Statistics and Computing (2020) 30:61–75
https://doi.org/10.1007/s11222-019-09865-1

Epidemiologic network inference

Pierre Barbillon[1] · Loïc Schwaller[2] · Stéphane Robin[1] · Andrew Flachs[3] · Glenn Davis Stone[4]

Received: 7 November 2018 / Accepted: 18 February 2019 / Published online: 1 March 2019

Abstract
In many epidemiologic models, a disease is assumed to spread along a contact network. We aim to infer this network, in addition to the epidemiologic model parameters, from the binary status of individuals observed throughout time. We perform an exact evaluation of the probability for each edge to be part of the network by using the matrix-tree theorem on the set of vertices made of the individual status at all times. This leads to a computational complexity of order $\mathcal{O}(mn^2)$, where n is the number of individuals and m the length of the time series. Simulations are provided to demonstrate the efficiency of the proposed method, and it is applied on data concerning seed choices by farmers in India and on data on a measles outbreak.

Keywords Contact network · Matrix-tree theorem · Propagation path · SIS Model

1 Introduction

The spread of epidemics within a given population has been extensively studied for many years in the pursuit of a better understanding of their dynamics and mechanisms. The spread is known to depend on numerous parameters such as contamination and extinction rates (Welch et al. 2011; Brauer et al. 2012). From a statistical point of view, a significant amount of work has been put into the estimation of such parameters in a series of classical models (Neal and Roberts 2004). Whenever the contacts between individuals can be organized according to a network, the topology of this network is also known to affect the dynamics of an epidemic (see Wang et al. 2003; Yang et al. 2015; Barbillon et al. 2015, and references therein).

In this paper, we are interested in both the estimation of the parameters of the epidemic and the reconstruction of the social network along the edges of which it spreads, with a special emphasis on the latter. More specifically, we consider the general framework of a susceptible–infected–susceptible (SIS) model where the status (sick or healthy) of each individual from a given population is observed at a series of discrete times. We aim to reconstruct the path of the spread or, at least, the contact network that actually structures the diffusion of the epidemic. Put in the framework of Welch et al. (2011), the disease data that we rely on are quite poor as we only observe the binary status of each individual throughout time, without any additional information about the pathogen or contacts. This framework can also apply to the propagation of goods or information.

The reconstruction of a contact network based on epidemic data has been previously considered in several pieces of work (see, e.g., Welch et al. 2011, for a review). Britton and O'Neill (2002) or Ray and Marzouk (2008) do not directly infer the network itself but assume it follows a random graph model and infer the parameters of this model, which encodes the global topological properties of the network. In both Groendyke et al. (2011) and Groendyke et al. (2012), a susceptible–exposed–infectious–removed (SEIR) model is assumed for the epidemic and combined with a random graph model for the contact network (Erdős–Rényi model in the former and Exponential Random Graph Model

This work has been partially supported by United States Department of Education Jacob K. Javits Fellowship, the National Geographic Society Young Explorer's Grant 9304-13, the John Templeton Foundation (Glenn Davis Stone PI), Washington University in St. Louis and MIRES funded by the Applied Mathematics and Informatics department of INRA.

✉ Pierre Barbillon
pierre.barbillon@agroparistech.fr

[1] UMR MIA-Paris, AgroParisTech, INRA, Université Paris-Saclay, 75005 Paris, France

[2] Mathematical Institute, Leiden University, P.O. Box 9512, 2300 Leiden, RA, The Netherlands

[3] Department of Anthropology, Purdue University, West Lafayette, Indiana, USA

[4] Department of Anthropology, Washington University in St. Louis, St. Louis, Missouri, USA

in the latter). By posterior sampling, they can recover posterior probabilities on the possible sources of infection for individuals.

We consider the case where cascades of propagation events are observed and where the goal is to estimate the contact network over which the cascades can propagate. In the setting where node status are observed at continuous times, Myers and Leskovec (2010) propose a penalized likelihood approach, where the ℓ_1 penalization encourages the sparsity of the inferred network. The inference algorithm takes advantage of the convexity of the penalized likelihood to estimate the parameters of the epidemic model and the edge probabilities.

In the context of a susceptible–infected (SI) model, all the information can be encoded in the times at which nodes get infected. This encoding imposes some constraints on the network to be inferred. Although the data are time stamped, the network reconstruction problem actually becomes static. In this framework, Gomez-Rodriguez et al. (2012) propose an efficient algorithm to search for the most likely network, faster than the ones developed in Myers and Leskovec (2010) and Gomez-Rodriguez et al. (2011) at the cost of a simpler context.

From a statistical point of view, the inference of a contact network can be seen as a special instance of a more general problem of network or structure inference. This problem can be stated as follows: based on observations collected on a set of nodes, one would like to infer the network encoding the interactions between these nodes. All network inference methods have to deal with a huge number of possible networks, that is $2^{n(n-1)/2}$ undirected structures on n nodes.

Chow and Liu (1968) proposed to circumvent the impossibility of an exhaustive exploration of this set by reducing the search space to the set of acyclic undirected connected graphs, also known as spanning trees, whose cardinality is only n^{n-2}. As the interaction network is unlikely to follow a single spanning tree, Meilă and Jaakkola (2006) and Kirshner (2007) suggested to consider an averaging over all possible spanning trees. Averaging over this set is feasible with a reasonable computational burden through closed-form expressions based on the matrix-tree theorem, which we recall in Sect. 3.1 as it constitutes the cornerstone of our approach. This theorem was previously used for contact network reconstruction in the essentially static context of SI models by Gomez-Rodriguez et al. (2012). The authors of this article were interested in retrieving the most likely contact network. We propose a different usage of the matrix-tree theorem to average over possible spanning trees, in order to estimate the probability for each possible edge to be part of the contact network. To our knowledge, this tool has not yet been used this way or in the context of a dynamic model, when data are collected on nodes throughout time.

In an SIS model, nodes can recover from an infection and then get infected several times. A propagation must therefore be expressed as a directed acyclic graph (DAG) on $n \times m$ vertices corresponding to n nodes observed at times $\{1, \ldots, m\}$. This differs from the Independent Cascade Model and its derivatives, where nodes can only get infected once (since the propagation follows an SI model) and where a cascade can be seen as a DAG on the nodes themselves.

Our main contribution is to leverage the Markovian structure of our model with respect to time, thereby making an efficient use of the matrix-tree theorem. As the states of nodes at time t can only be influenced by their states at time $t-1$, the Laplacian matrix used to sum up over all possible propagation paths (or trees) has a block upper triangular form and its determinant can be computed efficiently.

The main novelties of this contribution are (i) the inference of the contact network under an SIS model in discrete time, which allows for re-infection and (ii) the use of the matrix-tree theorem in a temporal setting which, to the best of our knowledge, is completely new and turns out to be very efficient. The inference of the contact network is model-free in the sense that we do not posit any random graph model upon the network underpinning the epidemic. We also show how the inference procedure we propose can be extended to learn prior weights on the edges of the contact network.

The paper is organized as follows. In Sect. 2, the models for the structure of the network and the diffusion are presented. The inference of the model parameters and the network are detailed in Sect. 3. Section 4 demonstrates the efficiency of the proposed method in contrasting situations. In Sect. 5, we apply the proposed methodology to data on annual seed choices collected from Telangana farmers in India to infer the network of influences between farmers, which might drive their collective seed choices. We also deal with a dataset on a measles outbreak. Some possible extensions are discussed in the concluding section.

2 Model

We consider the observation over time of n nodes. At any given time, each of these nodes can be in one of two states: 'infected' or 'susceptible'. We assume that infection events propagate along the edges of an unobserved contact network interconnecting the nodes. We consider the following SIS (susceptible–infected–susceptible) model. At each time step, an infected node can get cured with probability e or stay infected with probability $1-e$, while a susceptible node can be infected by one of its neighbors with probability c. The infection process is made explicit in the next paragraph. Our main goal is to infer the topology of the contact network, along the edges of which the infection events can occur.

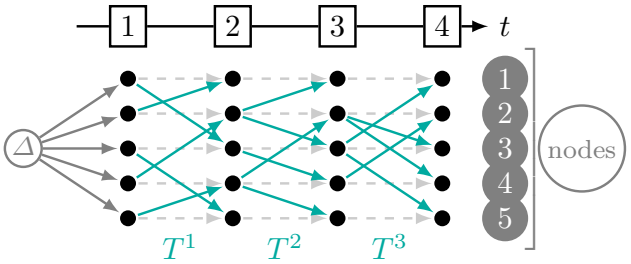

Fig. 1 Graphical model associated with an example of tree $T = (T^1, T^2, \dots)$

2.1 Network model

We consider n nodes $\{1, \dots, n\}$ observed at times $\{1, \dots, m\}$ and we let $V := \{(i,t) : 1 \le i \le n, 1 \le t \le m\}$ be the product set of the set of nodes with time. Our modeling of an epidemic is based on the assumptions that, (i) at each time step, each node can only be infected by one other node, which we call its parent, and taken among its neighbors in the contact network and that (ii) the parent of a node can change from one step to another. If node i is the parent of node j at time $t+1$, we put a directed edge from (i,t) to $(j, t+1)$. The infection path is the directed graph on the set of vertices V made of all these parent–child edges. An example of infection path is given in Fig. 1. This path is distinct from the contact network that we are trying to infer, the latter being defined on $\{1, \dots, n\}$ and not on V, but they are related as the contact network determines the set of possible parents for each node in the infection path.

Assumption (i) can seem quite strong, but it is consistent with the fact that in most epidemiological models, an individual is actually infected by only one individual among many possible candidates. This latter point is accounted for through assumption (ii), which allows the infectious contact to vary from one time to another.

We would like to emphasize that the infection path actually describes the way potential infection events could occur, as the transition model described in Sect. 2.2 allows for infections to miss with a given probability.

Because of assumption (i), the infection path is actually a tree, once a root vertex Δ and edges from Δ to vertices $(i, 1)$, $1 \leqslant i \leqslant n$, have been added (*cf.* Figure 1 for an illustration). Formally, we let $V^* := \{\Delta\} \cup V$ be the augmented set with $nm + 1$ elements and we let T denote the tree on V^* resulting from the completion of the infection path with the root Δ.

Because its edges only link vertices at time t to vertices at time $t+1$, T can be sliced into $m-1$ oriented bipartite graphs T^t $(1 \le t \le m-1)$, that each define the parents of nodes at time t. More specifically, we denote by $\{[ij] \in T^t\}$ the event that makes i the parent of j during the transition from time t to time $t+1$.

In the proposed modeling, the tree T is random and its distribution is defined as follows. We associate a prior weight β_{ij} with each oriented edge $[ij]$ and assume that, at each time t, each node j samples its parent i with probability proportional to β_{ij}. As a consequence, the probability of a tree T is

$$P(T) = B^{-1} \prod_{t=1}^{m-1} \prod_{[ij] \in T^t} \beta_{ij} \tag{1}$$

where $B := \sum_T \prod_{t=1}^{m-1} \prod_{[ij] \in T^t} \beta_{ij}$.

The weights β_{ij} can be seen as a way to account for some prior knowledge about the likelihood of each edge or as parameters of the model that need to be inferred. Here we will adopt the former point of view but our approach can easily be extended to the latter as discussed in Sect. 3.5.

2.2 Transition model

We now detail how each node may evolve conditionally on the current state of the network. We denote by Y_i^t the state (0 for susceptible and 1 for infected) of node i at time t. Thus $Y^t = (Y_i^t)_{i=1\dots n}$ summarizes the state of the whole network at time t and $Y = (Y^t)_{t=1\dots m}$ constitutes the whole observed dataset. We assume that the epidemiological process $(Y^t)_t$ is Markovian so its behavior is described by the transitions from Y^t to Y^{t+1}. We further assume that, at each time, nodes evolve independently from one another, conditionally on the preceding state of the network. Hence, we have that

$$\begin{aligned} P(Y \mid T) &= P(Y^1) \prod_{t=1}^{m-1} P(Y^{t+1} \mid Y^t, T^t) \\ &= \prod_{j=1}^{n} P(Y_j^1) \prod_{t=1}^{m-1} \prod_{i,j:[ij] \in T^t} P(Y_j^{t+1} \mid Y_j^t, Y_i^t). \end{aligned} \tag{2}$$

The transition probabilities are given by the terms $P(Y_j^{t+1} \mid Y^t, T^t)$ in (2). First, when infected, a node may stay infected (with probability $1-e$) or become susceptible (with probability e) independently from the other nodes. Second, a susceptible node can only be infected at time t if its parent in T^t is infected. If so, infection occurs with probability c. All these conditional probabilities are summarized in (4).

In the proposed modeling the marginal probability for a susceptible node to get infected depends on the fraction of infected nodes in an implicit manner, through the choice of its parent (which may or may not be infected). A more explicit dependence, such as $P(Y_j^{t+1} = 1 \mid Y_j^t = 0) = 1 - q^{I_t}$ $(I_t$

being the number of infected nodes at time t) cannot be cast in the tree-structured model we propose as it introduces a dependence with respect to the whole population.

3 Inference

3.1 Graphical model

In Sect. 2.1, we saw how the addition of a root vertex Δ turned the infection path into a spanning tree on the set of vertices $V^* = \{\Delta\} \cup \{(i,t) : 1 \leq i \leq n, 1 \leq t \leq m\}$. This property is the result of the Markovian structure of the model with respect to time and of the assumptions made on the infection model. This tree structure allows us to explore the set of possible infection paths at a low computational cost through an interesting algebraic result called the matrix-tree theorem.

Matrix-tree theorem. Let W be a square matrix with entries in $\mathbf{R}^+$ and whose rows and columns are indexed by a finite set U. The matrix W is taken so that, for all $v \in U$, $W_{v,v} = 0$. The general term of the Laplacian matrix Q associated with W is given by

$$Q_{u,v} := \begin{cases} -W_{u,v} & \text{if } u \neq v, \\ \sum_{k \in V} W_{kv} & \text{if } u = v. \end{cases}$$

Theorem 1 (Directed matrix-tree theorem, Chaiken (1982)) *Let C denote the cofactor matrix of Q and $\mathcal{T}_v$, $v \in V$, denote the set of directed trees on U rooted at vertex v. It holds that, for any $v \in U$,*

$$\sum_{T \in \mathcal{T}_v} \prod_{(k,l) \in T} W_{k,l} = C_{v,v}.$$

The definition of the cofactor matrix is given at the beginning of the Appendix. Theorem 1 can be used to sum over all infection paths by taking U equal to V^* and considering the trees rooted at Δ. The Markovian structure of the model with respect to time can be encoded in W as described in Fig. 2, where W^t contains the weights of all possible edges between times t and $t+1$. The row vector W^0 contains the weights of all edges between the root Δ and each node at times $t = 1$, and all its terms are usually taken equal to 1.

The following corollary will enable us to evaluate sums over all possible spanning trees in an efficient manner whenever W has this particular structure.

Corollary 1 *If W has the upper-diagonal block-structure described in Fig. 2, then*

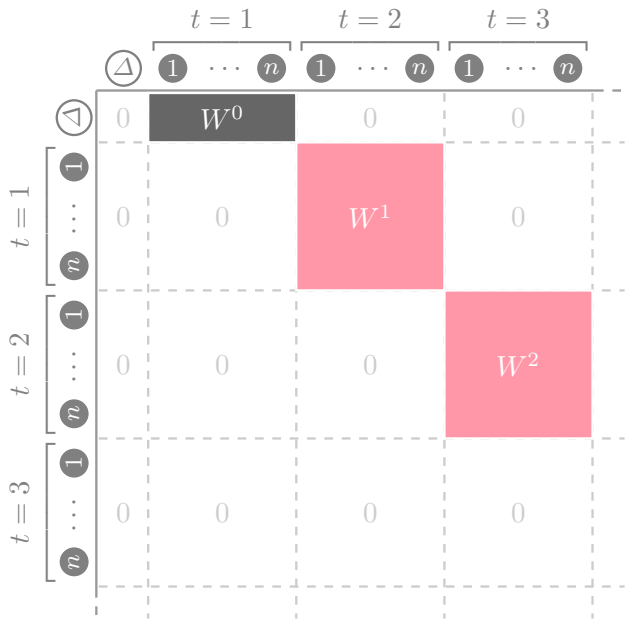

Fig. 2 General structure of matrix W for summing over all directed trees corresponding to the structures of our model

$$\sum_{T \in \mathcal{T}_\Delta} \prod_{(k,l) \in T} W_{k,l} = \prod_j W^0_{\Delta j} \prod_{t=1}^{m-1} \prod_j W^t_{+j}$$

where $W^t_{+j} = \sum_{i:i \neq j} W^t_{ij}$.

Proof This follows from Theorem 1 and from the definition of the Laplacian of W, which has the same structure as W plus nonzero entries on the diagonal. As a consequence, the minor $C_{\Delta\Delta}$ is the determinant of the Laplacian of W deprived of its first row and column. The computation of this determinant boils down to the product of all the diagonal terms of the Laplacian matrix except the first one. □

Complexity. The approach we propose takes advantage of Corollary 1. Indeed, Theorem 1 requires that we compute the determinant of the matrix W, which would result in a computational complexity of order $\mathcal{O}(n^3m^3)$. Since W is upper triangular, the complexity of the computation of the Laplacian matrix and the evaluation of its determinant reduce to $\mathcal{O}(mn^2)$.

3.2 Parameter inference

We consider here the weights β_{ij} to be fixed. Indeed, these weights can be used to introduce prior knowledge on the possibility for each edge to actually be part of the network. In Sect. 3.3, we will show how to compute the probability that an edge is part of the contact network, based on prior information and conditionally on the data. But the β_{ij} could also be

considered as parameters to be inferred and Sect. 3.5 shows that maximum likelihood estimates this can be obtained via a regular EM algorithm.

Regarding the estimation of parameters $r = (c, e)$, we also adopt a maximum likelihood approach.

Note that as the distribution of the data at time 1 does not bring any information about the transmission process we are interested in, we will work conditionally on Y^1, meaning that $P(Y)$ actually stands for $P(Y \mid Y^1)$.

Using Eqs. (1) and (2), we can write the likelihood of the data as

$$\begin{aligned} P(Y) &= \sum_{T \in \mathcal{T}} P(Y \mid T) P(T) \\ &= \sum_{T \in \mathcal{T}} \prod_{t=1}^{m-1} \prod_{[ij] \in T^t} \beta_{ij} \phi_{ij}^t / B, \end{aligned} \tag{3}$$

denoting

$$\phi_{ij}^t := P(Y_j^{t+1} \mid Y_j^t, Y_i^t, [ij] \in T^t).$$

When i is the parent of j in T^t, the following table summarizes all possible values for ϕ_{ij}^t:

ϕ_{ij}^t	$Y_j^{t+1} = 1$	$Y_j^{t+1} = 0$
$Y_j^t = 1\, Y_i^t = 1$	$1 - e$	e
$Y_j^t = 1\, Y_i^t = 0$	$1 - e$	e
$Y_j^t = 0\, Y_i^t = 1$	c	$1 - c$
$Y_j^t = 0\, Y_i^t = 0$	0	1

(4)

An obvious consequence of these transition parameters is that the time during which an infected node stays infected is geometrically distributed.

If we further denote $\psi_{ij}^t = \phi_{ij}^t \beta_{ij}$, the likelihood of the observed data can be written as

$$P(Y) = C / B \tag{5}$$

where $C := \sum_{T \in \mathcal{T}} \prod_t \prod_{[ij] \in T^t} \psi_{ij}^t$. Both B and C can be computed thanks to Corollary 1:

$$\begin{aligned} B &= \prod_{t=1}^{m-1} \prod_j \beta_{+j} = \prod_j (\beta_{+j})^{m-1}, \\ C &= \prod_{t=1}^{m-1} \prod_j \psi_{+j}^t, \end{aligned} \tag{6}$$

where $\beta_{+j} := \sum_i \beta_{ij}$ and $\psi_{+j}^t := \sum_i \psi_{ij}^t$, that is

$$\psi_{+j}^t = \begin{cases} (1-e) \sum_i \beta_{ij} & \text{if } y_j^t = 1,\ y_j^{t+1} = 1, \\ e \sum_i \beta_{ij} & \text{if } y_j^t = 1,\ y_j^{t+1} = 0, \\ c \sum_i \beta_{ij} y_i^t & \text{if } y_j^t = 0,\ y_j^{t+1} = 1, \\ -c \sum_i \beta_{ij} y_i^t + \sum_i \beta_{ij} & \text{if } y_j^t = 0,\ y_j^{t+1} = 0. \end{cases}$$

We now consider the inference of parameters e and c. According to Eq. (5), their maximum likelihood estimates (MLEs) are obtained via the maximization of $\log P(Y) = \log C - \log B$. These parameters are only involved in the term $\log C$, which can be split according to four subsets corresponding to the four possible configurations of $(y_j^t, y_j^{t+1}) \in \{0, 1\}^2$. For any $(a, b) \in \{0, 1\}^2$, we define $\mathcal{M}_{ab} := \{(j, t) : y_j^t = a, y_j^{t+1} = b\}$ and $M_{ab} := |\mathcal{M}_{ab}|$. So we have

$$\begin{aligned} BP(Y) &= \prod_{a,b} \prod_{(j,t) \in \mathcal{M}_{ab}} \psi_{+j}^t(r) \\ &= \left(B_{10} e^{M_{10}}\right) \left(B_{11} (1-e)^{M_{11}}\right) \left(S_{01} c^{M_{01}}\right) \\ &\quad \times \left(\prod_{(j,t) \in \mathcal{M}_{00}} (\beta_{+j} - c\, s_j^t) \right) \end{aligned}$$

where $B_{ab} := \prod_{(j,t) \in \mathcal{M}_{ab}} \beta_{+j}$, $s_j^t := \sum_i y_i^t \beta_{ij}$, and $S_{ab} := \prod_{(j,t) \in \mathcal{M}_{ab}} s_j^t$. The log likelihood is then

$$\begin{aligned} \log P(Y) = M_{10} \log e + M_{11} \log(1-e) + M_{01} \log c \\ + \sum_{(j,t) \in \mathcal{M}_{00}} \log\left(\beta_{+j} - c\, s_j^t\right) + \text{cst}. \end{aligned}$$

The MLE of e is straightforwardly $\widehat{e} = M_{10}/(M_{10} + M_{11})$. The MLE $\widehat{c}$ of c has no closed-form expression but satisfies

$$\sum_{(j,t) \in \mathcal{M}_{00}} \widehat{c}\, s_j^t \Big/ \left(\beta_{+j} - \widehat{c}\, s_j^t\right) = M_{01} \tag{7}$$

which has a unique solution as the left-hand-side term is monotonically increasing in c and is zero for $c = 0$. The MLE $\widehat{c}$ can therefore be easily obtained numerically.

3.3 Conditional edge probability

From a network point of view, we are interested in identifying the edges that are most likely to be part of the contact network. With this goal in mind, we define the two following complementary sets $\mathcal{T}_{ij} := \left\{T \in \mathcal{T} : \exists t, [ij] \in T^t\right\}$ and $\overline{\mathcal{T}}_{ij} := \left\{T \in \mathcal{T} : \forall t, [ij] \notin T^t\right\}$ and the corresponding events $\mathcal{E}_{ij} := \{T \in \mathcal{T}_{ij}\}$ and $\overline{\mathcal{E}}_{ij} := \{T \in \overline{\mathcal{T}}_{ij}\}$, where the latter states that $[ij]$ never appears along the tree T. To assess whether the edge $[ij]$ is part of the network, we want to compute the conditional probability

$$\mathbb{P}(\mathcal{E}_{ij} \mid Y) = 1 - \mathbb{P}(\overline{\mathcal{E}}_{ij} \mid Y).$$

Now we have that

$$\begin{aligned}\mathbb{P}(\overline{\mathcal{E}}_{ij} \mid Y) &= \mathbb{P}(\overline{\mathcal{E}}_{ij}, Y)/P(Y) \\ &= \sum_{T \in \overline{\mathcal{T}}_{ij}} \prod_t \prod_{(k,l) \in T^t} \psi^t_{k\ell} \Big/ \sum_{T \in \mathcal{T}} \prod_t \prod_{(k,l) \in T^t} \psi^t_{k\ell} \\ &= \prod_t \left(\prod_{k \neq j} \psi^t_{+k} \right) (\psi^t_{+j} - \psi^t_{ij}) \Big/ \prod_t \prod_k \psi^t_{+k} \\ &= \prod_t \left(1 - \frac{\psi^t_{ij}}{\psi^t_{+j}} \right).\end{aligned}$$

3.4 Multiple independent waves

We now extend the inference to the case where several independent spreads of the epidemic are observed. In the following, we refer to each spread as a 'wave'. Let us denote by H the number of waves and by $Y^{(1)}, \ldots, Y^{(H)}$ the series of observed infection waves, which are assumed to be i.i.d. Therefore, the likelihood is simply

$$P\left(Y^{(1)}, \ldots Y^{(H)}\right) = \prod_{h=1}^{H} P\left(Y^{(h)}\right) \tag{8}$$

Each log likelihood $\log P\left(Y^{(h)}\right)$ can be evaluated as described in Sect. 3.2. To infer the parameters e and c jointly on all waves, we denote by $M^{(h)}_{ab}$ the statistic M_{ab} (defined in Sect. 3.2) computed on wave $Y^{(h)}$ and by $\mathcal{M}^{(h)}_{ab}$ the corresponding set defined on the same wave. We further define the cumulated statistic $M^{(+)}_{ab} = \sum_{h=1}^{H} M^{(h)}_{ab}$.

Thus, the inference on parameters e and c can be conducted by replacing each M_{ab} with $M^{(+)}_{ab}$ in the previously described procedure. The estimate of e becomes $\widehat{e} = M^{(+)}_{10}/(M^{(+)}_{10} + M^{(+)}_{11})$ and, for the estimation of c, the estimation Eq. (7) becomes

$$\sum_h \sum_{(j,t) \in \mathcal{M}^{(h)}_{00}} \widehat{c}\, s^{t,h}_j \Big/ \left(\beta_{+j} - \widehat{c}\, s^{t,h}_j\right) = M^{(+)}_{01}. \tag{9}$$

Similarly, to compute the conditional edge probabilities, we define the general event $\mathcal{E}^{(+)}_{ij} = \{\exists h, \exists t : [ij] \in T^{t(h)}\}$, were $T^{t(h)}$ stands for t-th part of the tree $T^{(h)}$ along which the wave $Y^{(h)}$ takes place. Similarly, we define the complementary event $\overline{\mathcal{E}}^{(+)}_{ij} = \cap_h \overline{\mathcal{E}}^{(h)}_{ij}$ where $\overline{\mathcal{E}}^{(h)}_{ij}$ is the event $\overline{\mathcal{E}}_{ij}$ (defined in Sect. 3.3) for wave h. Because the waves are independent, so are the events $\overline{\mathcal{E}}^{(h)}_{ij}$ across waves. So we have that $\mathbb{P}\left(\overline{\mathcal{E}}^{(+)}_{ij}\right) = \prod_h \mathbb{P}\left(\overline{\mathcal{E}}^{(h)}_{ij}\right) = \mathbb{P}\left(\overline{\mathcal{E}}_{ij}\right)^H$,

$$P\left(Y^{(1)} \ldots Y^{(H)} \mid \overline{\mathcal{E}}^{(+)}_{ij}\right) = \prod_h P\left(Y^{(h)} \mid \overline{\mathcal{E}}^{(h)}_{ij}\right)$$

and

$$\mathbb{P}\left(\overline{\mathcal{E}}^{(+)}_{ij} \mid Y\right) = \frac{\mathbb{P}\left(\overline{\mathcal{E}}^{(+)}_{ij}\right)}{P(Y)} P\left(Y^{(1)} \ldots Y^{(H)} \mid \overline{\mathcal{E}}^{(+)}_{ij}\right)$$

and the conditional probability of an edge given the complete set of waves is $\mathbb{P}\left(\mathcal{E}^{(+)}_{ij}\right) = 1 - \mathbb{P}\left(\overline{\mathcal{E}}^{(+)}_{ij}\right)$.

Note that the waves can involve different set of nodes. However, the power of the statistical inference will only be improved for pairs of nodes which are present in several waves.

3.5 Alternative estimation procedures

Other strategies for the inference of the proposed model can be considered.

Bayesian inference. Parameters c and e can be inferred with a Bayesian approach. When using conjugate priors, the posterior distribution of both parameters can be established easily (see Appendix A.1 The posterior distribution of other quantities of interest, such as edge probabilities, can then be obtained via Monte Carlo sampling from the posterior distribution of parameters.

EM algorithm The proposed model can be seen as a mixture model with as many components as possible trees. In this setting, the weights β_{ij} act as the parameters ruling the proportions of the mixture components and their maximum likelihood estimates can be obtained via the EM algorithm Dempster et al. (1977). The update formulas are given in Appendix A.2. One interest of this approach is that it allows the estimation the weights of the edges β_{ij}, rather than keeping them fixed at a prescribed value.

In practice, none of these alternatives turned out to significantly improve edge retrieval.

4 Simulation study

We designed a simulation study to illustrate the ability of the proposed method to infer the edges of the contact network. We wanted to compare the accuracy of the inference under different network topologies, different settings for the infection parameters, and under both a unique long wave and several short waves.

Simulation settings Data were simulated according to the proposed model. For each dataset, a contact network

was drawn determining the edges that could be used in the simulated infection path. To consider different network topologies, we chose two random graph models and different densities (the proportion of actual edges of a network over the set of total possible edges). Networks were generated from the Erdős–Rényi (ER) and preferential attachment (PA) models. In the ER model (Erdős and Rényi 1959), the status of each edge in the network is drawn randomly and independently, the probability of inclusion is equal to the desired density. In the PA model (Barabási and Albert 1999), the network is constructed sequentially. A new node is attached to the previous network at a location randomly selected with probabilities proportional the degrees of former nodes. We chose these two generative models since they lead to contrasted structures. The ER model leads to networks well balanced with homogeneous degrees while the PA model leads to networks characterized by having few nodes with high degrees (hubs). In the simulations, we only considered undirected networks in which an edge between a pair of nodes indifferently makes any of the nodes the parent and the other the offspring. When performing the inference, we may or may not enforce the symmetry of edge probabilities.

For each topology, the density of the network varies in $d \in \{0.1, 0.2, 0.4\}$. Waves were initialized with a random set of nodes in state 1 and all the others set to 0. We compared the results obtained from one long wave of $m = 200$ time steps with the ones resulting from the observation of $H = 10$ waves of 20 time steps. We only considered simulations where the epidemic was active throughout the considered time period. Whenever the epidemic died out, we re-simulated the wave. We set the extinction probability to $e = 0.05$ and the infection probability varies in $c \in \{0.1, 0.2, 0.4\}$. We chose the parameter ranges so that the total extinction of the epidemic was rather unlikely. For the inference, all the weights β_{ij} were set to 1.

Results The conditional edge probabilities were computed as in Sect. 3.3 for uniwave simulations and as in Sect. 3.4 for multiwave simulations. The AUC were then computed by comparing the conditional edge probabilities to the actual edges of the simulated contact network. Figure 3 displays the boxplots of the AUC obtained over 100 simulations under each of the different settings.

A general remark that can be made is that all AUC are greater that .5, meaning that it is possible to recover the edges based only on the observation of nodes status along time.

Enforcing the symmetry improved the results, especially for large values of c. In all the settings, the multiwave simulations lead to a more accurate inference. This is also more contrasted when c is larger. This can be explained by the fact that the first time steps after the introduction of the "epidemic" (a node with state 1) are the more informative ones. Indeed, the possible edges used to propagate the epidemic are restricted to those which link the few already infected nodes with the new infected ones. We can also notice that the inference is easier in the case where the network is simulated from an ER model and when it is not too dense. The results on the inference of parameters e and c are provided in Appendix 1.

5 Illustration

5.1 Seed influence in Telangana region

Context Telangana is a region known for agrarian crisis in the form of pesticide overuse, seed uncertainty, and farmer suicide (Galab et al. 2009; Gutierrez et al. 2015; Stone 2007). Since 2002, more than 1,200 new genetically modified (GM) cotton seed brands have been released in India, with some achieving a robust but fleeting popularity (Stone et al. 2014). The causes driving these rises and falls in farmer seed certainty are not clear, but the ecological and socioeconomic stakes are high. Seed choices are a path-dependent decision (David 2007) in agriculture because a seed choice cannot be modified or re-selected once it has been sown. Scholarship in the spread of agricultural technologies emphasizes environmental learning, in which individuals try technologies and learn from the results (Griliches 1980; Herring and Rao 2012), social learning, in which individuals emulate others (Boyd et al. 2011; Henrich 2001), and didactic learning, in which individuals are influenced by larger institutions (Stone 2016). To understand how information flows across a community such as farmers learning which seeds to plant, it is first necessary to describe and model this community.

Data The data used for this dynamic model were collected by Andrew Flachs during a study of farmer decision-making in Telangana, India (Flachs et al. 2017). The data were drawn from annual surveys of seed choice in two Telangana villages where farmers grow commercial varieties of rice and hybrid genetically modified cotton. Farmers were asked to report on seeds sown in their fields, and Flachs collected other demographic and agricultural information pertaining to decision-making, including crop yields, spatial relationships of farmers, age, caste, and education. Because each of these socioeconomic and agronomic factors may influence how farmers choose seeds and the larger impact of GM crops in the developing world, it is important to model how influence travels across this network of farmers.

In order to have dense enough data, we focused on the $m = 3$ last years (2012 to 2014) where more data on seed choices were collected, on seeds grown at least 10 times over the three

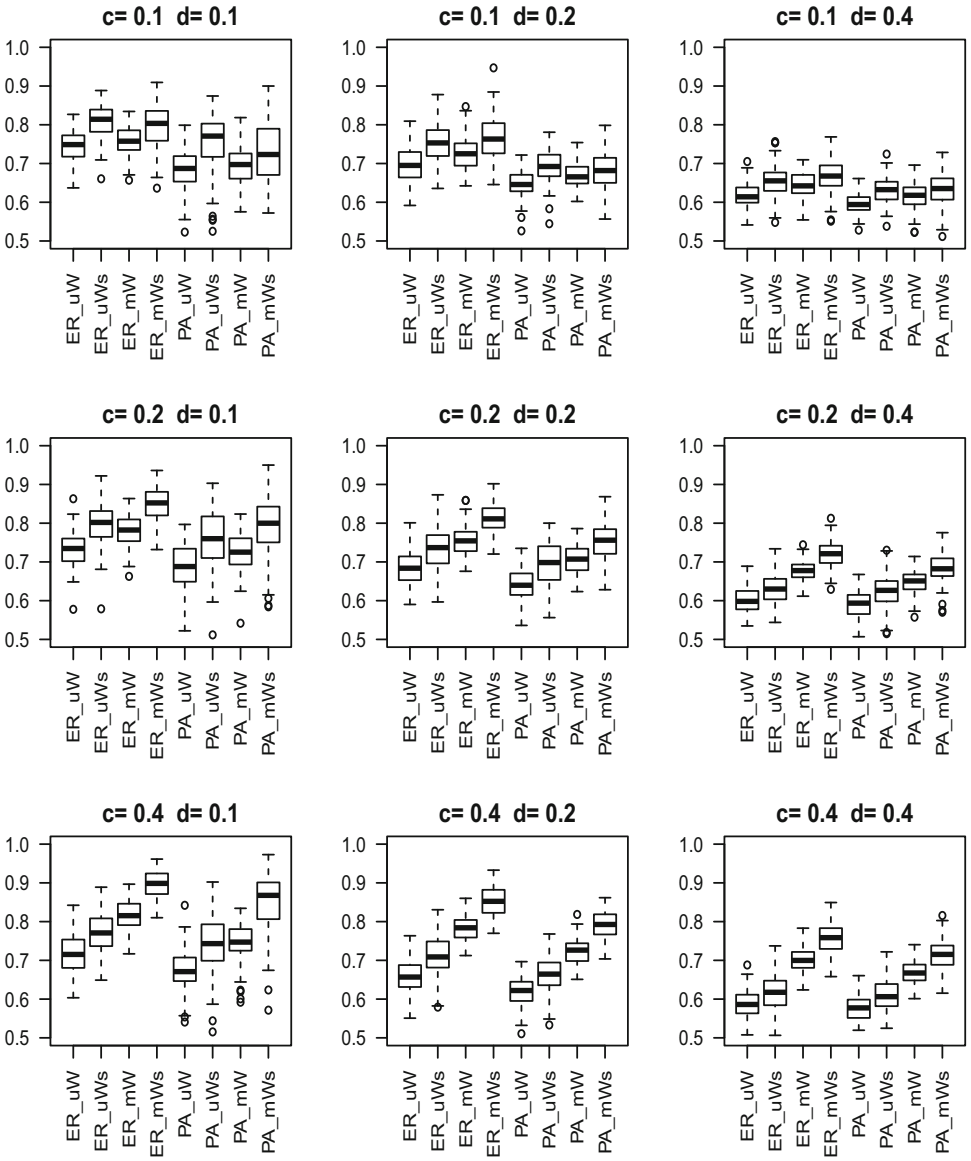

Fig. 3 Boxplots of areas under curve (AUC) computed for the conditional edge probabilities. Each boxplot is obtained from 100 simulations. In the simulations, the extinction and infection probabilities were set to $e = 0.05$ and $c \in \{0.1, 0.2, 0.4\}$. The number of nodes in the network was 20, networks were simulated with Erdős–Rényi (ER) and preferential attachment (PA) topologies, and the density of the network was set to $d \in \{0.1, 0.2, 0.4\}$. A simulation was always initialized with a unique node in state 1. A uniwave simulation (uW) consists of a wave of $m = 200$ time steps. A multiwave simulation (mW) consists of $H = 10$ waves of $m = 20$ time steps. Edge probabilities were inferred by enforcing or not enforcing the symmetry of the network (s on the x-axis means inference with symmetry enforced). For example, `PA_uWs` corresponds to a uniwave simulation over a preferential attachment network with symmetric edge inference

years by all the farmers and on seeds which are at least present two consecutive years. It resulted in data concerning $n = 127$ farmers growing $H = 14$ different seeds. The farmers were considered as nodes of an unknown network that we tried to infer from their seed choices. For each seed, we created a $n \times m$ matrix with $0 - 1$ entries, a 1 entry indicating that the corresponding farmer grows the seed the corresponding year, a 0 entry indicating that he/she does not. The discrete time model is well suited to these data since the seed choices are made for a whole season and new choices will only happen the following year. We considered the H seed diffusions as independent waves which is a strong assumption. However, since the farmers can grow more that one seed per year, it is not unrealistic to consider that between two years they may be influenced by several farmers concerning different seeds. Moreover, the waves were still related in the sense they share the same underlying network. We also chose to consider the parameters e and c as constant since the number of observations was limited. They can be seen as accounting for general trends in the farmer community regarding their willingness to try new seeds and to be influenced by others rather than by a seed's specific attribute. We computed the conditional probabilities for oriented edges as influence is an asymmetric phenomenon. Besides the seed data, the caste, village, acreage (agricultural exploited surface) and age of the farmers were available.

Results Transition parameters were estimated to $e = 0.69$ and $c = 0.88$. The high value for e shows that farmers are prone to change the seeds they grow often while the high value for c shows a rapid diffusion of seed choices. We chose to represent the results about the contact network by select-

ing the most probable edge for each farmer. The resulting network is displayed in Fig. 4. We favored this representation over a network obtained by thresholding the values of the computed conditional edge probabilities because the latter would have been hardly readable. We mainly noticed that some nodes seem to have a great influence, while some others are organized in small groups. We studied the links between the inferred probabilities of edges and the covariates. We compared the computed conditional edge probabilities between pairs of nodes within the same caste or village and between pairs of nodes from two different castes or villages by t tests for mean comparison. We found that edges within the same caste and the same village were more likely to occur (P values, respectively, 2.4×10^{-13} and 2.5×10^{-8}). This shows that farmers from the same castes and villages are more likely to influence one another than farmers that are not. Agricultural learning is often a highly localized process, where results from one area may not easily translate to results in another. Fields may differ by fertility, access to water, pest susceptibility, and a number of other factors. Furthermore, studies of social learning show that important and useful information may not cross key social dividing lines (Henrich 2001), such as caste, when members of a network must learn from others in different levels of a social hierarchy. Thus it is not surprising that farmers near in space and social status are likely to influence one another more than farmers of different castes and in different villages. The correlation between the differences of age (age of influencer minus age of influenced farmer) and the computed conditional edge probabilities was found significantly negative (P value 2.5×10^{-3}) which indicates that farmers tend to be influenced by farmers younger than they are. No significant correlation was found between the difference of total acreages exploited by the farmers (acreage of influencer minus acreage of influenced farmer) and the conditional edge probabilities.

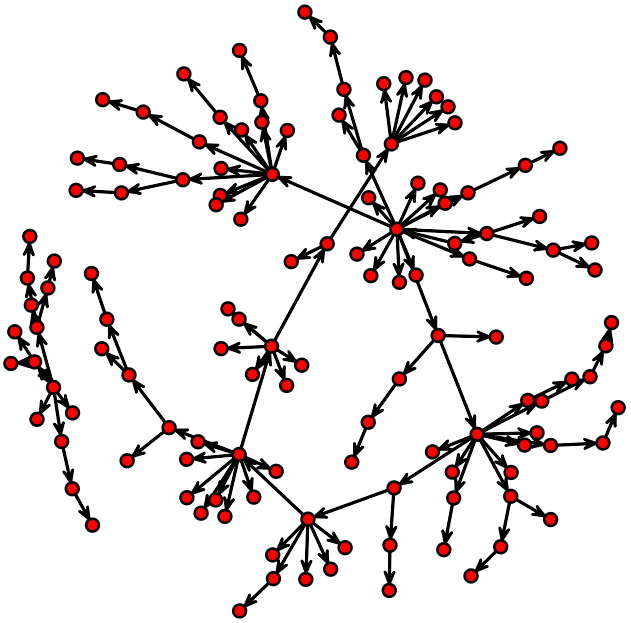

Fig. 4 Network of seed choice influences between farmers. This network was obtained by taking the most probable neighbor for each farmer from the inferred edge probabilities

5.2 Measles epidemic data

Dataset The Hagelloch dataset collected by Pfeilsticker (1863) and augmented by Oesterle (1993) is a detailed dataset on the 1861 measles outbreak in Hagelloch, Germany. Because of its completeness, it has been analyzed in several papers dealing with the inference of epidemic models (to cite but a few Neal and Roberts 2004; Groendyke et al. 2011, 2012). This dataset includes the dates of the prodromes, rash eruption and death (whenever it occurred) for each of the 188 children infected by the measles. In addition to these data on the epidemic, other information concerning the children such as their household with space location, school class, family, age, and gender are available. In order to fit our model, we transformed the data into a 46×187 matrix. One child was removed since he/she was infected long after the others (40 days after the last one infected). We considered an infectious period starting one day before the prodromes and ending 3 days after the rash or at the date of death whenever it occurred. These choices were guided by Neal and Roberts (2004). Since our model is SIS, at the end of the infectious period the children were set back to a susceptible state although a removed state would have been more relevant. We enforced symmetry for the computation of conditional edge probabilities since the contact network accounted for the interactions between children.

Results The transition parameters were estimated to $e = 0.11$ and $c = 0.18$. However, these estimates do not convey much information since we know that the SIS model is not totally suitable for these data. Nevertheless, we could still obtain the conditional edge probabilities which were our main interest. Figure 5 displays the quantiles of the computed conditional edge probabilities with respect to the delay between the starting dates of the infectious periods. For pairs of nodes with a delay of 1 to 9 days, the computed probabilities were among the highest. The edges between these pairs of nodes may then correspond to the edges of the contact network which are likely to have been part of the propagation path of the epidemic. To relate the conditional edges probabilities to the other data available on the children we used standard statistical tests: t tests for mean comparison to test whether the means of the computed edge probabilities were significantly higher or lower between children within

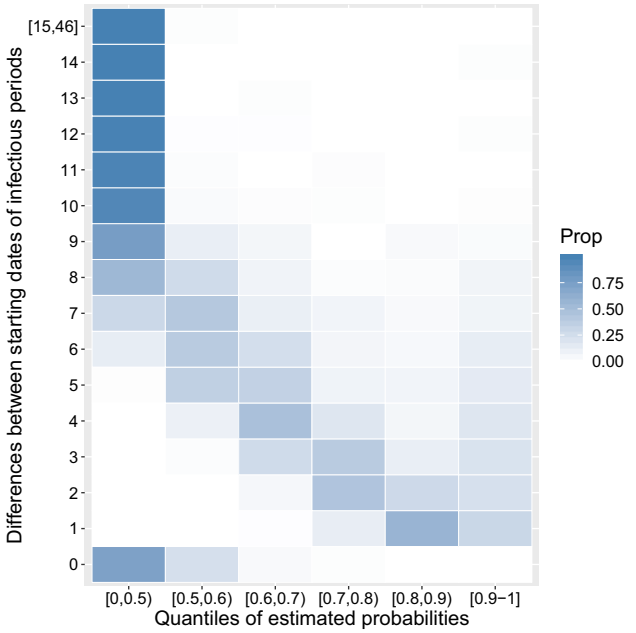

Fig. 5 Distribution of the computed conditional edge probabilities with respect to the delay between the starting dates of the infectious periods. The proportions are constrained to sum to one by row

Table 1 Relationships between covariates on children and computed conditional edge probabilities. NS means nonsignificant

Covariates	Effect	P value
distance	−	2.2×10^{-3}
age	−	$< 2.2 \times 10^{-16}$
family	+	1.2×10^{-2}
house	+	3.8×10^{-6}
sex	NS	4.4×10^{-1}
classroom	+	$< 2.2 \times 10^{-16}$

the same family, house, sex, classroom than between children from different ones, and correlation tests between the computed probabilities and the distances between the houses where the children live, and between the computed probabilities and the difference in ages of children (in absolute value). The sign of the effect and the P values are given in Table 1. These results are quite consistent with what was found in Neal and Roberts (2004) and in Groendyke et al. (2012), especially for the house and classroom effects.

6 Conclusion

We introduced a model for the spread of an epidemic and showed that the underlying social contact structure can be inferred from the observation of individual statuses along time. To perform the inference, we observed that the path of the epidemic can be assumed to be tree-shaped and we resorted to the matrix-tree theorem to average over all possible tree-structured paths. To our knowledge, this is one of the few uses of the matrix-tree theorem for the statistical inference of a dynamic model.

An important feature of the proposed approach is that the matrix-tree theorem gives rise to an inference procedure with very low computational complexity. This low complexity results from the combination of the Markov structure of the dynamic model with the tree structure of the infection path, which is summarized in Eq. (2). The marginal log likelihood of any model satisfying both properties will take the form

$$\prod_t \prod_j \psi_{+j}^t \Big/ \prod_t \prod_j \beta_{+j}^t$$

where $\psi_{ij}^t = \beta_{ij}^t \phi_{ij}^t$ (see Eqs. (3), (5) and (6)).

The model defined in Sect. 2 implies a geometric duration for contamination, but the second application shows that a fixed contamination time can also be considered. The proposed approach can be extended to more complex propagation models, the propagation rules being encoded in the terms ϕ_{ij}^t. More that two levels of infection ('sick' or 'healthy') can be considered: when the contamination duration is known, the model can be extended to an SIR model, by simply adding a 'recovered' state. From a general point of view, the model can be extended to an arbitrary number of states (for example, 'susceptible', 'incubating', 'contaminated', 'recovered') provided that all states are observed, which is not always the case. This generalization also allows for the joint modeling of several diseases (or seeds), using H binary states (H being the number of diseases) indicating which disease affects each node at each time. This representation allows us to model non-independent spreads, at the price of a larger number of parameters. In the same vein, the effect of environmental factors could also be accounted for via a regression term in the transition rates encoded in the ϕ_{ij}^t.
The difficulty of the parameter inference will mostly depend on the expression of ϕ_{ij}^t, but the complexity of the network reconstruction will remain the same and will still benefit from the computational efficiency achieved through the matrix-tree theorem.

In this work, we did not consider any additional information on contacts, which led to the choice $\beta_{ij} = 1$ for any pair (i, j). However, if some information is available, it could be encoded through particular choices of β_{ij}. For instance, if information on distances between individuals is available, a parametric form for β_{ij} linking the probability of contact to the distance could be assumed.

Acknowledgements The authors sincerely thank the Associate Editor and the two Reviewers for their careful reading, their comments and their advice, which obviously contributed to improve this manuscript.

A Appendix

Definition 1 (*Cofactor matrix*) Consider a $p \times p$ square matrix A. For any pair $1 \leq u, v \leq p$, the cofactor $C_{u,v}$ is defined as $C_{u,v} := (-1)^{u+v}|A^{u,v}|$, where $A^{u,v}$ stands for the matrix A deprived from its u-th row and v-th column. The cofactor matrix of A is the $p \times p$ square matrix C, with general term $C_{u,v}$.

A.1 Bayes inference for the parameters

Assume that parameters e and c have independent Beta prior distributions:

$$r = (c, e) \sim \text{Beta}(\beta_e, \beta'_e) \otimes \text{Beta}(\beta_c, \beta'_c). \tag{10}$$

In a Bayesian setting, the likelihood given in (5) corresponds to the conditional distribution $P(Y|r)$ so that the joint distribution of (Y, r) is given by

$$\begin{aligned} B\ P(Y, r) = &\frac{B_{10}B_{11}}{\text{B}(\beta_e, \beta'_e)} e^{M_{10}+\beta_e}(1-e)^{M_{11}+\beta'_e} \\ &\times \frac{S_{01}}{\text{B}(\beta_c, \beta'_c)} c^{M_{01}+\beta_c}(1-c)^{\beta'_c} \\ &\times \prod_{(j,t)\in\mathcal{M}_{00}} \left(\beta^t_{+j} - c \cdot s^t_j\right). \end{aligned}$$

The marginal likelihood of the data Y is obtained by integrating over c and e to get

$$\begin{aligned} B\ P(Y) = &\frac{B_{10}B_{11}}{\text{B}(\beta_e, \beta'_e)} \text{B}(M_{10} + \beta_e, M_{11} + \beta'_e) \\ &\times \frac{S_{01}}{\text{B}(\beta_c, \beta'_c)} \int c^{M_{01}+\beta_c}(1-c)^{\beta'_c} \\ &\times \prod_{(j,t)\in\mathcal{M}_{00}} \left(\beta^t_{+j} - c \cdot s^t_j\right) \text{d}c. \end{aligned}$$

Both $P(Y, r)$ and $P(Y)$ factorize with respect to e and c, e and c are therefore conditionally independent given Y. The posterior distribution of e is simply Beta:

$$P(e|Y) = \text{Beta}(\beta_e + M_{10}, \beta'_e + M_{11}) \tag{11}$$

whereas the posterior distribution of c does not have a closed-form expression:

$$P(c|Y) \propto c^{M_{01}+\beta_c}(1-c)^{\beta'_c} \prod_{(j,t)\in\mathcal{M}_{00}} \left(\beta^t_{+j} - c \cdot s^t_j\right) \tag{12}$$

but can be easily sampled via Monte Carlo or importance sampling, or numerically integrated. Observe that, in this setting, the weights β_{ij} act as hyper-parameters.

A.2 EM inference

If T is considered as a latent variable, the proposed model becomes an incomplete data model, for which maximum likelihood inference can be carried out via the EM algorithm (Dempster et al. 1977). Because the distribution of T is parameterized by β, the parameters to be inferred become $\theta = (\beta, c, e)$. To use the EM algorithm, we first need to write the complete log likelihood:

$$\begin{aligned} \log P_\theta(Y, T) &= \sum_t \log P_\theta(T^t) + \log P_\theta(Y^{t+1}|Y^t, T^t) \\ &= \sum_{i,j} N_{ij}(T) \log \beta_{ij} + \sum_{i,j} \sum_{t:T^t\ni(i,j)} \log \phi^t_{ij} - \log B \end{aligned}$$

where $N_{ij}(T) = \sum_t \mathbb{I}\{(i, j) \in T^t\}$. Then, we need the conditional expectation of this complete log likelihood $\mathbb{E}_{\theta^q}\left[\log P_\theta(Y, T)|Y\right]$, that is

$$\begin{aligned} &\sum_T P_{\theta^q}(T|Y) \log P_\theta(Y, T) \\ &= \sum_{i,j} \mathbb{E}_{\theta^q}[N_{ij}(T)|Y] \log \beta_{ij} \\ &\quad + \sum_{i,j,t} P_{\theta^q}\{(i, j) \in T^t|Y\} \log \phi^t_{ij} - \log B \end{aligned}$$

E step The E step requires to compute the conditional edge probability

$$P_\theta\{(i,j) \in T^t|Y\} = 1 - P_\theta\{(i,j) \notin T^t|Y\} = 1 - C^t_{ij}/C$$

where C^t_{ij} is computed in the same way as C in Eq. (6), setting the term ψ^t_{ij} to 0. As a consequence, we get $C^t_{ij}/C = \left(\psi^t_{+j} - \psi^t_{ij}\right)/\psi^t_{+j}$, so

$$P_\theta\{(i,j) \in T^t|Y\} = \psi^t_{ij}/\psi^t_{+j}.$$

This provides us with the conditional expected counts:

$$\mathbb{E}_\theta[N_{ij}(T)|Y] = \sum_t P_\theta\{(i,j) \in T^t|Y\} = \sum_t \psi^t_{ij}/\psi^t_{+j}.$$

M step The parameter estimates are updated by maximizing the conditional expectation $Q(\theta|\theta^q) := \mathbb{E}_{\theta^q}\left[\log P_\theta(Y,T)|Y\right]$. The terms of $Q(\theta|\theta^q)$ depending on e are

$$M_{11}\log(1-e) + M_{10}\log e$$

because, for all j and t, $\sum_i P_{\theta^q}\{(i,j) \in T^t|Y\} = 1$. (Each node has one and only one parent at each time.) Notice that the involved quantities do not depend on h, so we get the same estimate as in Sect. 3.2: $\widehat{e} = M_{10}/(M_{10}+M_{11})$.
The terms of $Q(\theta|\theta^q)$ depending on c are

$$E^q_{101}\log c + E^q_{100}\log(1-c),$$

where $E^q_{101} = \sum_{i,j,t} P_{\theta^q}\{(i,j) \in T^t|Y\}Y^t_i(1-Y^t_j)Y^{t+1}_j$ and $E^q_{100} = \sum_{i,j,t} P_{\theta^q}\{(i,j) \in T^t|Y\}Y^t_i(1-Y^t_j)(1-Y^{t+1}_j)$, so at iteration q, the estimate of c is updated to

$$\begin{aligned} c^{q+1} &= \arg\max_c \mathbb{E}_{\theta^q}\left[\log P_\theta(Y,T)|Y\right] \\ &= E^q_{101}/(E^q_{100}+E^q_{101}). \end{aligned}$$

Then we have to maximize $Q(\theta|\theta^q)$ w.r.t. β, which is equivalent to maximizing $\mathbb{E}_{\theta^q}[N_{ij}(T)|Y]\log\beta_{ij} - \log B$, where $\sum_{i,j}\mathbb{E}_{\theta^q}[N_{ij}(T)|Y]\log\beta_{ij} - m\sum_j \log\beta_{+j}$ is maximal for

$$\frac{\beta^{q+1}_{ij}}{\beta^{q+1}_{+j}} = \frac{\beta^{q+1}_{ij}}{\beta^{q+1}_{ij} + \sum_{u\neq i}\beta^{q+1}_{uj}} = \frac{\mathbb{E}_{\theta^q}[N_{ij}(T)|Y]}{m},$$

which leads to the update formula

$$\beta^{q+1}_{ij} = \frac{\mathbb{E}_{\theta^q}[N_{ij}(T)|Y]}{m - \mathbb{E}_{\theta^q}[N_{ij}(T)|Y]}\sum_{u\neq i}\beta^q_{uj}.$$

Estimates of the parameters *e* and *c* for the simulation study

Figures 6 and 7 display, respectively, the results on the estimates of the parameters e and c. The inference for e is mainly satisfactory except for some configurations where this parameter is underestimated in the multiwave setting. This appears in configurations with the lowest value for c ($c = 0.1$). In these configurations, a total extinction of the epidemic may happen (the lower c, the more likely the total extinction). Then the underestimation of e seems to be a consequence of our simulation choice which enforces that no total extinction occurs in any of the waves. This also leads to overestimate c for the very same configurations while for other configurations, the estimations of c suffers from a negative bias. This may be due to the fact that the edges were sampled in an underlying network instead among all possible pairs of nodes. Nevertheless, these estimates are still satisfactory since they do not prevent us from recovering the edges which is our main goal.

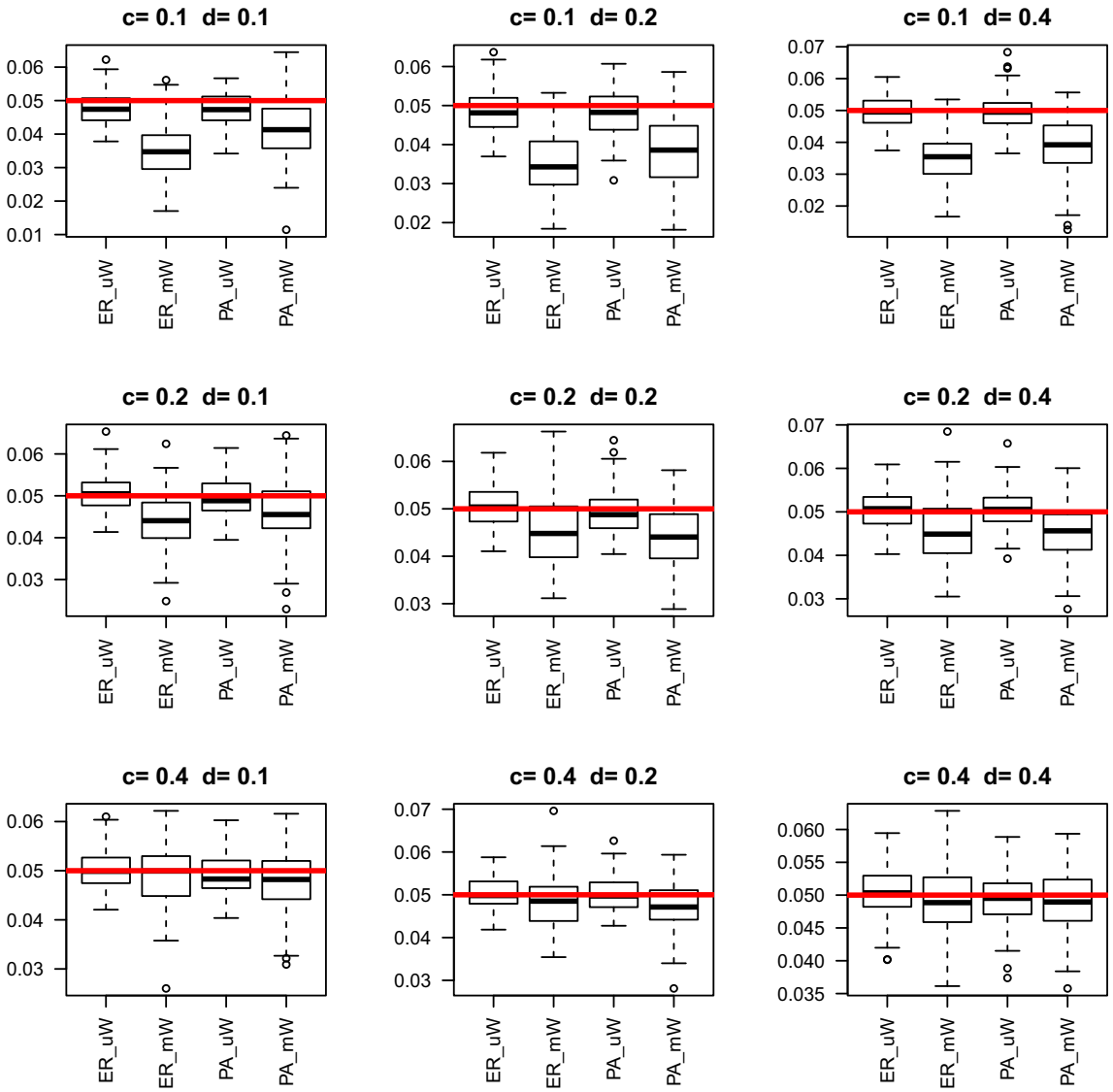

Fig. 6 Boxplots of estimates of e. Each boxplot is obtained from 100 simulations. In the simulations, the extinction and infection parameters were set to $e = 0.05$ and $c \in \{0.1, 0.2, 0.4\}$. The number of nodes in the network was 20, networks were simulated with Erdős–Rényi (ER) and preferential attachment (PA) topologies, and the density of the network was set to $d \in \{0.1, 0.2, 0.4\}$. A simulation was always initialized with a unique node in state 1. A uniwave simulation (uW) consists in a wave of $m = 200$ time steps. A multiwave simulation (mW) consists in $H = 10$ waves of $m = 20$ time steps. For example, `PA_uW` corresponds to a uniwave simulation over a preferential attachment network

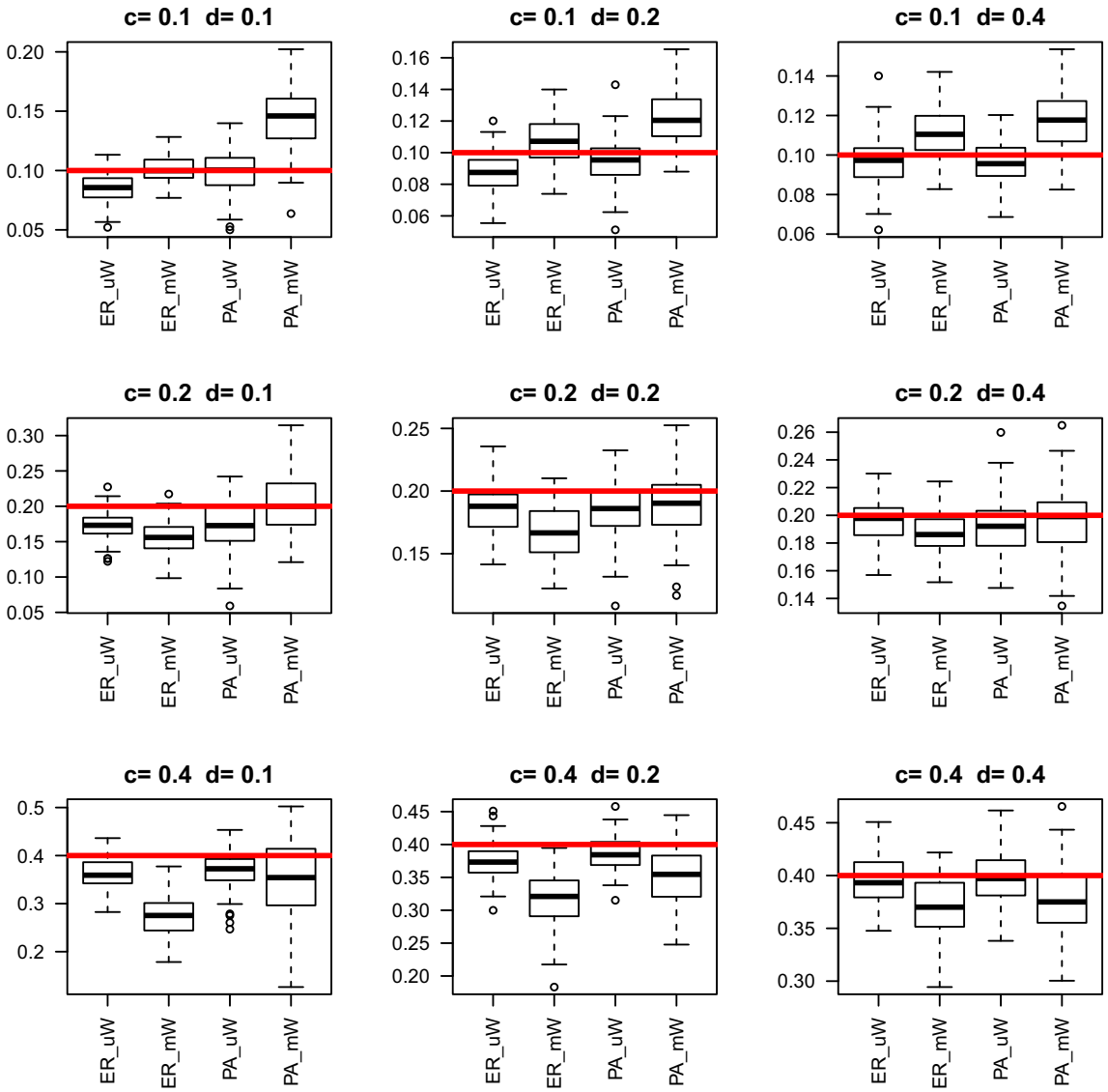

Fig. 7 Boxplots of estimates of c. Each boxplot is obtained from 100 simulations. In the simulations, the extinction and infection parameters were set to $e = 0.05$ and $c \in \{0.1, 0.2, 0.4\}$. The number of nodes in the network was 20, networks were simulated with Erdős–Rényi (ER) and preferential attachment (PA) topologies, and the density of the network was set to $d \in \{0.1, 0.2, 0.4\}$. A simulation was always initialized with a unique node in state 1. A uniwave simulation (uW) consists in a wave of $m = 200$ time steps. A multiwave simulation (mW) consists in $H = 10$ waves of $m = 20$ time steps. For example, `PA_uW` corresponds to a uniwave simulation over a preferential attachment network

References

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

Barbillon, P., Thomas, M., Goldringer, I., Hospital, F., Robin, S.: Network impact on persistence in a finite population dynamic diffusion model: application to an emergent seed exchange network. J. Theor. Biol. **365**, 365–376 (2015)

Boyd, R., Richerson, P.J., Henrich, J.: The cultural niche: why social learning is essential for human adaptation. Proc. Natl. Acad. Sci. **108**(Supplement 2), 10918–10925 (2011)

Brauer, F., Castillo-Chavez, C., Castillo-Chavez, C.: Mathematical Models in Population Biology and Epidemiology, vol. 40. Springer, Berlin (2012)

Britton, T., O'Neill, P.D.: Bayesian inference for stochastic epidemics in populations with random social structure. Scand. J. Stat. **29**(3), 375–390 (2002)

Chaiken, S.: A combinatorial proof of the all minors matrix tree theorem. SIAM J. Algebr. Discrete Methods **3**(3), 319–329 (1982)

Chow, C., Liu, C.: Approximating discrete probability distributions with dependence trees. IEEE Trans. Inf. Theory **IT–14**(3), 462–467 (1968)

David, P.A.: Path dependence: a foundational concept for historical social science. Cliometrica **1**(2), 91–114 (2007)

Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the EM algorithm. J. R. Stat. Soc. B. **39**, 1–38 (1977)

Erdős, P., Rényi, A.: On random graphs. I. Publicationes Mathematicae Debrecen. **6**, 290–297 (1959)

Flachs, A., Stone, G.D., Shaffer, C.: Mapping knowledge: gis as a tool for spatial modeling of patterns of warangal cotton seed popularity and farmer decision-making. Hum. Ecol. **45**(2), 143–159 (2017)

Galab, S., Revathi, E., Reddy, P.P.: Farmers' suicides and unfolding agrarian crisis in Andhra Pradesh. Agrar. Crisis India. 164–198 (2009). https://doi.org/10.1093/acprof:oso/9780198069096.001.0001

Gomez-Rodriguez, M., Balduzzi, D. and Schölkopf, B.: Uncovering the temporal dynamics of diffusion networks. (2011) Technical Report, arXiv:1508.00286

Gomez-Rodriguez, M., Leskovec, J., Krause, A.: Inferring networks of diffusion and influence. ACM Trans. Knowl. Discov. Data (TKDD) **5**(4), 21 (2012)

Griliches, Z.: Hybrid corn revisited: a reply. Econom. J. Econ. Soc. **48**, 1463–1465 (1980)

Groendyke, C., Welch, D., Hunter, D.R.: Bayesian inference for contact networks given epidemic data. Scand. J. Stat. **38**(3), 600–616 (2011)

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

Gutierrez, A.P., Ponti, L., Herren, H.R., Baumgärtner, J., Kenmore, P.E.: Deconstructing Indian cotton: weather, yields, and suicides. Environ. Sci. Eur. **27**(1), 12 (2015)

Henrich, J.: Cultural transmission and the diffusion of innovations: adoption dynamics indicate that biased cultural transmission is the predominate force in behavioral change. Am. Anthropol. **103**(4), 992–1013 (2001)

Herring, R.J., Rao, N.C.: On the'failure of bt cotton': analysing a decade of experience. Econ. Polit. Wkly. **47**, 45–53 (2012)

Kirshner, S.: Learning With Tree-averaged Densities and Distributions. In: NIPS, pp. 761–768 (2007)

Meilă, M., Jaakkola, T.: Tractable Bayesian learning of tree belief networks. Stat. Comput. **16**(1), 77–92 (2006)

Myers, S., Leskovec, J.: On the convexity of latent social network inference. In: Proceedings of the 23rd International Conference on Neural Information Processing System, vol. 2, pp. 1741–1749 (2010). http://papers.nips.cc/paper/4113-on-the-convexity-of-latent-social-network-inference

Neal, P.J., Roberts, G.O.: Statistical inference and model selection for the 1861 hagelloch measles epidemic. Biostatistics **5**(2), 249–261 (2004)

Oesterle, H.: Statistische Reanalyse einer Masernepidemie 1861 in Hagelloch. Ph.D. thesis, uitgever niet vastgesteld (1993)

Pfeilsticker, A.: Beiträge zur Pathologie der Masern mit besonderer Berücksichtigung der statistischen Verhältnisse

Ray, J., Marzouk, Y.M.: A Bayesian method for inferring transmission chains in a partially observed epidemic. In: Proceedings of the Joint Statistical Meeting (2008)

Stone, G.D.: Agricultural deskilling and the spread of genetically modified cotton in warangal. Curr. Anthropol. **48**(1), 67–103 (2007)

Stone, G.D.: Towards a general theory of agricultural knowledge production: environmental, social, and didactic learning. Cult. Agric. Food Environ. **38**(1), 5–17 (2016)

Stone, G.D., Flachs, A., Diepenbrock, C.: Rhythms of the herd: long term dynamics in seed choice by Indian farmers. Technol. Soc. **36**, 26–38 (2014)

Wang, Y., Chakrabarti, D., Wang, C. and Faloutsos, C.: Epidemic spreading in real networks: an eigenvalue viewpoint. In: 22nd International Symposium on Reliable Distributed Systems, 2003. Proceedings, pp. 25–34. IEEE (2003)

Welch, D., Bansal, S., Hunter, D.R.: Statistical inference to advance network models in epidemiology. Epidemics **3**(1), 38–45 (2011)

Yang, L.-X., Draief, M., Yang, X.: The impact of the network topology on the viral prevalence: a node-based approach. PLoS ONE **10**(7), e0134507 (2015)