Using Almost-Dynamic Bayesian Networks to Represent Uncertainty in Complex Epidemiological Models: a Proposal

Utilizzo di una Rete Bayesiana Quasi-Dinamica per Rappresentare l'Incertezza in Modelli Epidemiologici Complessi: una Proposta

Sabina Marchetti

Sommario We introduce a dynamic model to deal with uncertainty in complex epidemiological processes. Our proposal is based on the Dynamic Bayesian Networks formalism, where each node is associated a random variable, whose value specifies the state of an individual from a given population.
Sommario *Si introduce un modello dinamico per gestire l'incertezza nei processi epidemiologici complessi. La proposta presentata si basa sul formalismo delle Reti Bayesiane dinamiche, in cui ciascun nodo corrisponde ad una variable aleatoria, il cui valore definisce lo stato di un individuo in una data popolazione.*

Key words: SIR Model, Dynamic Bayesian Network, Propagation of Uncertainty

1 Introduction

Deterministic epidemiological models for the propagation of an infectious disease on a given population were first introduced by Kermack and McKendrick [5]. We propose a graphical implementation of a simple SIR model, to account for uncertainty in the propagation process. This is carried out via repeated simulations from what we call an *almost-dynamic* Bayesian network, whose pattern of mutual conditional relevances evolves with time.
Basic concepts on the SIR model and Dynamic Bayesian Networks are introduced in Sec. 2.1 and 2.2, respectively. Our proposal is sketched in Sec. 2.3. Some results, remarks and possible extensions are discussed in Sec. 3.

Sabina Marchetti
Dip. Scienze Statistiche, Università Sapienza, P.le A. Moro 5, 00185 Roma (Italy), e-mail: sabina.marchetti@uniroma1.it

2 Methods

2.1 Deterministic Epidemiological Models

Infectious disease models are based on ordinary differential equations (ODEs). Each equation is associated with a single layer of the population, that is thus partitioned into homogeneous compartments. As an example, consider the well-known SIR model [5], where a given population is constituted by three mutually exclusive and exhaustive components: susceptible (S), infectious (I) and recovered (R), to some infectious disease. At any time t, $t \geq 0$, people in S may acquire the disease, according to rate $\lambda(t)$, and enter compartment I, where they spend γ^{-1} units of time on average, before they move to R; see Fig. 1. The rate $\lambda(t)$ is called *force of infection* and is often defined as

$$\lambda(t) = \tau c \frac{I(t)}{N(t)}\,, \tag{1}$$

where $c > 0$ is the average number of contacts per individual per unit of time, E.g. per day, resulting in an infection according to $\tau \in [0,1]$, *transmissibility* parameter, and $I(t)/N(t)$ is the proportion of infectious individuals in the population, also called the *infection prevalence*, at t. If the population is stratified into k homogeneous classes, that specify different behavior among components, $\lambda(t) \in \mathbb{R}^k$ is derived from the product of τ and contact matrix $C \in \mathbb{R}^{k \times k}$, multiplied by vector $I_k(t)/N_k(t) \in \mathbb{R}^k$, whose elements correspond to the proportion of infectious individuals in each class.

γ is called *recovery rate*. ODEs describe the flows across compartments. At each time step, $S(t), I(t)$ and $R(t)$ report the number of individuals in each homogeneous component of the population, assumed constant, i.e. $\frac{\partial S(t)}{\partial t} + \frac{\partial I(t)}{\partial t} + \frac{\partial R(t)}{\partial t} = 0$, for all $t \geq 0$. No demographic dynamics (births, deaths, ageing of the population, etc.) are considered by a simple SIR model, whose flow diagram is depicted in Fig. 1. Also, by definition individuals are assumed to acquire lifelong immunity to the disease considered, once recovered. Thanks to its simple parametrization, the SIR model was applied in a number of works, E.g., [4]. Particularly, the related set of assumptions it applies to short-termed diseases, where dynamics strictly related to the population may be neglected, as well as total or partial waning of the immunity conferred by an infection.

When an infectious individual is introduced in a fully susceptible population, the number of secondary cases she produces in a unit of time is called *basic reproduction number* [5], denoted as R_0. It may be easily proved that in an SIR model, it corresponds to $R_0 = \tau c \gamma^{-1}$ (see, E.g., [1]).

Uncertainty and sensitivity analysis of deterministic epidemiological models is usually performed following what we call a *second-order* approach: each parameter value is varied within some range according to some distribution of uncertainty, either singularly or simultaneously. Stochastic approaches were also proposed, dating back to the Reed-Frost model, in 1928 (see [2] for a survey), to deal with uncertainty in such propagation processes. In Sec. 2.3, we propose a dynamic network model, that accounts for uncertainty

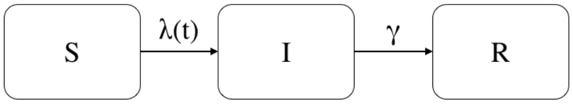

Figura 1 Flow diagram of a simple SIR model.

both in the parameters and in the pattern of contacts of any individual from the population. This way, while homogeneity of each compartment results in the parametrization of a collection of conditional probability tables, corresponding to transition matrices, the topology of the network increases (or decreases) risks at each time step. Our proposal is based on a graphical dynamic implementation of an SIR model. Our proposal is alternative to both existing epidemiological stochastic approaches, as it accounts for individual risks, and to the so-called *individual based* modeling [3], whose assumptions differ from those of an SIR.

2.2 Probabilistic Graphical Models

Probabilistic graphical models (PGMs) are used to represent the joint probability distribution of a (large) collection of random variables $\mathbf{V} = \{X_0, \ldots, X_n\}$, by a graph. Bold letters are used to denote sets of random variables. PGMs exploit conditional independence assumptions among pairs of random variables (in a one-to-one correspondence with the nodes of the graph) to reduce inferential complexity.

Nodes are connected by arcs ($-$) in an undirected graph, while edges ($\rightarrow$) are used to induce an ordering among the elements of $\mathbf{V}$ in a directed graph. We write $Adj(X)$ to denote the set of nodes *adjacent* to X in any graph, irrespective of their direction. Let X and Y be any two adjacent nodes in a directed graph, if there is an outgoing edge from X into Y, X is called a *parent* of Y, whereas Y is a *child* of X. $Pa(X)$ and $Ch(X)$ denote, respectively, the parents and children set of node $X \in \mathbf{V}$. Arcs are denoted as $((X,Y))$, while edges read (X,Y); i.e. let E be the set of links in the graph, $((X,Y)) \in E$ implies $X \in Adj(Y)$ and $Y \in Adj(X)$, while $(X,Y) \in E$ implies $X \in Pa(Y)$, $Y \in Ch(X)$, for any pair $X,Y \in \mathbf{V}$. Also, let $d_X = |Pa(X)|$ denote the in-degree of node X, i.e. cardinality of its parents set.

Bayesian networks are PGMs whose graphical component is an acyclic directed graph $\mathscr{G} = (\mathbf{V}, E)$. A Bayesian network is specified by the pair $(\mathscr{G}, P)$, where P is a strictly positive joint probability mass function over $\mathbf{V}$. By the Markov condition, for a given ordering in $\mathbf{V}$, each node is independent of its non-descendants in the graph, given its parents. It follows P may be equivalently represented by a collection of $n+1$ conditional probability tables (CPTs), whose columns correspond to distinct configurations of the parents of a node.

Dynamic Bayesian Networks (DBNs, [6]) are sequences of Bayesian networks, whose structure and/or parametrization change with time. In a DBN, conditioning always extends to each node's previous state, and possibly its parents'. As a result, the joint PMF at time t corresponds:

$$P(X_0^{t+1} = x_0^{t+1}, \ldots, X_n^{t+1} = x_n^{t+1}) = \prod_{i=0}^{n} P\left(X^{t+1} = x_i^{t+1} | X^t = x_i^t, Pa^{t+1}(X_i) = pa^{t+1}(X_i), Pa^t(X_i) = pa^t(X_i)\right) \tag{2}$$

where each configuration $(X_0^t = x_0^t, \ldots, X_n^t = x_n^t)$ belongs to product sample space $\Omega_{\mathbf{V}} = \times_{i='}^{n} \Omega_{X_i}$, and $Pa^t(X_i) = pa^t(X_i) \in \times_{X_j \in Pa^t(X_i)} \Omega_{X_j}$ is consistent with the first, $t \geq 0$. For a given event $\mathbf{x}^t \in \Omega_{X \in \mathbf{X}}$, we write $P(\mathbf{X}^t = \mathbf{x}^t) = P(\mathbf{x}^t)$ to simplify notation.

In graph theory, a population may be described by a network, whose nodes correspond to units, i.e. individuals, whereas in PGMs, nodes are random variables. In next section, we will introduce a simplified DBN whose nodes represent individuals. Each node X is associated a three-valued random variable, whose states, x_S, x_I and x_R, indicate her location in an SIR model.

2.3 Dynamic Probabilistic Modeling of an SIR Model

Let $\mathscr{B}^t = (\mathscr{G}^t, P^t)$ denote a graphical model at any time $t \geq 0$. In our formalism the graphical component is partially directed: with time, arcs may be changed into edges, and *vice versa*. Although conditioning ought to consider the whole adjacency set of each node, say X, as for undirected networks, relevance is restricted to those in $Pa(X) \subseteq Adj(X)$. Hence, at each t, $\mathscr{G}^t$ may be intended as an acyclic directed graph.

$\mathscr{G}^t = (\mathbf{V}, E^t)$, represents the pattern of contacts among units of a population. The graph may either be a given social network, or be randomly generated according to some known contact matrix C. Let $\mathbf{V}$ be the set of $(n+1)$ nodes (individuals in the network); as already mentioned, each random variable X_i takes values in $\Omega_{X_i} = \{x_{i,S}, x_{i,I}, x_{i,R}\}$, $i = 0, \dots, n$. State $x_{i,j}$ indicates X_i belongs to compartment j of an SIR structure, $j = S, I, R$. Also, each X_i is assigned label t_i, initialized as $t_i = -\infty$.

In our model, parameters of the model do not vary with time. Yet, if X_i takes value $x_{i,I} \in \Omega_{X_i}$, all incoming arcs are converted into $Ch^t(X_i)$[1] and t_j is updated to t. Let $d_{j,t} = |Pa^t(X_j)|$ be the *infectious-indegree* of node X_j, conditioning involves the subset of adjacent nodes of X_j in $Pa^t(X_j)$, and its corresponding state at $(t-1)$, for any $t \geq 1$.
Based on Eq. (1), we derive the individual FOI $\lambda_{j,t} = \tau d_{j,t-1}$, $j = 0, \dots, n$, $t \geq 0$. At each t, $P(X^t | X^{t-1}, Pa^t(X)) = \{P(x^t | x^{t-1}, pa^t(X)) : x^t, x^{t-1} \in \Omega_X, pa^t \in \Omega_{Pa^t(X)}\}$. The CPT of node X_j at t is specified in Table 1, that may be as well intended as a transition matrix, whose columns sum to one.

In details, the first column of the CPT represents the infection process, i.e. people moving from S to I in an SIR model. We assume the infectious-indegree of any node X_j serves as a proxy of the product $cI(t)/N(t)$ from Eq. (1). By definition, *zero-case* $(X = x_I)$, i.e. a single infectious individual in a fully susceptible population, is expected to produce $|Ch^0(X)|$ primary infections, at most. We define the basic reproduction number associated to $\mathscr{B} = \cup_{t=0}^{\infty} \mathscr{B}^t$ as follows:

$$R_0^{\mathscr{B}} = \min\left(|Ch^0(X)|, |Ch^0(X)|\tau\gamma^{-1}\right) \geq 0\,. \tag{3}$$

Let $\tau = 1$, if recovery is fast, i.e. γ is large, $R_0^{\mathscr{B}}$ will likely overestimate the number of secondary cases.

	$x_{j,S}^{t-1}; Pa^t(X_j)$	$x_{j,I}^{t-1}; Pa^t(X_j)$	$x_{j,R}^{t-1}; Pa^t(X_j)$
$x_{j,S}^t; t_j$	$e^{-\lambda_{j,t}}$	0	0
$x_{j,I}^t; t_j$	$1 - e^{-\lambda_{j,t}}$	$e^{-\gamma(t-t_j+\varepsilon)}$	0
$x_{j,R}^t; t_j$	0	$1 - e^{-\gamma(t-t_j+\varepsilon)}$	1

Tabella 1 CPT of node X_j, at time t in an SIR-Bayesian network. $\varepsilon \geq 0$, in our application (see Fig. 2) we set $\varepsilon = 1$.

3 Results and Discussion

As a toy example, we applied our proposal to a population of $(n+1) = 20$ individuals, whose pattern of contacts is depicted in Fig. 2(top-panel). The propagation schema was produced over $M = 1000$ simulations of the model with $\tau = 0.45$ and $\gamma = 0.60$. At each m, introduction of a randomly selected infectious zero-

[1] Arcs only are converted into outgoing edges: if X_i has already incoming edges, those are unchanged.

cases produced on average 3.56 secondary cases.[2] We compared our results with the corresponding SIR model; epidemic curves (and empirical quantile values of uncertainty) are depicted in Fig. 2(top-panel).

Let us state some general remarks. As a first, we stress recovery of some node X_i, at $t \geq 1$, acts as a blocking mechanism, flooring $P(x^t_{j,I}|X^{t-1}, Pa^t(X_j) = \{X_i\}, t_j = \infty)$ to zero.[3] Detection of paths originating from the zero-case, say X_0, that are likely to be blocked by recovery of a single node (or few of them) allows prior identification of critical subjects. Those subjects may be targeted by prevention strategies, tackling a minimal subset approach. Graphical tools may be used to detect *blocked* from *active* paths [7] for the propagation of a disease, such as the Bayes Ball algorithm [8].

In this direction, we argue a general evaluation of the topology of the network would critical in this sense. E.g., for a fixed transmissibility, a sparse graph will be more exposed to the blocking mechanisms mentioned above, compared to a denser one. Also, identification of *cliques*, i.e. maximally connected sets of nodes, may constitute valuable knowledge to policy planning.

As a second remark, repeated sampling allows to evaluate uncertainty in the overall propagation process. Several sampling procedures were proposed in the literature of DBNs, see, E.g., [6]. A efficient naive approach would simply take the so-called *maximum a posteriori* (MAP) configurations from $\Omega_{\mathbf{V}}$, at each time step, to update E^t.

Additionally, other than $R_0^{\mathscr{B}}$, measures on the disease propagation process may be derived from repeated iterations, as well as analytically. Among others, incidence and sero-prevalence of a disease [1]. Again, MAP configurations may be considered as average values prior to simulating.

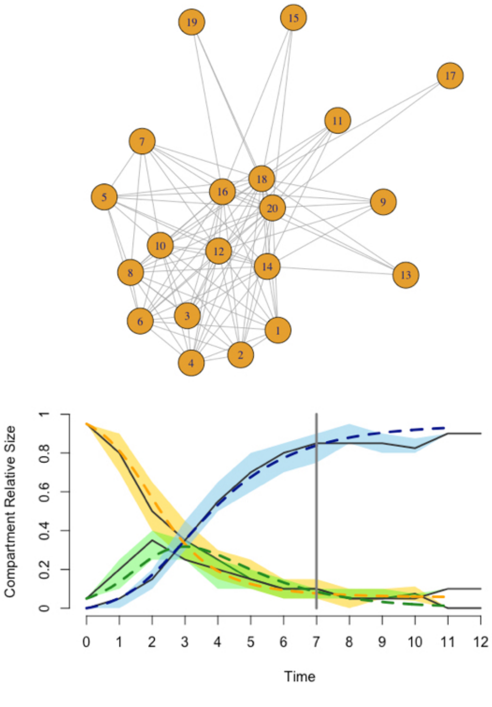

Figura 2 Top-panel: Network generated at random, with $|\mathbf{V}| = 20$, average in-degree 10 and $d = 19$. Bottom-panel: Epidemic curves resulting from $M = 1000$ simulations on the network above, $|\mathbf{V}| = 20$, $\tau = 0.45$, $\gamma = 0.60$. At every simulation, a node is selected at random as zero-case. Epidemic curves describe the relative size of compartments S (yellow), I (green) and R (light blue). Quantile bands and median are compared with the curves resulting from the SIR network, with $c = 0.20$; dashed lines orange, dark green and blue correspond, respectively to compartments S, I and R.

[2] We expected $R_0^{\mathscr{B}} \in [1.33, 10]$, by Eq. (3).

[3] Since $d_{j,t} = 0$.

As a further point, a straightforward extension might assume contacts are characterized by different strengths $w_{i,j} \in [0,1]$, such that $w_{i,j} \to 0$ indicates a almost *vacuos* contacts between nodes X_i and X_j, for any $t \geq 0$. Then, a more general definition of infectious-indegree may be introduced:

$$d_{j,t} = \sum_{X_i \in Pa^t(X_j)} w_{i,j} \mathbb{I}_{X_i^t = x_{i,I}^t} .$$

Finally, suppose we are interested in modeling an SIRS model, where recovered individuals move back to compartment S according to some rate $\phi \geq 0$, that is after ϕ^{-1} units of time in average. It suffices to replace the third column of Table 1 with $\left[e^{-\phi}, 0, 1 - e^{-\phi}\right]^T$.

4 Conclusions and Future Work

We proposed an almost-Dynamic Bayesian Network to efficiently deal with uncertainty in the propagation process of a given infectious disease in an SIR model. Particularly, our proposal models uncertainty in the propagation process by i) probabilistic modeling of the transitions across compartments (analogously to a stochastic SIR model), ii) accounting for the dynamic topology of the network (like any individual-based model). We stress our proposal is aimed to provide an intuition of the methodology: extension to models with several layers of complexity is straightforward, the increased complexity being restricted to the preliminary compilation process, i.e. to its parametrization, without affecting the inferential complexity. Future work will consider applications in this direction.

Future research will also consider a thorough approach to uncertainty, by incorporating uncertainty in the parameters, E.g. by means of auxiliary root nodes, to further extend stochastic epidemiological modeling based.

Riferimenti bibliografici

[1] Roy M Anderson, Robert M May, and B Anderson. *Infectious diseases of humans: dynamics and control*, volume 28. Wiley Online Library, 1992.

[2] Tom Britton. Stochastic epidemic models: a survey. *Mathematical biosciences*, 225(1):24–35, 2010.

[3] Volker Grimm and Steven F Railsback. Individual-based modeling and ecology:(princeton series in theoretical and computational biology). 2005.

[4] Matt J Keeling and Pejman Rohani. *Modeling infectious diseases in humans and animals*. Princeton University Press, 2011.

[5] W Kermack and A McKendrick. A contribution to the mathematical theory of epidemics. In *Proc. Roy. Soc. Lond*, pages 700–721, 1927.

[6] Kevin Patrick Murphy. Dynamic bayesian networks: representation, inference and learning. 2002.

[7] Judea Pearl. *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Elsevier, 2014.

[8] Ross D Shachter. Bayes-ball: Rational pastime (for determining irrelevance and requisite information in belief networks and influence diagrams). In *Proceedings of the Fourteenth conference on Uncertainty in artificial intelligence*, pages 480–487. Morgan Kaufmann Publishers Inc., 1998.