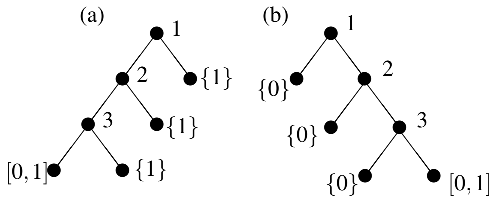

Figure 9: (a) The discrete transmission tree corresponding to the limiting case $\alpha \to \infty$ and $\beta \to -1$ is a left-branching comb, and (b) the discrete transmission tree corresponding to the limiting case $\beta \to \infty$ and $\alpha \to -1$ is a right-branching comb.

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 m with $m+1$ 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 $m+1$ leaves. See Fig. 9 for an example with $m = 3$. 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 at the source vertex. 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 $\prod_{z=1}^{m}\left(\frac{1}{n-z}\right)$ 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. Maximum likelihood estimation and sufficient statistics

In order to find the maximum likelihood estimates of α and β for the Beta-splitting model that give the most likely explanation for the transmission trees sampled from an arbitrary SICN (under the likelihood principle), we use the following inferential procedure:

Step 1: 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. Appendix A.1 r times.

Step 2: 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. Appendix A.2. Note that such additional information can be easily incorporated into more elaborate likelihood functions derived from Eq. (2.2) as outlined in Remark 6 and 7.

Theorem 2. *The likelihood of all r transmission tree topologies only depends on the sufficient statistic of* split-pair frequencies*:*

$$\{f(s^L,s^R):(s^L,s^R)\in\mathscr{S}_n\},$$
$$\textit{where }, \mathscr{S}_n := \{(s^L,s^R)\in\{0,1,\ldots,n-2\}^2 : s^L+s^R\leq n-2\} \ . \tag{3.11}$$

Therefore, the maximum likelihood point estimate for the Beta-splitting model based on r independent transmission trees, each with n leaves, is obtained by maximizing the per-vertex loglikelihood:

$$(\widehat{\alpha},\widehat{\beta}) = \operatorname*{arg\,max}_{(\alpha,\beta)\in(-1,\infty)^2} \sum_{(s^L,s^R)\in\mathscr{S}_n} \hat{P}(s^L,s^R)\log\left(\frac{\prod_{j=0}^{s^R}\frac{\beta+j}{\beta+j+\alpha}\prod_{i=0}^{s^L}\frac{\alpha+i}{\alpha+i+\beta+s^R+1}}{\frac{\alpha\beta}{(\alpha+\beta)(\alpha+\beta+1)}}\right) \ , \tag{3.12}$$

where, $\hat{P}(s^L,s^R) := f(s^L,s^R)/(n-1)r$ *and* $f(s^L,s^R)$ *is the frequency of the* $(n-1)r$ *many internal vertices across all r trees that have* s^L *and* s^R *many splits in their left and right sub-trees, provided* $n>1$.

Proof. From Eq. (3.4) in Theorem 1 and the assumption of independence across all r trees:

$$\prod_{i=1}^{r}\Pr(\tau_i;\alpha,\beta) = \prod_{i=1}^{r}\prod_{z\in\mathscr{I}(\tau_i)}\frac{B(s_z^L+\alpha+1,s_z^R+\beta+1)}{B(\alpha+1,\beta+1)} \tag{3.13}$$

$$= \prod_{(s^L,s^R)\in\mathscr{S}_n}\left(\frac{B(s^L+\alpha+1,s^R+\beta+1)}{B(\alpha+1,\beta+1)}\right)^{f(s^L,s^R)} \ . \tag{3.14}$$

Using the fact that s^L and s^R are non-negative integers we can exploit the properties of the beta function given by Eq. (3.7) to further simplify the likelihood function from Equation (3.14), as follows:

$$\prod_{i=1}^{r}\Pr(\tau_i;\alpha,\beta) = \prod_{(s^L,s^R)\in\mathscr{S}_n}\left(\frac{\prod_{j=0}^{s^R}\frac{\beta+j}{\beta+j+\alpha}\prod_{i=0}^{s^L}\frac{\alpha+i}{\alpha+i+\beta+s^R+1}}{\frac{\alpha\beta}{(\alpha+\beta)(\alpha+\beta+1)}}\right)^{f(s^L,s^R)} \ . \tag{3.15}$$

Thus, the likelihood function only depends on the transmission trees up to the *split-pair frequencies* as claimed above Eq. (3.11) in the Theorem.

Next we derive the maximum likelihood estimate (MLE) in Eq. (3.12). The MLE $(\widehat{\alpha},\widehat{\beta})$ obtained by maximizing the logarithm of the likelihood function in Eq. (3.15), for the purposes of point estimation, is equivalent to maximizing the per-vertex log-likelihood by ignoring the constant $(n-1)r$ as follows:

$$\begin{aligned}
&\underset{(\alpha,\beta)\in(-1,\infty)^2}{\arg\max}\ (n-1)r \sum_{(s^L,s^R)\in\mathscr{S}_n} \frac{f(s^L,s^R)}{(n-1)r} \log\left(\frac{\prod_{j=0}^{s^R}\frac{\beta+j}{\beta+j+\alpha}\prod_{i=0}^{s^L}\frac{\alpha+i}{\alpha+i+\beta+s^R+1}}{\frac{\alpha\beta}{(\alpha+\beta)(\alpha+\beta+1)}}\right)\\
&\quad = \underset{(\alpha,\beta)\in(-1,\infty)^2}{\arg\max} \sum_{(s^L,s^R)\in\mathscr{S}_n} \hat{P}(s^L,s^R) \log\left(\frac{\prod_{j=0}^{s^R}\frac{\beta+j}{\beta+j+\alpha}\prod_{i=0}^{s^L}\frac{\alpha+i}{\alpha+i+\beta+s^R+1}}{\frac{\alpha\beta}{(\alpha+\beta)(\alpha+\beta+1)}}\right) .
\end{aligned}$$

□

To focus on the jump chain's discrete structural information about the combinatorial skeletons of the contact network buried within the distribution over discrete transmission trees, as a necessary prelude to Theorem 3, our likelihood expressions in Eqs. (3.12), (3.13),(3.14) and (3.15) are *skeletal* and ignore leaf labels and the waiting times between events (see code in Sect. Appendix A.2). However, the likelihood function can be extended as discussed in Sect. 4.4.3.

The demonstration at the end of Sect. Appendix A.2 shows two independent MLE computations based on $r = 10$ independent transmission trees (without branch-lengths and leaf labels) that were sampled from the complete SICN on $n = 50$ vertices. The MLE $(\widehat{\alpha},\widehat{\beta})$ takes the following realizations: $(-0.0664,-0.0502)$ and $(0.0047,-0.0430)$. 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. The MLE is $(0.0279,0.0325)$ from another trial based on $r = 1000$ independent transmission trees drawn from the same complete SICN on 50 vertices. Figure 10 shows the sufficient statistics for the parameters α and β in these three trials.

3.4. Equivalence class of contact networks with the same Beta-splitting model

The maximum likelihood estimate and the sufficient statistics for the parameters from Sect. 3.3 finally lead to a partitioning of all contact networks by an equivalence relation of having the same *effective Beta-splitting model for transmission trees*. Consider $\mathscr{C}_n^0$, the set of all initial SICNs, i.e., the set of all SI-tagged contact networks on n vertices with a single initial infector labelled without loss of generality

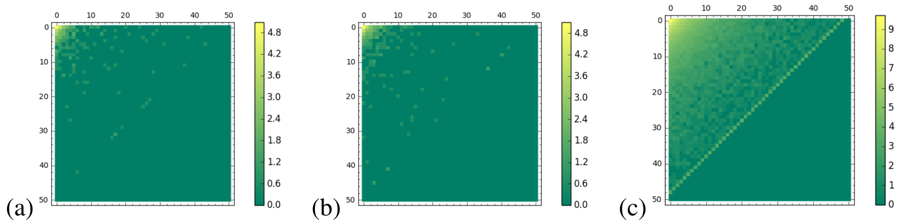

Figure 10: Sufficient statistic of split-pair frequencies $f(s^L, s^R)$ on $\mathscr{S}_{50}$ from transmission trees on the complete network over 50 vertices. Subplots (a) and (b) are for the two trials involving ten independent trees and subplot (c) is for a trial involving 1000 independent trees. The frequencies are displayed as $\log(1+f(s^L,s^R))$.

by ι_0 and from whom the infection can spread to the remaining $n-1$ individuals: $\iota_0, \iota_1, \ldots, \iota_{n-1}$. Thus, $\mathscr{C}_n^0$ is the set of all initial distributions starting from a single individual for our Markov chain in (2.11), i.e. with initial condition $(\tau(0), c(0))$, where $\tau(0)$ is the tree with ι_0 as its only vertex and $c(0)$ is the initial SICN. Note that $c(0)$, being an SICN, encodes that ι_0 is infected and all other vertices are uninfected initially. Let $\Pr(\tau; c(0))$ be the probability distribution on the space of transmission trees with n leaves (without branch-lengths) at the end of the transmission process (when all n individuals are infected) after starting from $c(0)$ according to (2.11). Finally let $\{\Pr((s^L,s^R); c(0)) : (s^L,s^R) \in \mathscr{S}_n\}$ be the probability distribution on the split-pairs in $\mathscr{S}_n$ that is further induced by $\Pr(\tau; c(0))$. Note that as the number of independent transmission trees r approaches infinity, the empirical relative frequency $\hat{P}((s^L,s^R); c(0)) := f(s^L,s^R)/(n-1)r$ that is obtained from r trees sampled from $\Pr(\tau; c(0))$, will converge with probability 1 by Borel's law of large numbers to the probabilities for each split-pair (s^L,s^R) in $\mathscr{S}_n$, as follows:

$$\lim_{r\to\infty} \hat{P}((s^L,s^R); c(0)) = \Pr((s^L,s^R); c(0)), \text{ for each } (s^L,s^R) \in \mathscr{S}_n \ .$$

Thus, the use of $\Pr((s^L,s^R); c(0))$ instead of $\hat{P}((s^L,s^R); c(0))$ in the per-vertex loglikelihood of Eq. (3.12) will produce an asymptotic or exact and possibly set-valued MLE in a deterministic manner without any standard error that is caused by finite r. We use this fact to prove Theorem 3.

Theorem 3. *Let the equivalence relation on $\mathscr{C}_n^0$ given by:*

$$c(0) \sim c'(0) \iff \left(\Pr((s^L, s^R); c(0)) = \Pr((s^L, s^R); c'(0)) \text{ for each } (s^L, s^R) \in \mathscr{S}_n\right) \ , \tag{3.16}$$

define the equivalence class $[c(0)] := \{c'(0) : c(0) \sim c'(0)\}$, such that the set of equivalence classes $\mathfrak{C}_n^0 := \{[c(0)]\}$ form a partition of $\mathscr{C}_n^0$ up to being identified by each distribution in

$$\mathfrak{S}_n^0 := \left\{\left\{\Pr((s^L, s^R); [c(0)]) : (s^L, s^R) \in \mathscr{S}_n\right\} : [c(0)] \in \mathfrak{C}_n^0\right\} \ , \tag{3.17}$$

the set of all distributions over the split-pairs in $\mathscr{S}_n$ that is generated by the transmission process unfolding on the product space of transmission trees and SICNs with any initial SICN in each $[c(0)] \in \mathfrak{C}_n^0$.

Then for each equivalence class $[c(0)]$ we can obtain its effective Beta-splitting model with parameters (α^*, β^*) given by the exact MLE involving $\Pr((s^L, s^R); [c(0)])$ *over each $(s^L, s^R) \in \mathscr{S}_n$ as specified by Eq.* (3.19)*. In other words, if two initial SICNs in $\mathscr{C}_n^0$ have the same distribution of split-pairs on $\mathscr{S}_n$ then they are indistinguishable by the exact MLE of the Beta-splitting model. We refer to the following map as the* Beta-projection *of the initial SICNs into the quarter-plane:*

$$\mathscr{B}^{\downarrow}(c(0)) = (\alpha^*, \beta^*) : \mathscr{C}_n^0 \to (-1, \infty)^2 \ , \tag{3.18}$$

and denote its inverse image by $\mathscr{B}^{\uparrow}(\alpha^*, \beta^*) : (-1, \infty)^2 \to \mathfrak{C}_n^0$.

Proof outline. The first part of the theorem is merely defining the equivalence class. The proof is a direct consequence of all initial SICNs in $[c(0)]$ being indistinguishable by the asymptotic maximum likelihood estimator that is confined to the information in $\{\Pr((s^L, s^R); [c(0)])\}$ under the Beta-splitting model:

$$(\alpha^*, \beta^*) := \underset{(\alpha,\beta)\in(-1,\infty)^2}{\arg\max} \prod_{(s^L,s^R)\in\mathscr{S}_n} \left(\frac{\prod_{j=0}^{s^R} \frac{\beta+j}{\beta+j+\alpha} \prod_{i=0}^{s^L} \frac{\alpha+i}{\alpha+i+\beta+s^R+1}}{\frac{\alpha\beta}{(\alpha+\beta)(\alpha+\beta+1)}} \right)^{\Pr((s^L,s^R);[c(0)])} . \tag{3.19}$$

We simply define the map and its inverse image in Eq. (3.18). □

Note that this equivalence relation is based on the rooted ranked planar binary topology of the tree and ignores other informative statistics of the transmission tree, including waiting times and individual labels.

Remark 4. *In practise, we will only be able to obtain the estimated effective Beta-splitting model associated with an initial SICN by finding $(\widehat{\alpha}, \widehat{\beta})$, the MLE from finitely many transmission trees drawn from the transmission process (i.e., with $r < \infty$ and positive standard error). As shown in Table 1 and Fig. 11, the MLEs can distinguish different underlying contact networks even when r is finite.*

3.4.1. Elements from the equivalence class of the source-initialized path network

To show that this equivalence relation is non-trivial, we give a concrete example of an equivalence class that not only contains the path network initialized at the source vertex (as in Sect. 3.2.3) but also n other initial SICNs. Recall that the only discrete transmission tree topology for the Beta-splitting tree with $(\alpha, \beta) \to (-1, \infty)$, is the right-branching comb, the same one obtained by assuming that the underlying SICN is the path network with the initial infection spreading from the individual, say ι_0, at the beginning of the path or the source vertex as in Sect. 2.1.3. To obtain other initial SICNs in the same equivalence class as the path network that is initially infected at the source vertex, let us consider the *unidirectional circular network* on n vertices given by $\{\iota_0, \iota_1, \dots, \iota_{n-1}\}$ and n directed edges given by $\{(\iota_0, \iota_1), (\iota_1, \iota_2), \dots, (\iota_{n-2}, \iota_{n-1}), (\iota_{n-1}, \iota_0)\}$. We can imagine the network being laid out on the plane along a circle. It is clear that we can have the infection initialized from any ι_i and it will spread sequentially along the circular path until all remaining individuals are infected in tandem, say anti-clockwise. Thus the transmission tree (ignoring leaf labels) generated on the unidirectional circular network by starting from any one of the n individuals is identical to the right-branching comb under the path network initialized at the source vertex. This simple construction gives us $n+1$ initial SICNs that belong to this equivalence class from two different underlying networks, namely circular and linear (unidirectional) path networks, but with the same relative split-pair frequencies given by the following uniform distribution on a side boundary of $\mathscr{S}_n$, provided $n \geq 2$:

$$\frac{f(s^L, s^R)}{(n-1)r} = \begin{cases} 1/(n-1) & \text{if } (s^L, s^R) \in \{(0, n-2), (0, n-3), \dots, (0,0)\} \subset \mathscr{S}_n \\ 0 & \text{otherwise} \ . \end{cases}$$

This is due to each of the r independent transmission trees being identically equal to the right-branching comb with split-pairs: $\{(0,n-2),(0,n-3),\ldots,(0,1),(0,0)\}$. Furthermore, in this simple example the probability of a split-pair is identical to its relative frequency for any $r \geq 1$ due to all probability being concentrated on one tree, i.e., $\Pr(s^L,s^R) = \hat{P}(s^L,s^R) := f(s^L,s^R)/(n-1)r = r/(n-1)r = 1/(n-1)$.

4. Classical families of contact networks and some inferential implications

We have already seen the values of α and β that prescribe the exact distribution over transmission trees when the SI-tagged contact network or SICN is the complete, path or star network. Here we explore other families of deterministic and random contact networks, primarily via simulations (using the generic code in Sects. Appendix A.1 and Appendix A.3), in order to obtain the sampling distributions they induce on the space of transmission trees. We further use samples from this distribution to obtain the maximum likelihood estimates (MLEs) for their corresponding effective Beta-splitting models (with code in Sect. Appendix A.2). These explorations are meant to strengthen one's intuition about the influence of various classical families of SICNs on the MLEs and their standard errors over $[-1,\infty)^2$, the shared parameter space of their effective Beta-splitting models. As discussed in Sect. 4.4, there are some natural inferential implications from these insights that can go well beyond the classical frequentist point estimation under the likelihood principle that is primarily pursued here.

More concretely, we simulate r transmission trees from various classical families of host contact networks using the Markov chain in Eq. (2.1) and tabulate the maximum likelihood estimates for α and β corresponding to their effective Beta-splitting models in the practical sense of Theorem 3 (with $r < \infty$) as per Remark 4. We study transmissions on three more deterministic and four random contact networks or parametric families of them. The mean and standard error (s.e.) of the MLEs $\hat{\alpha}$ and $\hat{\beta}$ based on transmission trees simulated from various contact networks over a few trials are tabulated in Table 1 with their IDs, and these IDs are depicted pictorially in Fig. 11. These simulation results, which we will try to make sense of below, do indicate that the MLEs can indeed distinguish different underlying contact networks to an extent even when r is finite. A more exhaustive study of other contact networks is possible by extending the generic code in Sects. Appendix A.1 and Appendix A.2 beyond the ten models studied here (as coded in Sect. Appendix A.3). We warn that the local optimization routines used are here are non-rigorous although we have been careful by using multiple initial conditions and choosing the numerically most stable expressions for the likelihood functions for each case. Ideally, the MLEs should be rigorously enclosed using interval arithmetic even through the natural interval extension [40] of the per-vertex likelihood function (which we have not done here).

4.1. Deterministic contact networks

We have already seen the complete, path and star networks as our guiding examples in Sect. 2.1 and their corresponding exact Beta-splitting models in Sect. 3.2. We also saw that the unidirectional circular network is in the same effective Beta-splitting equivalence class as the path network in Sect. 3.4.1. In this section we explore a few key families of deterministic contact networks to further extend our insights by interpolations of the ones already seen, when possible.

4.1.1. Bidirectional circular path network

Let us extend the unidirectional circular network of Sect. 3.4.1 to a bidirectional circular network by making each edge bidirected (or undirected). Thus $\{\imath_0, \imath_1, \ldots, \imath_{n-1}\}$ is the vertex set and $\{(\imath_0,\imath_1),(\imath_1,\imath_0),(\imath_1,\imath_2),(\imath_2,\imath_1),\ldots,(\imath_{n-1},\imath_0),(\imath_0,\imath_{n-1})\}$ is the edge set for this bidirectional circular network. With bidirectionality, we can have randomness in the transmission trees unlike the deterministic right-branching comb for the unidirectional case. This is because the next infection event can be either in the left or the right subtree of the root vertex encoding the first infection event. The initial infector could be any one of the vertices due to circular symmetry of the network and we take this to be $\imath_0$ without loss of generality. Thus the probability that the next infection is in the right or the left subtree of the root is equally likely, provided the current number of leaves (number of individuals infected) is less than n and each edge-weight in this contact network is 1. Finally each of these subtrees will be deterministically right-branching combs due to the fact that there is only one infected individual in each subtree that is

Table 1: The mean and standard error (s.e.) of the MLEs $\overline{\hat{\alpha}}$ and $\overline{\hat{\beta}}$ based on transmission trees simulated from various contact networks in replicated trials. Here, s.e. is the sample standard deviation over the trials.

ID	Contact network	n	r	trials	$\overline{\hat{\alpha}}$ (s.e.)	$\overline{\hat{\beta}}$ (s.e.)
1	Complete	1,000	1	5	-0.006952 (0.06853)	0.05208 (0.1005)
2	Star	1,000	1	5	∞ (0.0000)	-1.0000 (0.0000)
3	Path	1,000	1	5	-1.0000 (0.0000)	∞ (0.0000)
4	Bidirectional Circular	50	1	5	-0.9880 (0.0006)	1.4584 (0.1534)
5	Bidirectional Circular	50	100	5	-0.9879 (0.0000)	1.5189 (0.0067)
6	$\mathrm{BalancedTree}(2,9)$	1023	1	5	-0.4052 (0.0000)	-0.1477 (0.0000)
7	$\mathrm{BalancedTree}(3,6)$	1093	1	5	-0.06452 (0.0000)	-0.5215 (0.0000)
8	$\mathrm{BalancedTree}(4,5)$	1365	1	5	0.06556 (0.0000)	-0.7109 (0.0000)
9	$\mathrm{BalancedTree}(6,4)$	1555	1	5	0.2350 (0.0000)	-0.8510 (0.0000)
10	$\mathrm{BalancedTree}(10,3)$	1111	1	5	0.9249 (0.0000)	-0.9156 (0.0000)
11	$\mathrm{BalancedTree}(32,2)$	1057	1	5	1.1624 (0.0000)	-0.9853 (0.0000)
12	$\mathrm{BalancedTree}(999,1)$	1000	1	5	∞ (0.0000)	-1.0000 (0.0000)
13	2D toroidal grid	1024	1	5	-0.8612 (0.008425)	-0.5606 (0.03219)
14	2D toroidal grid	10000	1	5	-0.89346 (0.0022)	-0.6626 (0.0106)
15	3D toroidal grid	1000	1	5	-0.6849 (0.01479)	-0.3515 (0.03451)
16	3D toroidal grid	10648	1	5	-0.7628 (0.007956)	-0.4968 (0.01641)
17	$\mathrm{ER}(100,0.030)$	100	30	5	-0.6063 (0.01383)	-0.4052 (0.02710)
18	$\mathrm{ER}(100,0.040)$	100	30	5	-0.5179 (0.01855)	-0.3151 (0.02244)
19	$\mathrm{ER}(100,0.050)$	100	30	5	-0.4059 (0.02020)	-0.2246 (0.01952)
20	$\mathrm{ER}(100,0.10)$	100	30	5	-0.1997 (0.03106)	-0.1280 (0.03063)
21	$\mathrm{ER}(100,0.20)$	100	30	5	-0.1074 (0.03961)	-0.06166 (0.03020)
22	$\mathrm{ER}(100,0.40)$	100	30	5	0.02247 (0.06603)	0.01541 (0.05499)
23	$\mathrm{ER}(100,0.64)$	100	30	5	-0.01097 (0.03984)	0.01046 (0.05112)
24	$\mathrm{ER}(100,1.0)$	100	30	5	-0.001787 (0.04347)	-0.01555 (0.04019)
25	$\mathrm{RandReg}(1000,3)$	1000	1	5	-0.7504 (0.004186)	-0.06260 (0.06322)
26	$\mathrm{RandReg}(1000,4)$	1000	1	5	-0.5530 (0.04513)	-0.002305 (0.09785)
27	$\mathrm{RandReg}(1000,6)$	1000	1	5	-0.3520 (0.03464)	0.06042 (0.06586)
28	$\mathrm{RandReg}(1000,10)$	1000	1	5	-0.1939 (0.06167)	0.07274 (0.1238)
29	$\mathrm{RandReg}(1000,100)$	1000	1	5	0.06378 (0.04519)	0.1084 (0.05844)
30	$\mathrm{RandReg}(1000,999)$	1000	1	5	-0.01496 (0.08893)	0.006464 (0.04166)
31	$\mathrm{SWRN}^{*,\circ}(50,2,0.0)$	50	30	5	-0.9878 (0.0001516)	1.514 (0.01222)
32	$\mathrm{SWRN}^{*}(50,2,0.1)$	50	30	5	-0.9618 (0.003047)	-0.4147 (0.03203)
33	$\mathrm{SWRN}^{\circ}(50,2,0.1)$	50	30	5	-0.9652 (0.002863)	-0.3828 (0.1171)
34	$\mathrm{SWRN}^{*}(50,2,0.2)$	50	30	5	-0.9375 (0.004620)	-0.5683 (0.0193)
35	$\mathrm{SWRN}^{*}(50,2,0.5)$	50	30	5	-0.8632 (0.008181)	-0.6471 (0.03722)
36	$\mathrm{SWRN}^{*}(50,5,0.1)$	50	30	5	-0.7530 (0.01572)	-0.4751 (0.04671)
37	$\mathrm{SWRN}^{\circ}(50,5,0.1)$	50	30	5	-0.7918 (0.01596)	-0.5130 (0.03323)
38	$\mathrm{SWRN}^{\circ}(50,5,0.2)$	50	30	5	-0.6881 (0.03277)	-0.3595 (0.06002)
39	$\mathrm{SWRN}^{\circ}(50,5,0.5)$	50	30	5	-0.5264 (0.04687)	-0.2138 (0.09471)
40	$\mathrm{SWRN}^{\circ}(100,2,0.2)$	100	1	5	-0.9479 (0.01509)	-0.3991 (0.5065)
41	$\mathrm{SWRN}^{\circ}(100,2,0.2)$	100	30	5	-0.9493 (0.003869)	-0.6027 (0.03475)
42	$\mathrm{SWRN}^{*}(100,2,0.5)$	100	1	5	-0.9023 (0.03411)	-0.7139 (0.03929)
43	$\mathrm{SWRN}^{*}(100,2,0.5)$	100	30	5	-0.8878 (0.006687)	-0.6821 (0.02189)
44	$\mathrm{SWRN}^{\circ}(100,2,0.5)$	100	1	5	-0.8714 (0.05584)	-0.6533 (0.09257)
45	$\mathrm{SWRN}^{\circ}(100,2,0.5)$	100	30	5	-0.8920 (0.005128)	-0.6786 (0.02189)
46	$\mathrm{SWRN}^{\circ}(100,5,0.99)$	100	30	5	-0.5079 (0.02371)	-0.2290 (0.03059)
47	$\mathrm{SWRN}^{\circ}(100,10,0.99)$	100	30	5	-0.2027 (0.07641)	-0.05611 (0.06949)
48	$\mathrm{PrefAttach}^{*}(100,1)$	100	30	10	-0.3275 (0.04932)	-0.8215 (0.01121)
49	$\mathrm{PrefAttach}^{*}(100,2)$	100	30	10	-0.2443 (0.03283)	-0.6647 (0.01294)
50	$\mathrm{PrefAttach}^{\circ}(100,1)$	100	30	10	-0.3813 (0.04908)	-0.8254 (0.005460)
51	$\mathrm{PrefAttach}^{\circ}(100,2)$	100	30	10	-0.3339 (0.03884)	-0.6743 (0.01657)
52	$\mathrm{PrefAttach}^{\circ}(100,3)$	100	30	10	-0.2545 (0.04181)	-0.5863 (0.01652)
53	$\mathrm{PrefAttach}^{\circ}(100,5)$	100	30	10	-0.1748 (0.04214)	-0.4698 (0.03110)
54	$\mathrm{PrefAttach}^{\circ}(100,10)$	100	30	10	-0.1196 (0.03449)	-0.3089 (0.02663)
55	$\mathrm{PrefAttach}^{\circ}(100,1)$	100	1	5	-0.2472 (0.2698)	-0.7993 (0.05843)

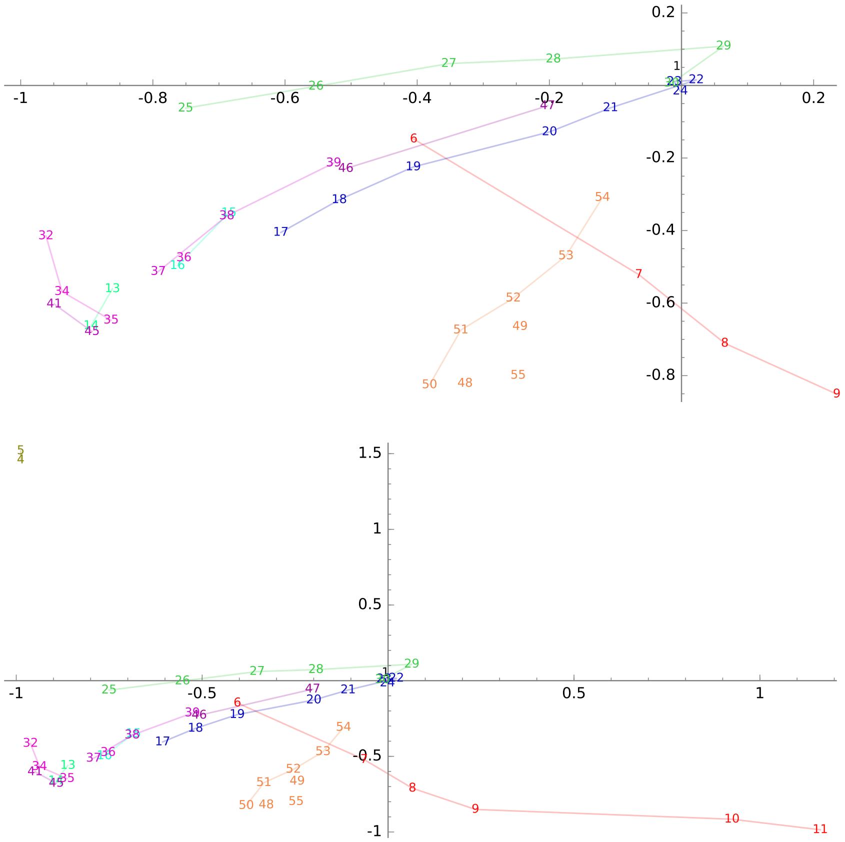

Figure 11: A pictorial depiction of the mean MLEs $\overline{\hat{\alpha}}$ along x-axis and $\overline{\hat{\beta}}$ along y-axis based on transmission trees simulated from various contact networks indexed by their ID from Table 1. Different IDs from the same parametric family are shown in the same color with lines connecting them if there is a natural parametric interpretation within the family (see description in text). The top figure is a zoom-in of the bottom one.

capable of infecting its only uninfected out-neighbor, if any. We call such trees with *o*nly *r*ight-*b*ranching *s*ubtrees *o*f the *r*oot vertex as *orbsor* trees. Two such orbsor trees generated from the bidirectional contact network with $n = 6$ and initialized from ι_0 are shown in Fig. 12.

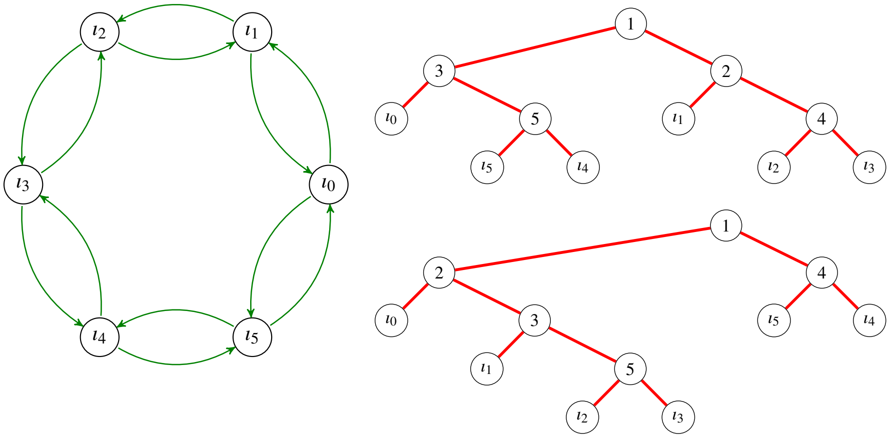

Figure 12: The bidirectional circular contact network (left) and two sampled transmission trees from it (right).

Putting all these facts together we can directly see that the probability distribution on transmission trees is uniformly distributed over the 2^{n-1} orbsor trees that form as subset of the $n!(n-1)!$ many possible transmission trees, as follows:

$$\Pr(\tau) = \begin{cases} \frac{1}{2^{n-1}} & \text{if } \tau \text{ is an orbsor tree with } n \text{ leaves,} \\ 0 & \text{otherwise.} \end{cases} \tag{4.1}$$

For comparison with our three core examples, recall that the distribution on transmission trees induced by the complete, star and path networks are also uniformly distributed on the following subsets of the set of transmission trees: the entire set, the set of $(n-1)!$ left-branching comb trees and the singleton set of the right-branching comb tree, respectively. Thus, our first four deterministic contact networks induce uniform distributions on various subsets of the set of all transmission trees.

As shown in Table 1, $(\overline{\widehat{\alpha}}, \overline{\widehat{\beta}})$, the mean MLE of the effective Beta-splitting model based on $r = 1$ or $r = 100$ transmission trees drawn from such a network over $n = 50$

individuals is close to $(-0.98, 1.5)$. They are depicted by IDs 4 and 5 on the top left corner of the zoomed-out image at the bottom of Fig. 11. Although some combinatorial book-keeping may allow one to analytically pursue the transformation of the distribution in Eq. (4.1) to $\mathscr{S}_n$ (see last paragraph of Sect. 4.4.1 and Sect. 5.3), we begin to content ourselves with mere simulation results in preparation for more complicated networks that are not easily amenable to extractions of exact expressions. The mean MLE of $(-0.98, 1.5)$ for this bidirectional path network, that is binomially composing the right-branching comb of the path network on either side of its root vertex, makes intuitive sense because it is not as extreme as $(-1, \infty)$, the MLE of the path network at the boundary of the parameter space in its limit (with ID 3 that can be imagined in Fig. 11).

4.1.2. Balanced tree network

$\text{BalancedTree}(d,h)$ is the perfectly balanced tree of height $h \geq 1$ and whose root has degree $d \geq 2$. The number of vertices in this network is $n = 1 + d + d^2 + \cdots + d^h = (d^{h+1} - 1)/(d-1)$ and the number of edges is $n-1$. Balanced tree networks can be thought of as a biparametric extension of the star network which is equivalent to $\text{BalancedTree}(n-1, 1)$.

The transmission tree generated on such perfectly balanced tree contact networks is unique if we ignore vertex labels and branch-lengths. We refer to them as left-branching d-sharks in the visual spirit of left-branching combs for star networks. Instead of giving a recursive formula for these trees we illustrate them by examples for $\text{BalancedTree}(3,2)$ and $\text{BalancedTree}(2,3)$ in Fig. 13. Thus, the $\text{BalancedTree}(d,h)$ contact network produces a distribution on transmission trees that is concentrated on a single left-branching d-shark tree (ignoring all vertex labels and branch-lengths) and when $d = n-1$ and $h = 1$ the left-branching $(n-1)$-shark tree is the left-branching comb tree for the star contact network. The mean MLE of the effective Beta-splitting models are shown in Table 1 for various values of d and h. As d and h approach $n-1$ and 1, respectively, while keeping the population size n as close to 1000 as possible, the corresponding mean MLEs are approaching $(\infty, -1)$, the limiting MLE of the star network with ID 2 and the $\text{BalancedTree}(999, 1)$ with ID 12 as expected. This tendency

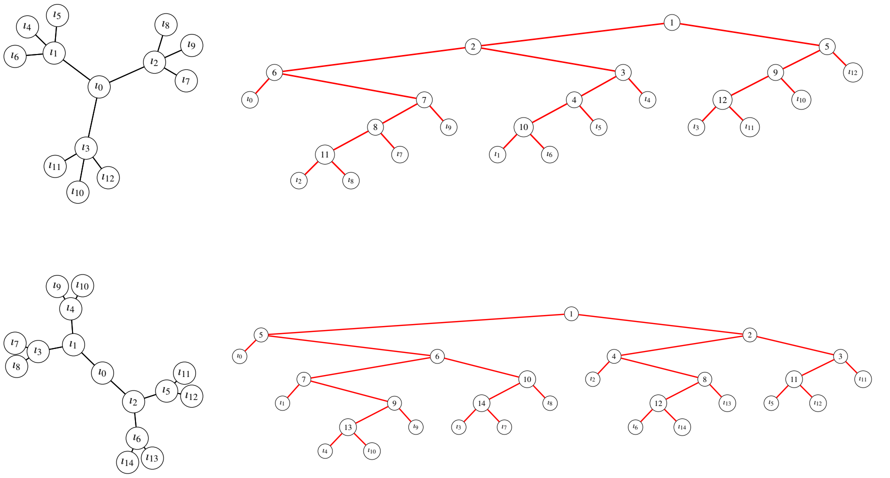

Figure 13: The contact networks $\text{BalancedTree}(3,2)$ (top left) and $\text{BalancedTree}(2,3)$ (bottom left) and their corresponding transmission trees initialized from ι_0 are given by the left-branching 3-shark (top right) and 2-shark (bottom right) trees, respectively

toward $(\infty, -1)$ is depicted by the sequence of IDs 6–11 in Fig. 11. The standard errors are zero due to the uniqueness of the transmission tree (at the sufficient but not minimally sufficient resolution of rooted, unranked, planar and leaf-unlabelled tree) that is realized under each $\text{BalancedTree}(d,h)$ contact network.

4.1.3. Toroidal regular grid network

We identify the vertices along the two pairs of opposite edges and the three pairs of opposite faces in the regular finite 2-dimensional (2D) square grid with $\sqrt{n} \times \sqrt{n} = n$ individual vertices and the 3-dimensional (3D) cube grid with $\sqrt[3]{n} \times \sqrt[3]{n} \times \sqrt[3]{n} = n$ individual vertices, respectively. A 2D toroidal grid with $n = 9$ is shown in Fig. 14. The Figure also shows the sampling variation in three independent transmission trees grown on the network with initial infection at ι_0.

Due to the toroidal structure, the transmission tree distributions are invariant to the initial infection (ignoring leaf labels). The mean MLEs of the effective Beta-splitting model corresponding to 2D and 3D toroidal grids with n around 10^3 and 10^4 are depicted in Table 1. The mean MLEs based on one transmission tree seem to be fairly concentrated about $(-0.9, -0.66)$ and $(-0.76, -0.5)$ for contact networks on toroidal

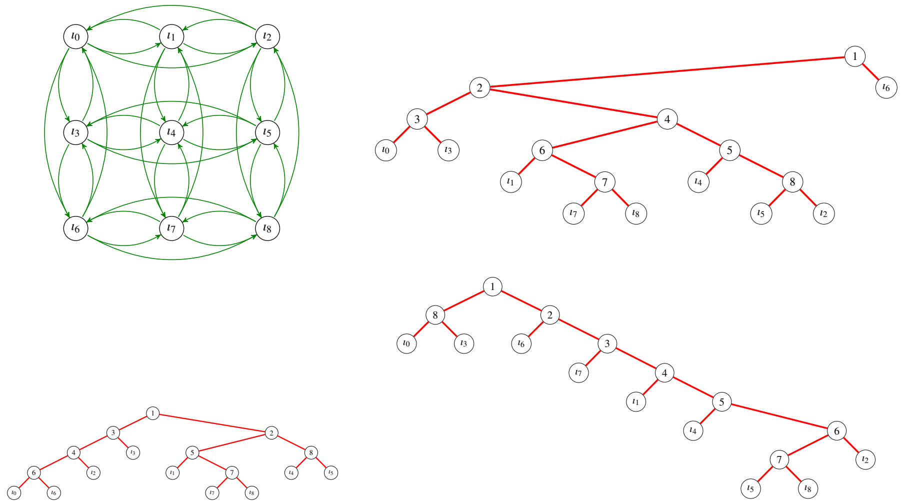

Figure 14: Toroidal 2D grid contact network with $n = 9$ (top left) and three sample transmission trees from it with initial infection at ι_0.

2D and 3D grids with about $n = 10^4$ individuals, respectively. There is also an effect towards $(-1,-1)$ for the 2D and 3D cases as n increases from about 10^3 to 10^4 as depicted by the ID pairs (13, 14) and (15,16) in Fig. 11, respectively.

4.2. *Random contact networks*

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 network model [3, 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 network 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 vertices. We explore some basic random graph models to gain insights into the distribution of transmission trees and the MLEs of their effective Beta-splitting models.

Remark 5. *Let us note that one may also study the distribution of transmission trees for a specific realization of a given random contact network or its partially observed sub-network. Such subtle distinctions, which in turn will depend on the exact decision problem and the available data at hand, can be pursued by modifying our basic code in Appendix Appendix A. But here we limit ourselves to the random sense involving multiple independent trials such that each trial is a realization of a full transmission tree with n leaves on each realization of a given random contact network with an initially infected node (i.e., a static initial SICN).*

4.2.1. Erdős-Rényi random network

The Erdős-Rényi random network denoted by $\mathrm{ER}(n, p)$ on n vertices is obtained by inserting each of the $n(n-1)/2$ undirected edges independently with probability p [41, 42]. We interpret the undirected edges as being bidirected to obtain our random contact network $\mathrm{ER}(n, p)$ to study transmission trees evolving on the connected component of $\mathrm{ER}(n, p)$ containing the initially infected individual ι_0. Thus $\mathrm{ER}(n, p)$ becomes more dense and approaches the complete network as the edge probability $p \to 1$ or equivalently as the average vertex degree $\lambda := np \to n$. We observed more than 90 vertices on average in the connected component containing ι_0 if $p > 0.03$ when $n = 100$. This is sensible because $\mathrm{ER}(n, p)$ is known theoretically to have a unique giant component containing a positive fraction of the vertices almost surely if $\lambda > 1$. We are primarily interested in the regime where $p > \log(n)/n = \log(100)/100 \approxeq 0.0461$ (in Table 1) or $\lambda > \log(100) \approxeq 4.61$ (in Fig. 15) when $\mathrm{ER}(n, p)$ is known theoretically to be connected almost surely so that all n individuals can be eventually infected from the initial infection at ι_0.

Maximum likelihood estimates $\widehat{\alpha}$ and $\widehat{\beta}$ of the effective Beta-splitting model based on $r = 30$ independent transmission trees that were grown on independent realizations of $\mathrm{ER}(100, \lambda/100)$ and replicated in five independent trials are shown in Fig. 15 as a function of λ. The mean and standard errors of the MLEs are also given in Table 1. Note that the standard errors are higher when compared to those of the deterministic contact networks if $r = 1$ (results not shown). This is naturally due to the additional randomness introduced by distinct realizations of the $\mathrm{ER}(n, p)$ random contact network across the trials. The standard errors in Table 1 have been reduced by increasing r to 30. As expected, the MLEs plotted as points in Fig. 15 and the mean MLEs in Table 1 and their corresponding IDs 17–24 in Fig. 11 approach the origin as λ approaches 100 and the contact network $\mathrm{ER}(100, 1)$ becomes the complete network.

Interestingly ID 6 of the deterministic $\mathrm{BalancedTree}(2, 9)$ has its MLE fairly close to the mean MLE of $\mathrm{ER}(100, 0.050)$ with ID 19. When h is increased by 1 from 6 (with $n = 127$) to 9 (with $n = 1023$), the mean MLE of $\mathrm{BalancedTree}(2, 9)$ starts from $(-0.3259, -0.05752)$ and reaches $(-0.4052, -0.1477)$ piece-wise linearly with different slopes slightly over 1. These results are in sync with a standard probability trick of approximating a sparse connected ER using trees.

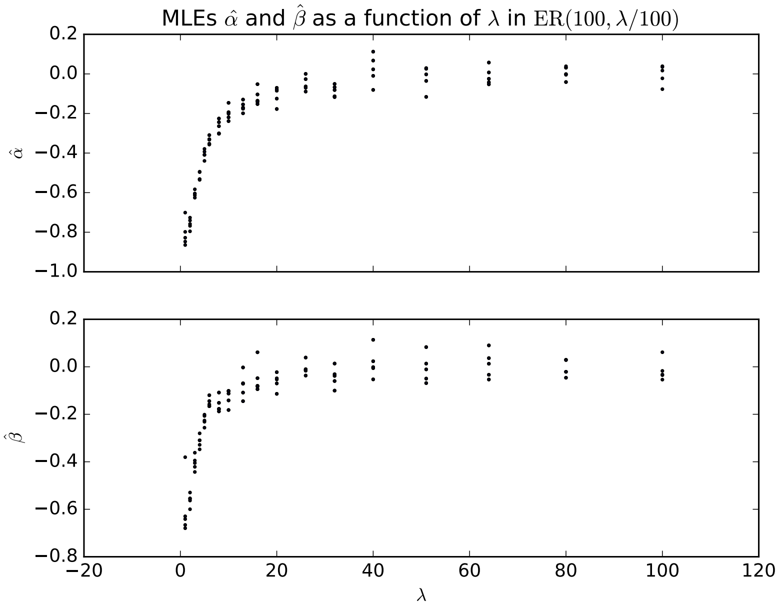

Figure 15: The Maximum likelihood estimates $\hat{\alpha}$ and $\hat{\beta}$ as a function of average vertex degree λ in the Erdős-Rényi random network $\mathrm{ER}(n, p)$ on $n = 100$ vertices with edge probability $p = \lambda/n$. The estimates are based on $r = 30$ independent transmission trees grown on independent realizations of $\mathrm{ER}(100, \lambda/100)$ in five replicate trials.

4.2.2. Random regular network

Each vertex in a *d-regular graph* has the same degree d or number of neighbors. A d-regular directed graph or network must also satisfy the stronger condition that the indegree and outdegree of each vertex are equal to each other. A random d-regular graph is a graph drawn uniformly at random from the set of all d-regular graphs on n vertices, where $3 \leq d < n$ and nd is even [43]. We merely interpret the undirected edges as being bidirected to obtain $\mathrm{RandReg}(n,d)$, the random d-regular network on n vertices. Note that $\mathrm{RandReg}(n,d)$ approaches the complete network on n vertices as d approaches $n-1$. This is evident in the behavior of the mean MLEs obtained from $r = 1$ transmission tree grown on the $\mathrm{RandReg}(n,d)$ contact network on $n = 1000$ vertices for values of d in $\{3,4,6,10,100,999\}$ as shown in Table 1 and by their IDs 25–30 approaching ID 0 of the complete network with the same n and r at the origin in Fig. 11. Note that Model ID 28 with $d = 10$ and $n = 1000$ nearly satisfies the condition that $d = O(n^{1/3-\varepsilon})$ for the underlying randomized algorithm [44] to generate an asymptotically uniform random d-regular graph on n vertices [45].

4.2.3. Connected small-world random network

The small-world random graph model of [46] on n vertices is constructed by first creating a ring over n vertices (undirected circular path graph). Then each vertex in the ring is connected with its m nearest neighbors if m is even (and its $m-1$ nearest neighbors if m is odd). Finally edges are rewired as follows: for each edge (u,v) in the underlying *n-ring with m nearest neighbors* graph, with probability p rewire (u,v) as the new edge (u,w) with randomly-chosen existing vertex w. The undirected edges are interpreted as bidirected and we repeatedly sample until we obtain a *connected small world random network* $\mathrm{SWRN}(n,m,q)$ that is specified by the three parameters: n for the number of vertices, m for the number of nearest neighbors each vertex is connected to and q for the probability of rewiring each edge. Thus, $\mathrm{SWRN}(n,m,q)$ accounts for clustering and at least partially explains the “small-world” phenomena that is observed in a variety of real-world networks while retaining the short average path lengths of the

Erdős-Rényi random graph $\mathrm{ER}(n,p)$.

We consider two possible initializations, i.e. two different initial SICNs for a given realization of $\mathrm{SWRN}(n,m,q)$. In the first case, denoted by $\mathrm{SWRN}^*(n,m,q)$, we grow the transmission tree from the vertex with the largest out-degree. If more than one vertex has the maximal out-degree then we choose the vertex with the smallest label. In the second case, denoted by $\mathrm{SWRN}^\circ(n,m,q)$, we grow the transmission tree from a randomly chosen vertex. The maximum likelihood estimates for their effective Beta-splitting models are obtained from r independent transmission trees grown on these random networks with the two initializations having different parameters as shown in Table 1 with their IDs. The two initializations basically coincide when $q = 0$.

We mainly explore the case with m equalling 2 and 5 with 1 and 2 neighbors, respectively, on either side of each vertex initially, for a few values of the rewiring probability $q \in \{0.2, 0.5, 0.99\}$ with $n \in \{50, 100\}$ and $r \in \{1, 30\}$ as shown in Table 1 and depicted with their IDs 31–47 in Fig. 11. More interpretable IDs, in the sense of having variation in only one parameter in the family, are shown by the same color with lines connecting them.

$\mathrm{SWRN}(n,m,q)$ interpolates between the *n-ring with m nearest neighbors* network and $\mathrm{ER}(n,p)$, such that, as $q \to 1$, $\mathrm{SWRN}(n,m,q) \to \mathrm{ER}(n, nm/(n(n-1)))$. When q is close to 1, we do find that the MLEs of $\mathrm{SWRN}^\circ(100,5,0.99)$ with ID 46, with initial infection chosen uniformly at random just as in the $\mathrm{ER}(n,p)$ contact network, are roughly closer to those for $\mathrm{ER}(100, 0.050)$ with ID 19 since $nm/(n(n-1)) \approxeq 0.05$ when compared with $\mathrm{SWRN}^\circ(100,5,q)$ with smaller values of q (results not shown).

An interesting observation is the proximity of the mean MLEs for certain 2D and 3D toroidal grids to certain SWRNs: (i) ID 15 of the 3D toroidal grid with $n = 1000$ vertices (each connected to its six nearest neighbors) and ID 38 of $\mathrm{SWRN}^\circ(50,5,0.2)$ (each of its 50 vertices initially connected to its four nearest neighbors before being rewired with probability 0.2), (ii) ID 16 of the 3D toroidal grid with $n = 10648$ and IDs 36 and 37 corresponding to $\mathrm{SWRN}^*(50,5,0.1)$ and $\mathrm{SWRN}^\circ(50,5,0.1)$, respectively, and (iii) ID 14 of the 2D toroidal grid with $n = 10^4$ and $\mathrm{SWRN}^\circ(100,2,0.5)$ with ID 45. More extensive simulations are necessary to systematically understand these proximities (using rigorous global interval optimization techniques, say in [40], for the MLEs as opposed to the local optimizations used here). Insights from such rigorous simulations may lead to further analysis towards understanding the nature of such proximities between these distinct model families under the Beta-projections of Eq. (3.18), especially for different values of n. Other insights from Fig. 11 include the effect of changing n and the initialization strategy.

4.2.4. Preferential attachment random network

Next we explore the random network created using the preferential attachment model of [47]. Real-world networks are best described by a scale-free power-law distribution for their degrees. The preferential attachment model, unlike the other random graph models here, produces such a power-law degree distribution through two generic mechanisms: (i) networks are grown by the addition of new vertices, and (ii) new vertices attach preferentially to exiting vertices that are already well connected (i.e. with a high degree). The randomized algorithm for the construction of the preferential attachment network $\text{PrefAttach}(n,m)$ is as follows. First, a graph with m vertices and no edges is initialized, and a graph of n vertices is grown by attaching new vertices, each with m edges, to existing vertices independently according to probability proportional to their degrees. This results in a preferential attachment behavior whereby new vertices are attached to existing vertices that already have a high degree. We interpret the undirected edges as bidirected to obtain a network.

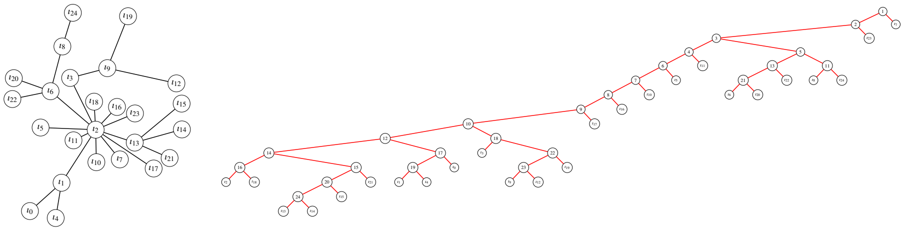

Figure 16: A realization of the $\text{PrefAttach}^*(n,m)$ contact network with $n = 25$ and $m = 1$ (left) and a realization of the transmission tree grown on it (right) with the infection spreading from the maximal degree vertex ι_2.

We also use two different initializations to grow a random transmission tree on a realization of this random contact network. In the $\text{PrefAttach}^\circ(n,m)$ random SICN we initialize the infection uniformly at random from one of the initial m individuals with no edges at the beginning of its construction. For the $\mathrm{PrefAttach}^*(n,m)$ contact network we find the smallest vertex label with the largest out-degree, i.e. the most preferentially attached vertex (which may not necessarily be the first vertex to be attached), as the initial infected individual. We make the distinction because transmission trees can only be grown on $\mathrm{PrefAttach}^*(n,m)$ after the preferential attachment model has completed its construction of the network with all n vertices. In contrast, for $\mathrm{PrefAttach}^\circ(n,m)$ we can grow the transmission tree even while the preferential attachment model is constructing the underlying contact network. The mean MLEs of the effective Beta-splitting model corresponding to $r = 30$ independent transmission trees grown over independent realizations of these two variants of the preferential attachment contact networks is shown in Table 1. The bursty or starry nature of the hubs or popular vertices is evident in the left-branching tendency of the transmission trees grown on such preferential attachment networks as shown in Fig. 16. The IDs 50–54 correspond to $\mathrm{PrefAttach}(100,m)$ where m takes the values 1, 2, 3, 5 and 10, respectively, show the mean MLEs approach toward the origin where the mean MLE of the complete SICN occurs in Fig. 11 – it cannot reach the origin since $m << n$ for $\mathrm{PrefAttach}(n,m)$. This is sensible since more of the m vertices will get attached to in the initial steps of the algorithm that is constructing the preferential attachment network and thereby reduce the preferential attachment tendency for larger values of m.

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

In Sect. 3.2 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 exactly 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 $(-1,\infty)$. Using the inferential procedure of Section 3.3 we can 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. Next we present a family of SICNs that interpolate our three SICNs: one at the origin and two at extremes of our parameter space $(-1,\infty)^2 \ni (\alpha,\beta)$.

A circulant network or digraph on n vertices labelled by $\mathbb{V} = \{0,1,\dots,n-1\}$ 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 $\mathsf{C}(n,\mathbb{A})$. First note that $\mathsf{C}(n,\{1,2,\dots,n-1\})$ is the complete network k_n and $\mathsf{C}(n,\{d\})$ has constant degree sequence with degree d since each vertex 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 $\mathsf{C}(n,\{1\})$ when the infection starts at vertex 0 (at the source vertex of p_n) since the extra directed edge $(n-1,0)$ in $\mathsf{C}(n,\{1\})$ 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 $\mathbb{A}_i = \{1,2,\dots,i\}$ we get the sequence of circulant graphs to interpolate from the path network to the complete network:

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

This sequence is shown for $n=5$ in the bottom row of Fig. 17 (going from right to left). To achieve an interpolating sequence from the star network to the complete graph we note that $\mathsf{C}(n,\emptyset)$ has no edges. By letting $\mathbb{A}_0 = \emptyset$, we can obtain the desired sequence by simply adding the edges of the star network, $\{(0,i) : i \in \{1,2,\dots,n\}$, to the edge set of each $\mathsf{C}(n,\mathbb{A}_i)$ in $(\mathsf{C}(n,\mathbb{A}_i))_{i=0}^{n-2}$, as shown in the top row of Fig. 17 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. Appendix A.4.

We can finally see in Fig. 18 how the probability density function (PDF) of the MLEs, $(\widehat{\alpha},\widehat{\beta})$, change as we sequentially vary the SICN in the family that interpolates from 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 simu-

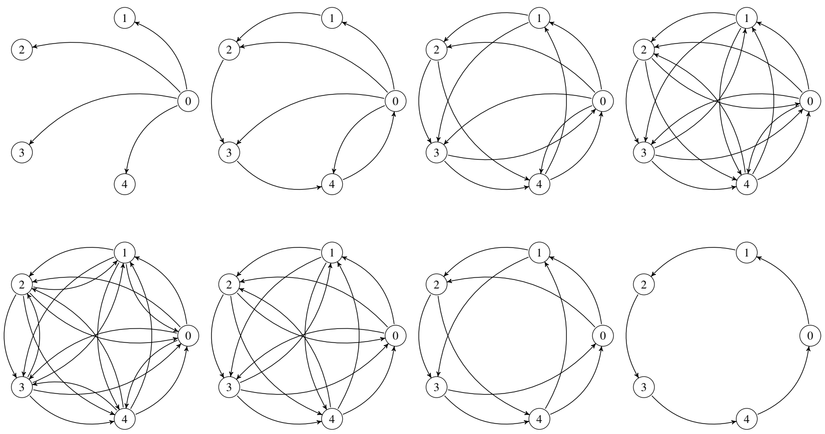

Figure 17: A path from star network to circular path network through the complete network with 5 vertices (in z-pattern layout).

lated from each SICN in the sequence of 98 SICNs over a population of size $n = 50$. In Fig. 18, 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 $\mathsf{C}(n, \mathbb{A}_i)$. 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.

The sufficient statistics of split-pair frequencies for α and β from various stages of the interpolating family of SICNs spanning the star, complete and path network are shown in Fig. 19. Note how these sufficient statistics are also changing gradually as

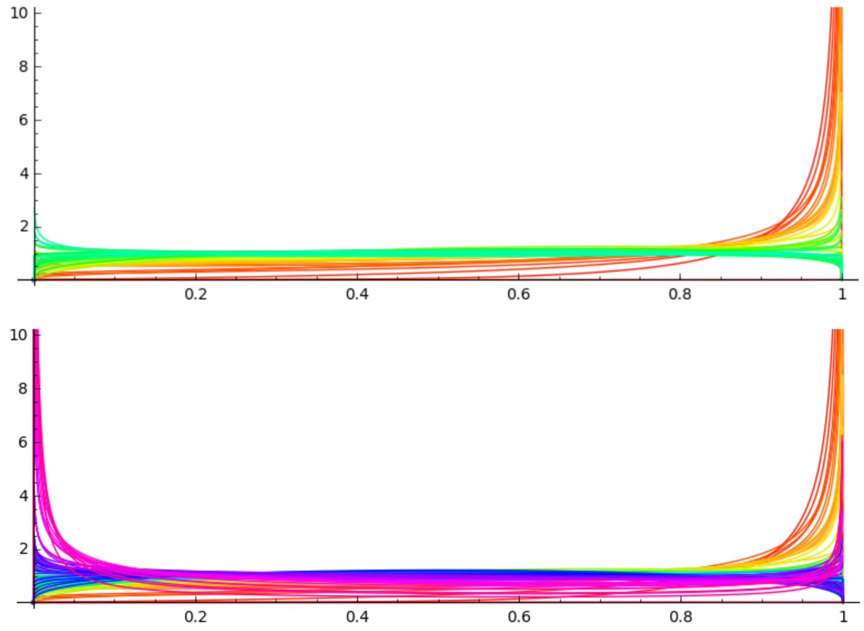

Figure 18: Probability density function (PDF) of the $\mathscr{B}(\alpha+1,\beta+1)$ distribution at the maximum likelihood estimates of α and β based on 10 sampled transmission trees from each SICN in the sequential family that interpolates from the star network (red hue) to the circular path network (pink hue) via the complete network (blue hue) with $n = 50$ vertices. 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.

expected.

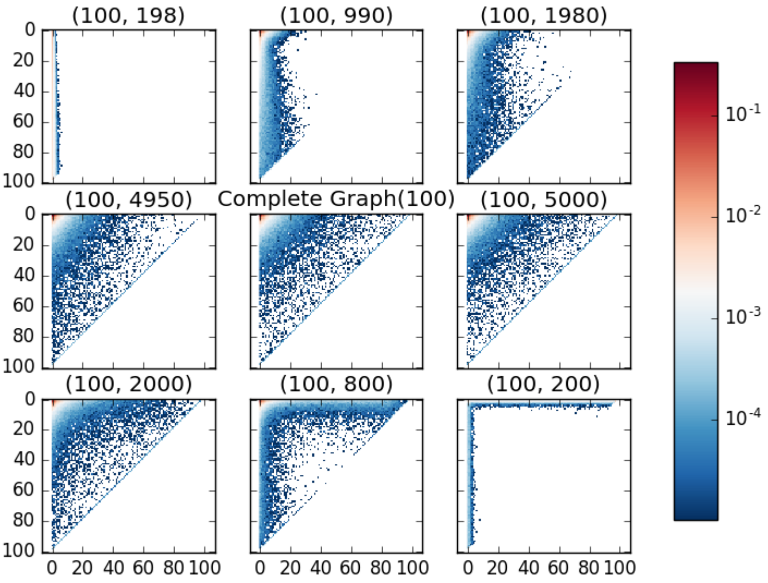

Figure 19: The sufficient statistics of split-pair frequencies for α and β shown as the empirical mass function from 1000 independent transmission trees over a population of size $n = 100$ at various stages of the interpolating family spanning the star, complete (in the center of the panel with 100 vertices and 9900 edges) and path networks (the sub-plots are labelled by the number of vertices and edges in the host contact network).

4.4. Implications for statistical inference

We have deliberately confined the inferential aspects of this study to the classical maximum likelihood estimator in the simplest sampling setting involving fully grown transmission trees with n leaves over r independent trials. This merely amounts to inferring the underlying contact network via multiple random breadth-first expansions that are encoded through rooted, ranked, planar, labelled, and binary (spanning) trees, i.e. our transmission trees. We now give some basic insights into two other natural approaches to inference for applied network scientists and also provide some natural extensions of the likelihood function.

4.4.1. An L_1 perspective

We could have also defined the effective Beta-splitting model by minimizing the total variation distance between $\Pr((s^L, s^R); [c(0)])$, the probabilities on $\mathscr{S}_n$ generated by the transmission process unfolding on the product space of transmission trees and SICNs with initial SICNs in $[c(0)]$, and $\Pr((s^L, s^R); \alpha, \beta)$, the probabilities generated by the Beta-splitting model as follows:

$$(\alpha^*, \beta^*) := \underset{(\alpha,\beta)\in(-1,\infty)^2}{\arg\min} \quad \frac{1}{2} \sum_{(s^L,s^R)\in\mathscr{S}_n} \left| \Pr((s^L, s^R); [c(0)]) - \Pr((s^L, s^R); \alpha, \beta) \right| \ . \tag{4.2}$$

One elegant aspect of this L_1-minimizing rule, as opposed to the likelihood maximizing rule behind MLE, is its desirable non-parametric decision-theoretic properties, such as being a metric on all probability distributions over $\mathscr{S}_n$, having easily interpretable distances in the universal scale of $[0,1]$, and having well-known relations to other statistical notions of divergence. For these reasons we believe that the L_1-minimizing rule, if efficiently implementable, would lead to better statistical properties for the task of characterizing the equivalence class based on empirical $\hat{P}$ of relative frequencies on $\mathscr{S}_n$ as opposed to the exact $\Pr((s^L, s^R); [c(0)])$ in Eq. (4.2).

Unfortunately, an efficient numerical procedure for the minimization problem above is not yet available as it would require expressions for $\{\Pr((s^L, s^R); \alpha, \beta)\}$ as a function of $(\alpha, \beta) \in (-1, \infty)^2$ analogous to the loglikelihood expressions. Such efficient expressions require further transformation of the probabilities at the rooted unranked planar and leaf-unlabelled resolution of transmission trees [35, Lemma 4.1] onto $\mathscr{S}_n$, which is nothing but the planar unordered cousin of Aldous' shape statistics sequence [48, Eqn. (4.1)]. Note that Aldous' shape statistics sequence, even in its original nonplanar form [37], can itself be further projected to various tree shape statistics [48, p. 1231] used routinely in simulation-based studies of transmission trees sketched earlier in Sect. 1.

4.4.2. A Bayesian perspective

We saw that the L_1 perspective has some merits, provided appropriate expressions can be obtained. In defense of the likelihood principle on the other hand, the exact likelihood expressions developed here are most useful to Bayesian methods of inference that further incorporate prior information with the likelihood function based on finite samples. For example, if we know that the underlying contact network is complete then we can invoke appropriate parametric families of prior distributions centred at $(0,0)$, the known MLE for the complete network, with variance in priors reflecting our extent of this prior knowledge. Using the simulation-based approach of Sect. 4 we can obtain the effective Beta-splitting model parameters with standard errors that can inform the formulation of appropriate prior distributions. Such a prior formulation can not only reflect the sample sizes, number of leaves in possibly partially observed transmission trees, etc. at our disposal, but also various aspects of the contact networks from other sources of information such as (i) transportation networks and cell-signal triangulated or GPS-based location networks, etc. that underpin coarse structural information of the host contact process in the context of a communicable disease or (ii) Twitter follower networks in the context of "cultural" transmission events that are defined merely to be retweets or mentions, for example (see Sect. 5.2).

Thus, inducing priors over the parameter space $[-1,\infty)^2$ for more complex analytically intractable models through their effective Beta-splitting models will now be possible, due to our efficient likelihood expressions, using current semi-parametric and variational Bayes methods, including the composition of different effective Beta-splitting models as finite mixtures with possibly unknown number of components over different sub-network neighbourhoods covering the entire contact network.

4.4.3. Natural extensions of the likelihood function

Next we provide some natural extensions of the likelihood expressions developed here.

Remark 6. *The likelihood function can be extended to include branch-lengths – especially under the independent but possibly non-identical exponential waiting-times assumption in the generator of the encompassing continuous time Markov chain given by Eq.* (2.2)*, as is the case for the complete, star and path networks studied here. Leaf labels can also be added according to a tractable labelling process of the population as per Eq.* (3.6)*.*

Remark 7. *More crucially, the likelihood expressions naturally generalize to transmission trees that are only partially grown or in the process of growing at the time of observation so that they have fewer than n leaves encoding all currently infected individuals. For such* partial likelihood *expressions one merely needs to specify $\mathscr{I}(\tau_i)$, the set of internal vertices in the i-th partially grown transmission tree τ_i, in Eq.* (3.13) *and let this specification imply to its consequent expressions. Also note that the second product over internal vertices within tree τ_i in Eq.* (3.13) *allows each $\mathscr{I}(\tau_i)$ to be distinct with its own size, i.e., $|\mathscr{I}(\tau_i)| \in \{1,2,\ldots,n-1\}$. Such natural extensions of the likelihood expressions may be necessary in applied contexts and do produce maximum likelihood estimates with larger standard errors at least in some cases (results not shown here). However, their asymptotic consistency is expected to depend on the details of the SICN itself and on how often one samples trees with nearly n leaves and/or on the initialization mechanisms for the first infector across the r independently drawn partial trees. These partial likelihoods are implementable by modifying the lists being comprehended over vertices in Sect. Appendix A.1 and/or Sect. Appendix A.2, for instance.*

In this work, for concreteness and clarity, we keep the transmission trees full with n leaves representing the final state when the infection has invaded the whole population. Due to this assumption, the *unordered* split-fair frequencies in Eq. (3.11) become sufficient for α and β. When allowing for multiple but independent partial trees of possibly different sizes less than n as per Remark 7, the unordered split-fair frequencies will still remain sufficient, provided the sampling strategy is asymptotically consistent for the underlying initial SICN, albeit one may only be observing split-pair frequencies over a subset of $\mathscr{S}_n$.

The sufficiency of unordered split-fair frequencies over $\mathscr{S}_n$ may no longer hold for (i) more complex extensions of the likelihood which may involve non-static or temporally varying networks or for (ii) multiple simultaneous partial transmission trees grown on the same static network up to possibly random discrete stopping time M from a set of initially infected vertices (especially when the set of their infected vertices become mutually incident before time M). Simplest extensions of the Beta-splitting model via finite mixtures for instance could be constructed to approximate such more realistic models by limiting oneself to the sufficient statistics of *ordered split-pair frequencies*:

$$\{f(s^L,s^R,z) : (s^L,s^R,z) \in \mathscr{S}_n \times \{0,1,2\ldots,M\}\} \ , \tag{4.3}$$

where z is the discrete time index, and over products of such ordered split-pair frequencies. We could allow z to belong to $\mathbb{Z}_0$ and even allow $n \to \infty$ in order to study contact networks in the large population limit with or without appropriate rescaling of discrete time into the continuum, for the purposes of ignoring combinatorial complications of the modelled *individuals* in such limits, where possible and sensible.

5. 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 (SICNs) 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. Appendix 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 with each weight set to 1 in order to focus on the combinatorial skeletons of their associated topological Markov chains. It is relatively straightforward to consider the dynamics on more general networks using the richer language for digraphs [27, 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 vertices in either the strongly connected giant component or in a giant in-component.

We then develop a biparametric Beta-splitting family of models for the growth of transmission trees in Sect. 3 that gives the exact probability of any transmission tree as a function of $\alpha > -1$ and $\beta > -1$. 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 underlying contact network up to how its contact structure affects the infection potential of the infector and infectee after the infection event. The approach avoids the explicit modeling of the underlying contact network (that is typically unobserved or only partially observed) 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 shown 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 the maximum likelihood estimator and sufficient statistics of split-pair frequencies for the Beta-splitting model from independent observations of the transmission trees. Using the distributions on split-pairs we specify equivalence classes of initial SICNs that are indistinguishable by their Beta-splitting transmission trees with the same effective Beta-splitting model through their Beta-projections into the quarter-plane $(-1, \infty)^2$ as conjectured in [49]. We have also shown by simulations coupled with an inferential maximum likelihood procedure that the best-fitting parameters of the effective Beta-splitting models based on samples of trees grown over (i) six deterministic contact networks and four random contact networks seem to be well-separated under their Beta-projections into $(-1, \infty)^2$ and (ii) the Beta-projections of a sequential family of SI-tagged contact networks from $\star_n$ to k_n to p_n do indeed change gradually in $(-1, \infty)^2$ from $(\infty, -1)$ to $(0,0)$ to $(-1, \infty)$. Various natural implications for statistical inference were outlined for future work. In the following sections we discuss some obstacles that need to be overcome for a few other important extensions of this work.

5.1. Towards births and deaths

Recall from Sect. 1 that our construction needs a *tagged contact network*. We only focus on the simplest binary tags (S and I) here and thus limited ourselves to SI-tagged contact networks (SICNs). This simple setting allowed us to obtain our main results here because of the underlying core process being one of pure birth events, whereby individuals ‘are born’ from the set of vertices tagged by S into the set tagged by I at some rate. Thus, there is no ‘death’ here, in the sense of vertices with tag I being retagged for ‘removal’ with tag R (SIR model). The tag R denotes recovery from the infection (after having learnt how to fight the infection at some other rate). In the SIR model the R-tagged individual does not get retagged by S or I. One could represent such a sequence of birth and death events by a binary sequence that could be encoded by Dyck paths that satisfy natural conditions depending on the encoding for the SIR model. By further incorporating such Dyck paths and the three tags one could extend the Markov chains defined here and try to study the distributions they induce on trees with leaves recruited from the vertices with the tag S into vertices with tag I and those with tag I replaced by tag R (and frozen, in the sense of not having any further descendents) at different rates. Extensions to SIR model which allows for the ‘removal’ of infected individuals from the population at a given rate is also conceivable via mapping to percolation on semi-directed networks (see [27, V.B.4] and the references therein).

By considering birth and death processes, as opposed to a pure birth process, one can also make progress on developing transmission processes with only two tags 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. Such as model is interesting from a discrete dynamics viewpoint since one could allow the discrete time z to continue to infinity and study time-asymptotic distributions over SIS transmission trees – the same set of transmission trees for the Markov chains developed here, but with the number of leaf nodes being allowed to be any appropriate number in $\{0,1,\ldots,n\}$ at time $z \in \mathbb{Z}_0$, and with all leaf vertices tagged by I (as implicitly done here) with an absorbing state when the tree has 0 leaves with no infected individual or by conditioning on specific set of Dyck paths [50] that remove such absorption events. The Markov chains for these problems will need more complicated state spaces and immediate precedence rules for state transitions that take the possibly conditioned Dyck paths into account. It is not clear how one could extend the Beta-splitting construction in an interpretable manner for these settings.

5.2. Towards cultural transmissions

The insights developed here through the SICNs and their transmission trees are also applicable to the simplest models of “meme” [1, p. 192] evolution in online social media networks [2] through transmission events that can be distilled (admittedly naively) from observable actions such as ‘likes’, ‘mentions’, ‘retweets’ and ‘+1s’ along with any concomitant comments. See a dataset spotlight [51] in the official blog of Kaggle.com for a specific extremist cultural context. We would like to point out that the 55 models and their parameter choices were partly informed by empirical insights from extracting, transforming, loading and exploring the raw tweets available to Twitter developers via DataFrame and GraphFrame APIs in Apache Spark [52].

Although the SI model studied here is the simplest two-state Finite Markov Information Exchange (FMIE) process [3, Sec. 2.2] called the Pandemic Process [3, Sec. 7], it is shown to be a fundamental building-block [3, Sec. 3.2,7] for a large class of FMIE processes which includes various classical epidemic models [see 3, Sec. 8,9] and has some remarkable properties: (i) SI model exhibits the fastest possible spread of information in any FMIE model [3, Sec. 3.2] and (ii) 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 [27, II.A]. These properties of the model make the Beta-splitting tree distributions we provide here particularly useful for extensions into applied operations research along Markov control processes that are aimed at influencing the growth of certain undesirable aspects of transmission trees, over carefully filtered [51] extremist cultural networks, through interventions orchestrated from appropriate control spaces, including artificially intelligent chatbots, for instance.

5.3. Towards other tree resolutions

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 [35]. 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.

5.4. Towards dynamic contact networks

The jump Markov chain of the transmission process on static SI-tagged contact networks (SICNs) is a prerequisite for contemplating appropriate partial orders on the set of all SICNs 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 [27, VII.B.7] and the references therein) in complete agreement with Theorem 3 on equivalence classes over initial SICNs.

Let $\mathscr{N}_n$ be 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. By fixing an initial infector, say $\imath_0$, we can use the Beta-projection: $\mathscr{B}^{\downarrow}(w) : \mathscr{N}_n \to (-1,\infty)^2$, to map each (connected) contact network w in $\mathscr{N}_n$ to the exact maximum likelihood estimate $(\alpha,\beta) \in (-1,\infty)^2$ while maintaining the partial ordering between contact networks in $\mathscr{N}_n$. Such a planar geometric embedding of the contact networks, given by $\mathscr{B}^{\downarrow}(\alpha,\beta)$ 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 at (α,β) and that specified directly by the initial contact networks in $\mathscr{B}^{\uparrow}(\alpha,\beta)$. Future research on Markov chains with transitions over partially ordered contact networks as well as transmission trees could build upon insights from our simpler setting here.

Authors' Contributions

DW introduced RS to transmission trees in May 2014. Both authors set out the research programme and formalized the Markov chains. DW made the crucial observation of complementary combs under star and path networks and this allowed RS to expand and adapt the work done in [35] for the epidemic context. RS extended the three Theorems based on comments and suggestions by anonymous reviewers, wrote all the accompanying SageMath/Python code and conducted experiments with tweets that informed the simulation study in Table 1. DW conducted the simulations in Fig. 5 and all calculations involving continuous time branch-lengths. Both authors revised the paper multiple times.

Acknowledgements

RS thanks Robert C. Griffiths for combinatorial guidance on planar binary trees and pointers to ballot theorems, Bhalchandra Thatte and Charles Semple for pointers on partially ordered sets of networks, Anusha Raazesh for left-branching *d*-shark trees, and Daniel Stouffer and Tanja Stadler for encouragement. This work grew out of a lecture in a postgraduate biometrics course at Cornell University (2014), refined by feedback from Tom Britton, Mia Deijfen, Federica Giardina and Pieter Trapman at an Epidemics Group meeting in Stockholm University (2015) and extended significantly due to suggestions by an editor and two anonymous reviewers (2016). 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. Twitter data analytics by RS is supported by Databricks Academic Partners Program (under Allison Kunz). DW was partially supported by Marsden grant UOA1324 from the Royal Society of New Zealand.

[1] R. Dawkins, The Selfish Gene, Oxford University Press, Oxford, UK, 1976.

[2] O. Solon, Richard Dawkins on the internet's hijacking of the word 'meme' (2013).
URL http://www.wired.co.uk/news/archive/2013-06/20/richard-dawkins-memes

[3] D. Aldous, Interacting particle systems as stochastic social dynamics, Bernoulli 19 (4) (2013) 1122–1149.

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

[5] M. J. Sackin, "Good" and "bad" phenograms, Systematic Zoology 21 (1975) 225–226.

[6] D. H. Colless, Review of phylogenetics: the theory and practice of phylogenetic systematics, Systematic Zoology 31 (1982) 100–104.

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

[8] D. Ludwig, Final size distribution for epidemics, Mathematical Biosciences 23 (1) (1975) 33–46.

[9] L. Pellis, N. M. Ferguson, C. Fraser, The relationship between real-time and discrete-generation models of epidemic spread, Mathematical Biosciences 216 (1) (2008) 63–70.

[10] T. House, J. V. Ross, D. Sirl, How big is an outbreak likely to be? methods for epidemic final-size calculation, Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences 469 (2150).

[11] D. T. Haydon, M. Chase–Topping, D. J. Shaw, L. Matthews, J. K. Friar, J. Wilesmith, M. E. J. Woolhouse, 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) (2003) 121–127.

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

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

[14] R. J. F. Ypma, W. M. van Ballegooijen, J. Wallinga, Relating phylogenetic trees to transmission trees of infectious disease outbreaks, Genetics 195 (3) (2013) 1055–1062.

[15] B. T. Grenfell, O. G. Pybus, J. R. Gog, J. L. Wood, J. M. Daly, J. A. Mumford, E. C. Holmes, Unifying the epidemiological and evolutionary dynamics of pathogens, science 303 (5656) (2004) 327–332.

[16] S. D. Frost, O. G. Pybus, J. R. Gog, C. Viboud, S. Bonhoeffer, T. Bedford, Eight challenges in phylodynamic inference, Epidemics 10 (2015) 88–92, challenges in Modelling Infectious Disease Dynamics.

[17] G. E. Leventhal, R. Kouyos, T. Stadler, V. von Wyl, S. Yerly, J. Böni, C. Cellerai, T. Klimkait, H. F. Günthard, S. Bonhoeffer, Inferring epidemic contact structure from phylogenetic trees, PLoS Computational Biology 8 (3) (2012) e1002413.

[18] S. D. W. Frost, E. M. Volz, Modelling tree shape and structure in viral phylodynamics, Philosophical Transactions of the Royal Society of London B: Biological Sciences 368.

[19] E. B. O'Dea, C. O. Wilke, Contact heterogeneity and phylodynamics: how contact networks shape parasite evolutionary trees, Interdisciplinary perspectives on infectious diseases 2011.

[20] D. Welch, Is network clustering detectable in transmission trees?, Viruses 3 (6) (2011) 659–676.

[21] C. Colijn, J. Gardy, Phylogenetic tree shapes resolve disease transmission patterns, Evolution, Medicine, and Public Health.

[22] J. O. Lloyd-Smith, S. J. Schreiber, P. E. Kopp, W. Getz, Superspreading and the effect of individual variation on disease emergence, Nature 438 (7066) (2005) 355–359.

[23] A. J. Leigh Brown, S. J. Lycett, L. Weinert, G. J. Hughes, E. Fearnhill, D. T. Dunn, Transmission network parameters estimated from hiv sequences for a nationwide epidemic, Journal of Infectious Diseases 204 (9) (2011) 1463–1469.

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

[25] C. Groendyke, D. Welch, D. R. Hunter, Bayesian inference for contact networks given epidemic data, Scandinavian Journal of Statistics 38 (3) (2011) 600–616.

[26] C. Groendyke, D. Welch, D. R. Hunter, A network-based analysis of the 1861 hagelloch measles data, Biometrics 68 (3) (2012) 755–765.

[27] R. Pastor-Satorras, C. Castellano, P. Van Mieghem, A. Vespignani, Epidemic processes in complex networks, Rev. Mod. Phys. 87 (2015) 925–979.

[28] P. Holme, Modern temporal network theory: a colloquium*, Eur. Phys. J. B 88 (9) (2015) 234.

[29] J. F. C. Kingman, The coalescent, Stochastic Processes and their Applications 13 (1982) 235–248.

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

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

[32] T. Stadler, S. Bonhoeffer, Uncovering epidemiological dynamics in heterogeneous host populations using phylogenetic methods, Philosophical Transactions of the Royal Society of London B: Biological Sciences 368 (1614).

[33] T. G. Vaughan, D. Kühnert, A. Popinga, D. Welch, A. J. Drummond, Efficient bayesian inference under the structured coalescent, Bioinformatics 30 (16) (2014) 2272–2279.

[34] D. A. Rasmussen, E. M. Volz, K. Koelle, Phylodynamic inference for structured epidemiological models, PLoS Comput Biol 10 (4) (2014) e1003570.

[35] R. Sainudiin, A. Véber, A beta-splitting model for evolutionary trees, Royal Society Open Science 3 (5).

[36] T. S. Developers, Sage Mathematics Software (Version 6.8) (2015).
URL `http://www.sagemath.org`

[37] D. J. Aldous, Stochastic models and descriptive statistics for phylogenetic trees, from Yule to today, Statist. Sci. 16 (1) (2001) 23–34.

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

[39] P. Flajolet, R. Sedgewick, Analytic Combinatorics, 1st Edition, Cambridge University Press, New York, NY, USA, 2009.

[40] W. Hofschuster, W. Krämer, C-xsc 2.0: A c++ library for extended scientific computing, in: Numerical Software with Result Verification, 2003, pp. 15–35.

[41] P. Erdős, A. Rényi, On random graphs. I, Publ. Math. Debrecen 6 (1959) 290–297.

[42] E. N. Gilbert, Random graphs, Ann. Math. Statist. 30 (4) (1959) 1141–1144.

[43] B. Bollobás, Random Graphs, 2nd Edition, Cambridge University Press, 2001, cambridge Books Online.

[44] A. Steger, N. C. Wormald, Generating random regular graphs quickly, Comb. Probab. Comput. 8 (4) (1999) 377–396. `doi:10.1017/S0963548399003867`.

[45] J. H. Kim, V. H. Vu, Generating random regular graphs, in: Proceedings of the Thirty-fifth Annual ACM Symposium on Theory of Computing, STOC ’03, ACM, New York, NY, USA, 2003, pp. 213–222. `doi:10.1145/780542.780576`.

[46] D. J. Watts, S. H. Strogatz, Collective dynamics of ‘small-world’ networks., Nature 393 (6684) (1998) 409–10.

[47] A.-L. Barabási, R. Albert, Emergence of scaling in random networks, Science 286 (5439) (1999) 509–512.

[48] R. Sainudiin, T. Stadler, A. Véber, Finding the best resolution for the kingman–tajima coalescent: theory and applications, Journal of Mathematical Biology 70 (6) (2015) 1207–1247.

[49] R. Sainudiin, D. Welch, The transmission process: A combinatorial stochastic process on binary trees over the contact network of hosts in an epidemic, UCDMS Research Report 2015/4 (2015) 1–31.
URL `http://www.math.canterbury.ac.nz/~r.sainudiin/preprints/20151210_transmissionProc.pdf`

[50] L. Addario-Berry, B. A. Reed, Ballot Theorems, Old and New, Springer Berlin Heidelberg, Berlin, Heidelberg, 2008, pp. 9–35.

[51] M. Risdal, Dataset Spotlight: How ISIS Uses Twitter — Khuram Zaman (2016).
URL `http://blog.kaggle.com/2016/06/03/dataset-spotlight-how-isis-uses-twitter/`

[52] M. Zaharia, M. Chowdhury, M. J. Franklin, S. Shenker, I. Stoica, Spark: Cluster computing with working sets, in: Proceedings of the 2Nd USENIX Conference on Hot Topics in Cloud Computing, HotCloud'10, USENIX Association, Berkeley, CA, USA, 2010, pp. 10–10.

Appendix 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.

Appendix A.1. Simulating the transmission process

```
def CountsDict(X):
    '''convert a list X into a Dictionary of counts or frequencies’’’
    CD = {}
    for x in X:
        CD[x] = (CD[x] + 1) if (x in CD) else 1
    return CD

def markAsInfected(C,v,m):
    '''mark vertex 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 vertex 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 ’’’
    LBT = LabelledBinaryTree
    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
    #print ascii_art(Ttree)
    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 vertex 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 vertex initialI'''
    #initialisation of SICN
    z=0 # infection event count
    markAsInfected(C,initialI,'infected')
    infectedIs = [initialI]
    popSize=C.order()
    # initialisation of Transmission Tree
    pathsDict={} # dictionary of vertices -> paths from root in tree
    LBT = LabelledBinaryTree
    # 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)
        numberInfected=len(currentSOE)
        nextEdge = currentSOE[randrange(0,numberInfected)]
        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]))
        #comment the next line in large simulations!
        print "step z = ",z; print ascii_art(T); print "--------------------"
    return T.as_ordered_tree(with_leaves=False)
# 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]]
```

Appendix A.2. Likelihood of Beta-splitting transmission trees

```
def splitsSequence(T):
    '''return a list of tuples (left,right) split sizes at each split vertex'''
    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)))

def negLogLkl_SplitPairCounts(spc,a,b):
    '''-log likelihood of multiple independent ranked planar trees
       through their sufficient statistics of the frequence of
       split-pair counts spc= [(nL_i,nR_i,c_i): i=1,..,K]
       under Beta-splitting model
       a,b>-1, where (a+1,b+1) are the parameters of the beta
       distribution. This implements first Equation in Thm 1 involving beta functions'''
    return -RR(sum(map(lambda x:
                x[2]*log(1.0*beta(x[0]+a+1,x[1]+b+1)/beta(a+1,b+1)), spc)))

def splitPairsCounts(TS):
    '''list of the frequency of all distinct split-pairs,
       i.e. (# of left splits, # right splits)
       below each internal vertex in each transmission tree
       in the list TS of transmission trees'''
    splitPairCounts=sorted(CountsDict(flatten([splitsSequence(t) \
                                          for t in TS],max_level=1)).items())
    return [(x[0][0],x[0][1],x[1]) for x in splitPairCounts]

def splitPairsCountsDict(TS):
    '''dictionary of the frequency of all distinct split-pairs,
       i.e. (# of left splits, # right splits)
       below each internal vertex in each transmission tree in the list
       TS of transmission trees'''
    sD = CountsDict(flatten([splitsSequence(t) for t in TS],max_level=1))
    return sD

def logLklOfASplitPair(a,b,nL,nR):
    '''beta(nL+a+1,nR+b+1)/beta(a+1,b+1) without beta functions via Eqn 2 in Thm 1'''
    A1=sum([log((b+j)/(b+j+a)) for j in range(nR+1)])
    A2=sum([log((a+i)/(a+i+b+nR+1)) for i in range(nL+1)])
    A3=log(b*a/((a+b)*(a+b+1)))
    return A1+A2-A3

def negLogLkl_SplitPairCounts2(spc,a,b):
    '''-log likelihood of multiple independent ranked planar trees
       through their sufficient statistics of the frequency of
       split-pair counts spc= [(nL_i,nR_i,c_i): i=1,..,K]
       under Beta-splitting model
       a,b>-1, where (a+1,b+1) are the parameters of the beta
       distribution. This implements second Equation in Thm 1 without beta functions'''
    return -(sum(map(lambda x:
                x[2]*logLklOfASplitPair(a,b,x[0],x[1]), spc)))

def LklOfASplitPair(a,b,nL,nR):
    '''beta(nL+a+1,nR+b+1)/beta(a+1,b+1) without beta functions via Eqn 2 in Thm 1'''
    A1=prod([((b+j)/(b+j+a)) for j in range(nR+1)])
    A2=prod([((a+i)/(a+i+b+nR+1)) for i in range(nL+1)])
    A3=(b*a/((a+b)*(a+b+1)))
    return (A1*A2)/A3

def negLogLkl_SplitPairCounts2Prod(spc,a,b):
    '''- log likelihood of multiple independent ranked planar trees
       through their sufficient statistics of the frequency of
       split-pair counts spc= [(nL_i,nR_i,c_i): i=1,..,K]
       under Beta-splitting model
       a,b>-1, where (a+1,b+1) are the parameters of the beta
       distribution -- This implements second Equation in Thm 1 without beta functions'''
    return -(sum(map(lambda x:
                x[2]*log(LklOfASplitPair(a,b,x[0],x[1])), spc)))

# demo of the mle for a complete graph with 50 vertices 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
# simulation settings
n=50
reps=10
# trial 1: make a list of 10 independent transmission trees on complete graph with 50 vertices
ts1=[transmissionProcessTC(graphs.CompleteGraph(n).to_directed(),0)  for _ in range(reps)]
```

```
spc1=splitPairsCounts(ts1)
# trial 2: make another list of 10 indep transmission trees on complete graph with 50 vertices
ts2=[transmissionProcessTC(graphs.CompleteGraph(n).to_directed(),0)  for _ in range(reps)]
spc2=splitPairsCounts(ts2)

# fastest and numerically most robust MLE computation using sufficient statistics
# USE this for larger simulations
def negLkl1(AB):
        return negLogLkl_SplitPairCounts2(spc1,AB[0],AB[1])

def negLkl2(AB):
        return negLogLkl_SplitPairCounts2(spc2,AB[0],AB[1])

%time mle=minimize_constrained(negLkl1,[c_1,c_2],[0.0,0.0],disp=0) #MLE for trial 1
print [n,reps,mle]
%time mle=minimize_constrained(negLkl2,[c_1,c_2],[0.0,0.0],disp=0) #MLE for trial 2
print [n,reps,mle]

# output
CPU time: 4.73 s, Wall time: 9.52 s
[50, 10, (-0.06644703305884028, -0.05022885910313697)]
CPU time: 9.34 s, Wall time: 18.87 s
[50, 10, (0.004665943054046424, -0.043007797370934346)]
```

Appendix A.3. Transmission trees under various contact networks

```
n = 50 # number of individuals in the population
reps = 1 # number of independent transmission trees

# Bidirectional Circular Network
ts=[transmissionProcessTC(digraphs.Circulant(n,[1,-1]),0)  for _ in range(reps)]

# Balanced Tree Network
ts=[transmissionProcessTC(graphs.BalancedTree(3,2).to_directed(),0)  for _ in range(5)]
ascii_art(ts[0]) # labelled transmission tree is a "left-branching 3-shark"
                     _____________1______________
                    /                           /
         __________2___________              __3___
        /                     /             /     /
  _____5______             __4____        _6___  i4
 /           /            /      /       /    /
i0        __7___       __8___   i10     12_  i6
         /     /      /     /          /  /
       _9___  i7     10_   i11        i1 i5
      /    /        /  /
```

```
     11_  i9       i3 i12
    /  /
   i2 i8

# Toroidal Grid Network

# 2D
def toroidal2DGrid(n):
    G = graphs.GridGraph([n,n])
    #G.show()  # long time
    G.add_edges([((i,0),(i,n-1)) for i in range(n)])
    G.add_edges([((0,i),(n-1,i)) for i in range(n)])
    G.relabel()
    return G
ts=[transmissionProcessTC(toroidal2DGrid(10).to_directed(),0)  for _ in range(10)]

# 3D
def toroidal3DGrid(n):
    G = graphs.GridGraph([n,n,n])
    #G.show()  # long time
    G.add_edges([((0,i,j),(n-1,i,j)) for i in range(n) for j in range(n)])
    G.add_edges([((i,0,j),(i,n-1,j)) for i in range(n) for j in range(n)])
    G.add_edges([((i,j,0),(i,j,n-1)) for i in range(n) for j in range(n)])
    G.relabel()
    return G
ts=[transmissionProcessTC(toroidal3DGrid(m).to_directed(),0)  for _ in range(10)]

# Some Random Contact Networks

## Erdos Reyni Network
def ErdosReyniConnectedCompOf0(n,p,reps):
    '''return reps many transmission trees from the connected component
    containing initial infection vertex 0'''
    ts=[]
    i=0; MAXTrials=10000; successfulTrials=0;
    while (successfulTrials<reps or i>MAXTrials):
        i=i+1
        g0=graphs.RandomGNP(n,p).to_directed()
        g=g0.subgraph(g0.connected_component_containing_vertex(0))
        if g.order()>1:
            #print g.order(), g.size()
            ts.append(transmissionProcessTC(g,0))
            successfulTrials=successfulTrials+1
    return ts

# demo of the mle
```

```
mleList=[]
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=100
#Lambdas=[floor(RR(n/2^i)) for i in range(ceil(log(n,2)))]; Lambdas.reverse(); #Lambdas
Lambdas=sorted(Set([floor(RR(n/(5/4)^i)) for i in range(ceil(log(n,5/4)))]));
for L in Lambdas:
    # edge prob in Erdos-Reyni random graph on n vertices = lambda/n
    # where lambda = expected degree of a vertex
    prob=RR(L/n)
    reps=30 # all reps have the same unlabelled transmission tree topology
    for i in range(5):
        ts= ErdosReyniConnectedCompOf0(n,prob,reps)
        spc=splitPairsCounts(ts)
        def negLkl(AB):
            return negLogLkl_SplitPairCounts2(spc,AB[0],AB[1])
        mle=minimize_constrained(negLkl,[c_1,c_2],[0.0,0.0],disp=0)
        # mean number of connected components
        # = pop size = # leaves in transmission tree
        meanNumCc=RR(mean([(t.node_number()+1)/2 for t in ts]))
        mleL=[n, prob, L, reps, meanNumCc, mle]
        print mleL
        mleList.append(mleL)

## Random Regular
d,n=4,100 # n>d>2, and n*d is even
reps=10
ts=[transmissionProcessTC(graphs.RandomRegular(d,n).to_directed(),0) \
                                                  for _ in range(reps)]

## Small World Model

def findMaxDegAndItsVertex(gr):
    '''find the maximum degree in a graph and its vertex with smalelst ID'''
    maxD=0; maxDv=0;
    for vd in gr.degree_iterator(range(n),True):
        #print vd
        if vd[1] > maxD:
            maxD=vd[1]; maxDv=vd[0];
    return maxD,maxDv

import networkx # so we use the networkx implementation
def ConnectedSmallWorldFromMostpopular(n,k,p,reps):
    '''return reps many transmission trees from the connected small-world network
       initialized from the most popular smallest-labelled vertex'''
    ts=[]
    i=0; MAXTrials=10000; successfulTrials=0;
    while (successfulTrials<reps or i>MAXTrials):
        i=i+1
        g0 = DiGraph(networkx.watts_strogatz_graph(n,k,p))
        if g0.is_connected(): # just making sure we have connected network
            #print g0.order(), g0.size()
            maxDV=findMaxDegAndItsVertex(g0)
            # to initialize from the most popular smallest labelled vertex
            ts.append(transmissionProcessTC(g0,maxDV[1]))
            successfulTrials=successfulTrials+1
    return ts

def ConnectedSmallWorldFromAnywhere(n,k,p,reps):
    '''return reps many transmission trees from the connected small-world network
       initialize from a random vertex'''
    ts=[]
    i=0; MAXTrials=10000; successfulTrials=0;
    while (successfulTrials<reps or i>MAXTrials):
        i=i+1
        g0 = DiGraph(networkx.watts_strogatz_graph(n,k,p))
        if g0.is_connected(): # just making sure we have connected network
            #print g0.order(), g0.size()
            # to initialize at a random vertex, say 0 - it's too noisy!
            ts.append(transmissionProcessTC(g0,0))
            successfulTrials=successfulTrials+1
    return ts

def spc2spcRF(spc):
    '''to turn the split-pair counts into split-pair relative frequencies to
       interrogate local optimization madness... - should be really using rigorous
       interval global optimization here!'''
    s=sum([x[2] for x in spc])
    spcRf=[]
    for i in range(len(spc)):
        spcRf.append((spc[i][0],spc[i][1],spc[i][2]/s))
    return spcRf

# demo of the mle
c_1 = lambda p: p[0]+0.9999999 # constraint for alpha > -1
c_2 = lambda p: p[1]+0.9999999 # constraint for beta > -1
# simulation settings
mles=[]
n,k,p,reps,repeatMLE=50,5,0.10,30,5
for i in range(repeatMLE):
    # small world initialized from the maximal degree vertex with the smallest label
    #ts=ConnectedSmallWorldFromMostpopular(n,k,p,reps)
    # small world initialized from a random vertex (actually label 0)
    ts=ConnectedSmallWorldFromAnywhere(n,k,p,reps)
    #--use MLE-equivalent line line below if spc-freqs need normalization for numerics...
    #spc=spc2spcRF(splitPairsCounts(ts))
    spc=splitPairsCounts(ts)
    def negLkl(AB):
        #--our standard method - not ok for n ~ 50 due to
        #--SICN-circularity’s (k=2) implications for number screen wrt log...
        #--our standard method below not ok for n>=50, k=2
        #return negLogLkl_SplitPairCounts2(spc,AB[0],AB[1])
        #--this product likelihood method ok for n=50 but not n=100,k=2 -->boundary...
        return negLogLkl_SplitPairCounts2Prod(spc,AB[0],AB[1])
    mle=minimize_constrained(negLkl,[c_1,c_2],[.0,.0],disp=0)
    x = [i, n, k, p, reps, mle, negLkl(mle)]
    print x
    mles.append(x)
y=mean([x[5][0] for x in mles]),std([x[5][0] for x in mles]), \
mean([x[5][1] for x in mles]),std([x[5][1] for x in mles]); [x.N(digits=4) for x in y]

## Preferential Attachment Model

def findMaxDegAndItsVertex(gr):
    ’’’find the maximum degree in a graph and its vertex with smalelst ID’’’
    maxD=0; maxDv=0;
    for vd in gr.degree_iterator(range(n),True):
        #print vd
        if vd[1] > maxD:
            maxD=vd[1]; maxDv=vd[0];
    return maxD,maxDv

def PrefAttachmentFromMostPopular(n,m,reps):
    ’’’return reps many transmission trees from the preferential attachment model
    with initial infection from the smallest vertex with the maximum degree’’’
    ts=[]
    for i in range(reps):
        g=graphs.RandomBarabasiAlbert(n,m)
        maxDV=findMaxDegAndItsVertex(g)
        ts.append(transmissionProcessTC(g.to_directed(),maxDV[1]))
    return ts

# demo of the mle
c_1 = lambda p: p[0]+0.9999999 # constraint for alpha > -1
c_2 = lambda p: p[1]+0.9999999 # constraint for beta > -1
# simulation settings
mles=[]
n,m,reps,repeatMLE=100,1,30,10
```

```
for i in range(repeatMLE):
    ts=PrefAttachmentFromMostPopular(n,m,reps)
    spc=splitPairsCounts(ts)
    def negLkl(AB):
        return negLogLkl_SplitPairCounts2(spc,AB[0],AB[1])
    mle=minimize_constrained(negLkl,[c_1,c_2],[0.0,0.0],disp=0)
    x = [i, n, m, reps, mle]
    print x
    mles.append(x)

# to initiliaze from any one of the initial m vertices
# (0 is equally likely in initial vertex set {0,...,m-1})
ts=[transmissionProcessTC(graphs.RandomBarabasiAlbert(n,m).to_directed(),0) \
                                                        for _ in range(reps)]
```

Appendix A.4. 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
```

Appendix A.5. Transmission tree distributions

See [35, Appendix: Algorithms] for simulating transmission trees and obtaining the probability for various equivalence classes of trees under the Beta-splitting model.