INRA
Institut National de la Recherche Agronomique

MIA
Mathématiques et Informatique Appliquées

EVALUATION OF THE EXPECTED SIZE OF A SIR EPIDEMICS ON A GRAPH

BY

Nathalie Peyrard
Régis Sabbadin

Rapport de Recherche RR-2012-1
Février 2012

Unité de Biométrie et Intelligence Artificielle
Institut National de la Recherche Agronomique
Toulouse, France
http://carlit.toulouse.inra.fr/wikiz/index.php/Accueil

Evaluation of the expected size of a SIR epidemics on a graph

N. Peyrard[a], R. Sabbadin[a]

[a] UR 875 - Unite de Biométrie et Intelligence Artificielle
24 chemin de Borde Rouge Auzeville CS 52627
31320 Castanet Tolosan - FRANCE

Abstract

In this article, we study the problem of evaluating the expected cost (generalizing the notion of expected size) of an epidemics spreading over a finite network of individuals according to a discrete-time SIR dynamics. Studies from the literature provide results in expectation over random graphs of infinite size, or rely on Monte Carlo simulations. Our results are threefold. (i) We prove that the evaluation problem for a finite graph is #P-complete. (ii) We propose an exact "divide and conquer" evaluation algorithm, for which we provide upper bounds on its time complexity. In particular, we prove that it can be polynomial when the graph is a tree. (iii) We propose an approximation algorithm, based on the mean field principle, with polynomial time complexity for any graph. An experimental comparison of the exact and Monte Carlo evaluations with the results of the mean field algorithm show that the latter provides a significant gain in computational time while leading to satisfying approximation quality.

Keywords: divide and conquer algorithm, mean field approximation.

1 INTRODUCTION

Stochastic spatio-temporal models on graphs offer powerful tools for understanding disease propoagation among humans or animals. Two classical models are the Susceptible-Infected-Susceptible (SIS) and the Susceptible-Infected-Removed (SIR) models. These models have more often been studied in their continuous time version (see House and Keeling, 2011, Peyrard et al., 2008 and references therein). In this article, having in perspective the question of disease control and the fact that control strategies are often the result of successive decisions taken at regular time points (vaccination plans, public places closings), we consider a discrete representation of time. We also focus on the SIR model, which provides a reasonable representation of several human diseases (Salathé and Jones, 2010), and in particular childhood diseases (Ferrari et al., 2006).

The objective of this work is to study the problem of evaluating the Expected Epidemics Cost (EEC) corresponding to the situation where different costs of infection may be attached to distinct individuals. A particular case is when all costs are equal, leading to the Expected Epidemics Size (EES) problem. Under the hypothesis of one-step infection duration, we use similarities with network reliability computation (Ball, 1980) to establish the computational complexity of EEC evaluation. Then we address the question of exact and approximate computation of this quantity, and thus of EES. Exact results are scarce in literature. For a general SIR model, Newman (2002) has shown how to compute exactly EES numerically, using tools from percolation theory. These results are in average over (infinite size) random graphs with fixed degree distribution. Neal (2003) establishes the asymptotic epidemics size distribution of a SIR model on a Bernouilli random graph. When dealing with a concrete interaction network, we are interested in the evaluation of EEC/EES for this specific network.

We propose an exact algorithm to achieve this task. This algorithm is defined recursively, and uses the fact that Removed individuals may split the initial interaction network into disjoint networks with independent spread dynamics. Then, to be able to handle large graphs, we propose an alternative to Monte Carlo simulations (Ferrari et al., 2006; Salathé and Jones, 2010), relying on the mean field principle (Chandler, 1987). When applied to spatio-temporal models on graph, it amounts to the approximation of the joint distribution of n individuals by n independent random chains. Usually, in the mean field approximation the n independent random chains have identical distributions, or these distributions only depend on the number of neighbors of each individual in the network. the In order to improve the approximation quality, we derive a mean field approximation with different distributions for the n individuals.

The SIR model is described in Section 2. We establish the complexity of EEC evaluation in Section 3. Then the exact and approximate solution algorithms are presented in Sections 4 and 5. Their relative performances are studied and compared to Monte Carlo simulation results in Section 6.

2 SIR EPIDEMICS ON A GRAPH

A stochastic SIR model on a graph can be used to model the spread of a disease on a network of individuals. A group of n individuals is considered, and a random variable X_i^t is attached to each one, representing the sanitary status of individual i at time t. Each individual can be in one of the three states, Susceptible (S), Infected (I) or Removed (R), so $X_i^t \in \mathcal{X} = \{S, I, R\}$.

A directed graph $G = (V, E)$ is used to model the possible transmission paths between individuals. An individual is represented by a vertex $i \in V = \{1 \ldots, n\}$ in the graph. We assume that the graph G is connected. If $(j, i) \in E$, it means that direct contamination from j to i is possible. The neighborhood $N(i) \subseteq V$ of a vertex i is the set of vertices which can contaminate i: $N(i) = \{j \in V, (j, i) \in E\}$.

Then SIR dynamics are as follows. If ρ_{ji} is the probability that the infection is transmitted from vertex $j \in N(i)$ in state I to vertex i in state S, then, for a given configuration $x_{N(i)}^t$ of $X_{N(i)}^t$ we have

$$\begin{aligned} &p_i(X_i^{t+1} = I | X_i^t = S, X_{N(i)}^t = x_{N(i)}^t) = \\ &1 - \prod_{j \in N(i), x_j^t = I} (1 - \rho_{ji}). \end{aligned}$$

In other terms, we assume that disease transmission events are independent. Other transition probabilities are deterministic:

$$p_i(X_i^{t+1} = R | X_i^t = I) = 1,\ p_i(X_i^{t+1} = R | X_i^t = R) = 1$$

We make the assumption of a one-step duration of infection, which means that the time unit considered is the period during which an infected vertex can infect its neighbors. The state R is absorbing.

We are interested in the problem of computing the Expected Epidemics Size (EES) of a SIR process with one-step infection duration, or more generally the Expected Epidemics Cost (EEC). Indeed, different costs may be assigned to infected vertices. We define the cost vector $c = \{c_1, \ldots, c_n\}$ where $c_i \geq 0$ is the cost incurred when $X_i^\infty = R$. If we denote by $\mathcal{I} \subseteq V$ (resp. $\mathcal{S} \subseteq V$) the set of infected (resp. susceptible) vertices at the beginning of the epidemics (time $t = 0$, where it is assumed that there are no R vertices), EEC is equal to

$$EEC(G, \mathcal{I}, \mathcal{S}, \rho, c) = E\left[\sum_{i \in V} c_i \mathbb{1}_{[X_i^\infty = R]} \mid G, \mathcal{I}, \mathcal{S}, \rho, c\right].$$

The particular case where all c_i are equal to 1 leads to the EES value, the expected number of R vertices at the end of the epidemics. EES is also equal to the expected number of vertices which are infected during the course of the epidemics.

3 COMPUTATIONAL COMPLEXITY OF EEC EVALUATION

In this section, we study the computational complexity of EEC. An EEC evaluation problem is defined as a pair $< \mathcal{P}, EEC >$, where $\mathcal{P}$ is an instance of the problem, and $EEC(\mathcal{P})$ is the measure of $\mathcal{P}$, which has to be computed.

EEC Evaluation Problem

- Problem instance $\mathcal{P} =< G = (V, E), \mathcal{I}, \mathcal{S}, \rho, c >$
- Problem measure:

$$EEC(\mathcal{P}) = E\left[\sum_{i \in V} c_i \mathbb{1}_{[X_i^\infty = R]} \mid \mathcal{P}\right].$$

In the following, we show that the epidemics evaluation problem $\mathcal{P}$ is $\#P$-complete, where $\#P$ is the counterpart of the complexity class NP for counting problems. Note that EEC is not an integer-valued function, so it is not in $\#P$, stricto-sensu. However, it can be easily shown that, provided that the ρ_{ij}s and c take rational values only, computing EEC comes down to computing an integer-valued function. Therefore, it is meaningful to explore $\#P$-completeness.

3.1 THE EEC EVALUATION PROBLEM IS IN $\#P$

To show the $\#P$ membership and hardness of the EEC evaluation problem, we will use its close similarity to the Source-to-Terminal Reliability problem (Ball et al., 1992), defined as:

Source-to-Terminal Reliability Problem

- Problem instance $\mathcal{R}_{s,t} =< G = (V, E), s, t, p >$ where:
  - $G = (V, E)$ is a directed graph.

  - $s \in V$ is a source vertex.
  - $t \in V$ is a terminal vertex.
  - $p : E \to [0,1]$ is a reliability function. $p(e)$ is the probability that edge $e \in E$ does not fail. All edges states (failing or not) are assumed to be independent.

  The reliability function p defines a probability measure Pr over the set of subgraphs of G: if $G' = (V, E')$ where $E' \subseteq E$, then, under the assumption of edges states independence, the probability that only the edges in E' do not fail is

$$Pr(G') = \left(\prod_{e \in E'} p_e \right) \left(\prod_{e \in E \setminus E'} (1 - p_e) \right).$$

- The problem measure $Rel(\mathcal{R}_{s,t})$ is the probability that there exists at least one path from s to t in G which contains only edges which have not failed:

$$Rel(\mathcal{R}_{s,t}) = \sum_{G' \in Connect(G,s,t)} Pr(G'),$$

  where $Connect(G, s, t)$ is the set of subgraphs of G in which s and t are connected.

The Source-to-Terminal Reliability problem is $\#P$ complete (Ball, 1980). In order to show that the computation of EEC belongs to $\#P$, we are going to show that for any $\mathcal{P}$, $EEC(\mathcal{P})$ can be computed by $O(|V|)$ calls to an oracle computing $Rel(\mathcal{R}_{s,t})$, for $\mathcal{R}_{s,t}$ instances which are easily (in polynomial time) computed from a $\mathcal{P}$ instance.

Proposition 1 *The EEC evaluation problem belongs to $\#P$.*

Proof: Let $\mathcal{P} :< G = (V, E), \mathcal{I}, \mathcal{S}, \rho, c >$ be a problem instance. Remark that

$$EEC(\mathcal{P}) = \sum_{j \in \mathcal{I}} c_j + \sum_{i \in \mathcal{S}} c_i E \left[\mathbb{1}_{[X_i^\infty = R]} \mid \mathcal{P} \right].$$

Let us write $\mu_i = E \left[\mathbb{1}_{[X_i^\infty = R]} \mid \mathcal{P} \right]$: μ_i is the probability that the infection spreads from $\mathcal{I}$ to vertex i in the directed graph G. Under the one-step infection duration hypothesis, μ_i is exactly the probability that there exists a path composed of unfailing edges from a vertex in $\mathcal{I}$ to i, in the *unreliable network* G where the edge reliability function is $\rho = p$. Indeed, i can get infected if and only if there is a subset of edges $E' \subseteq E$ forming a path from $\mathcal{I}$ to i, which is actually followed by the infection. This is equivalent to the fact that edges $E' \subseteq E$ do not fail in the unreliable network G. So, the probability of infection of i can be computed by solving a network reliability problem. The only difference between computing μ_i and solving a network reliability problem is that the source is not a single vertex s, but instead a set of source vertices $\mathcal{I}$.

Let us denote $\mathcal{R}_{\mathcal{I},t} =< G = (V,E), \mathcal{I}, t, p >$ the problem of computing the probability that there exists a path composed of unfailing edges from a vertex in $\mathcal{I}$ to vertex t. It can be transformed in polynomial time into a classical reliability problem $\mathcal{R}_{s,t} =< G' = (V', E'), s, t, p' >$, where

- $V' = (V \setminus \mathcal{I}) \cup \{s_{\mathcal{I}}\}$, where $s_{\mathcal{I}}$ is an additional vertex obtained by merging all vertices in $\mathcal{I}$.
- $E' = E \setminus \Big\{(j,k) \in E, j \in \mathcal{I}\Big\} \cup \Big\{(s_{\mathcal{I}}, k),$ where $k \in V \setminus \mathcal{I}$ and $\exists j \in \mathcal{I}, (j,k) \in E\Big\}$. E' is obtained by merging all edges linking vertices in $\mathcal{I}$ to the same vertex $k \in V \setminus \mathcal{I}$.
- p' is obtained from p, by letting $p'_e = p_e$ when $e = (i,j)$ is such that $\{i,j\} \cap \mathcal{I} = \emptyset$, and

  $$\forall (s_{\mathcal{I}}, j) \in E', p'_{(s_{\mathcal{I}},j)} = 1 - \prod_{i \in \mathcal{I}, (i,j) \in E} (1 - p_{(i,j)}).$$

  $p'_{(s_{\mathcal{I}},j)}$ is the probability (given p) that at least one edge linking a vertex in $\mathcal{I}$ to j does not fail.

Obviously, $Rel(\mathcal{R}_{s_{\mathcal{I}},t}) = \mu_t$. Thus, $\forall i \in \mathcal{S}$, μ_i can be computed by a single call to a $\mathcal{R}_{s,t}$ oracle, and $EEC(\mathcal{P})$ can be computed in polynomial time, provided that there exists a polynomial time oracle for computing $\mathcal{R}_{s,t}$. Thus, the EEC evaluation problem belongs to $\#P$. □

3.2 THE EEC EVALUATION PROBLEM IS $\#P$-COMPLETE

We show that the EEC evaluation problem is $\#P$-hard, by reduction of the Source-to-Terminal Reliability problem to the EEC evaluation problem.

Proposition 2 *The EEC evaluation problem is $\#P$-hard.*

Proof: Consider the following instance $\mathcal{R}_{s,t} =< G = (V, E), s, t, p >$. The measure of this instance can be computed from the value of the following instance $\mathcal{P} =< G, \mathcal{I}, \mathcal{S}, \rho, c >$ of the EEC evaluation problem, where $\rho = p$, $\mathcal{I} = \{s\}$, $\mathcal{S} = V \setminus \{s\}$, $c_i = 0, \forall i \neq t$ and $c_t = 1$. Indeed, in that case

$$EEC(\mathcal{P}) = E\left[\mathbb{1}_{[X_t^\infty = R]} \mid \mathcal{P} \right] = Rel(\mathcal{R}_{s,t}).$$

So, the EEC evaluation problem is $\#P$-hard. □

Then, from Propositions 1 and 2:

Proposition 3 *The EEC evaluation problem is $\#P$-complete.*

4 EXACT COMPUTATION OF EEC

The EEC value can be computed exactly by a recursive *divide and conquer* strategy. Let us assume that the configuration of the SIR process at time $t = 0$ has the structure shown in Figure 1 (ignoring R vertices, which do not play any role in the epidemics).

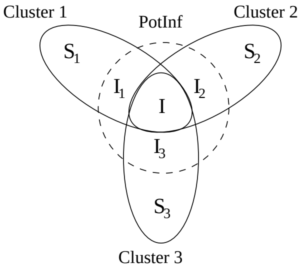

Figure 1: Divide and conquer strategy for computing the expected epidemics cost.

In this example, $\mathcal{I}$ being given, the remaining set of $\mathcal{S}$ vertices can be partitioned into three clusters of vertices[1], Cl_1, Cl_2, Cl_3, such that there exists

[1] For an arbitrary set $\mathcal{I}$, a partition always exists, but in the worst case, there is a single cluster.

no path in the graph G linking two S vertices from distinct clusters without going through vertices in $\mathcal{I}$. This decomposition will allow us to decompose the computation of $EEC(G, \mathcal{I}, \mathcal{S}, \rho, c)$. Indeed, consider $PotInf$, the set of vertices which could get infected between $t = 0$ and $t = 1$: $PotInf = \{j \in \mathcal{S}, N(j) \cap \mathcal{I} \neq \emptyset\}$.

In the next time step, all vertices in $\mathcal{I}$ will become R, thus splitting the problem into three independent problems with disjoint vertices sets: $\{Cl_k = S_k \cup I_k\}_{k=1\ldots3}$, where $I_k \subseteq PotInf \cap Cl_k$ is the set of vertices in cluster k which will actually get infected, and $S_k = Cl_k \setminus I_k$. Denoting $G^{\downarrow Cl_k}$ the graph restricted to vertices in Cl_k, the expected epidemics cost can be computed through the following recursive equations:

$$\begin{aligned}
&EEC(G, \mathcal{I}, \mathcal{S}, \rho, c) = 0 \text{ if } \mathcal{I} = \emptyset \text{ and else}\\
&EEC(G, \mathcal{I}, \mathcal{S}, \rho, c) = c(\mathcal{I}) + \sum_k \sum_{I_k \subseteq PotInf \cap Cl_k}\\
&\left(p(I_k|\mathcal{I}, G) EEC(G^{\downarrow Cl_k}, I_k, S_k, \rho^{\downarrow Cl_k}, c^{\downarrow Cl_k})\right).
\end{aligned}$$

where $\rho^{\downarrow Cl_k}$ and $c^{\downarrow Cl_k}$ are respectively the restrictions of ρ and c to the subgraph Cl_k. Algorithm 1 is the implementation of this recursive procedure.

Note that the time complexity $T(n)$ of Algorithm 1 can greatly vary with the graph structure. We show in the Supplementary Material that, in the case of a single initial infected vertex, $T(n)$ is linear in n if the graph is a tree , and $T(n) = O(2^{\frac{n(n+1)}{2}})$ when the graph is a clique. The $\#P$-hardness of the problem makes the existence of a time-efficient exact solution algorithm very unlikely. Therefore, in the following we present an approximation algorithm based on the mean field approximation.

5 MEAN FIELD APPROXIMATION

When applied to the SIR model, the mean field approximation amounts to replace a Markov chain with n state variables with its best approximation by n independent Markov chains over a single variable. The quality of the approximation is measured by the Kullback-Leibler divergence. The solution of this optimization problem is presented in Section 5.1. However, the computation of this global mean field approximation is too complex. Therefore, we propose an algorithm which computes an iterative mean field solution (Section 5.2). This procedure amounts to computing iteratively a local mean field approximation at each time step of the process.

```
Algorithm: SIR_EEC_Exact
Data : $\{G = (V, E), \mathcal{I}, \mathcal{S}, \rho, c\}$
Result: $\{EEC\}$
% Initialization;
$EEC \leftarrow c(\mathcal{I})$;
CLUSTERS $\leftarrow$ Partition of $\mathcal{S}$ into subsets, each subset being
connected to $\mathcal{I}$ but disconnected to the others by $\mathcal{I}$ vertices;
POTINF $\leftarrow$ $\mathcal{S}$ vertices which have $\mathcal{I}$ vertices as neighbors;
if CLUSTERS$\neq \emptyset$ then
    for $Cl_k \in$ CLUSTERS do
        for $I_{next} \subseteq$POTINF$\cap Cl_k$ do
            $EEC_{next} \leftarrow$
            SIR_EEC_Exact($G^{\downarrow Cl_k}, I_{next}, Cl_k \setminus I_{next}, \rho^{\downarrow Cl_k}, c^{\downarrow Cl_k}$);
            $EEC \leftarrow EEC + p(I_{next}|G, \mathcal{I}) \times EEC_{next}$;
        end
    end
end
```

Algorithm 1: SIR_EEC_Exact

5.1 GLOBAL MEAN FIELD

Let us consider an instance $\mathcal{P} =< G = (V, E), \mathcal{I}, \mathcal{S}, \rho, c >$, and a vertex i in $\mathcal{S}$,

$$\begin{aligned} & E[\mathbb{1}_{\{X_i^\infty = R\}} \mid \mathcal{P}] = p(X_i^\infty = R \mid \mathcal{P}) \\ = \; & p(\exists\, 1 \leq t < \infty \; s.t. \; X_i^t = I \mid \mathcal{P}) \\ = \; & \sum_{1 \leq t < \infty} p(X_i^t = I \mid \mathcal{P}) \end{aligned} \quad (1)$$

The last equality holds since events $\{X_i^t = I\}$ are incompatible under the one-step infection duration assumption. A vertex in $\mathcal{S}$ cannot be reached anymore by the epidemics after a certain number of time steps, bounded by n. Thus the sum (1) has at most n non-zero terms. Equality (1) shows that the complexity of the computation of $EEC(\mathcal{P})$ is due to the computation of the marginal probabilities $p(X_i^t = I \mid \mathcal{P})$. In the following, we build the mean field approximation of the joint spatio-temporal distribution of the SIR process, from which we will derive an approximation of these marginals probabilities.

Let $X^t = \{X_1^t, \ldots, X_n^t\}$ represent the state of all vertices at time t. In the SIR model, the joint spatio-temporal distribution of $\{X^0, X^1, \ldots, X^T\}$ $(1 \leq T \leq n)$ is given by: $\forall \{x^0, x^1, \ldots, x^T\} \in \mathcal{X}^{n \times T}$

$$p(x^0, x^1, \ldots, x^T) = p^0(x^0) \prod_{t=1}^{T} \prod_{i=1}^{n} p_i(x_i^t \mid x_i^{t-1}, x_{N(i)}^{t-1}). \quad (2)$$

(From now on the conditioning on $\mathcal{P}$ is omitted in the notations.) Let $\mathcal{Q}$ be the family of distributions of n independent Markov chains, such that $q^0(x^0)$ is equal to the Dirac distribution $p^0(x^0)$ defined by $\mathcal{I}$ and $\mathcal{S}$ and transition probabilities are of the form given in Table 1.

	S	I	R
S	$1 - q_i^t(X_i^t = I \mid X_i^{t-1} = S)$	0	0
I	$q_i^t(X_i^t = I \mid X_i^{t-1} = S)$	0	0
R	0	1	1

Table 1: Transition probabilities for distribution q.

For $q \in \mathcal{Q}$, the joint spatio-temporal distribution becomes, $\forall \{x^0, x^1, \ldots, x^T\} \in \mathcal{X}^{n \times T}$,

$$q(x^0, x^1, \ldots, x^T) = q^0(x^0) \prod_{t=1}^{T} \prod_{i=1}^{n} q_i^t(x_i^t \mid x_i^{t-1}). \tag{3}$$

Note that for distributions in $\mathcal{Q}$, the transition probabilities can depend on time. We then define the mean field approximation of the joint spatio-temporal distribution (2) as the distribution q^* in $\mathcal{Q}$ which minimizes the Kullback-Leibler divergence between q and p: $q^* = \arg\min_{q \in \mathcal{Q}} KL(q||p)$, with

$$KL(q||p) = \sum_{x^0,\ldots,x^T} q(x^0, \ldots, x^T) \log \frac{q(x^0, \ldots, x^T)}{p(x^0, \ldots, x^T)}.$$

Minimizing this expression over the distribution q is a complex optimization problem. Indeed, $KL(q||p)$ is equal to:

$$\sum_{t=1}^{T} \sum_{x^{t-1},x^t} q^{t-1}(x^{t-1}) q^t(x^t \mid x^{t-1}) \log \left(\frac{q^t(x^t \mid x^{t-1})}{p^t(x^t \mid x^{t-1})} \right).$$

Since $q^{t-1}(x^{t-1})$ can be obtained by marginalization of $q(x^0, \ldots, x^{t-1}) = q^0(x^0) \prod_{s=1}^{t-1} q^s(x^s \mid x^{s-1})$ over the set of variable $\{x^0, \ldots, x^{t-2}\}$, $KL(q||p)$ is equal to

$$\sum_{t=1}^{T} \sum_{x^0,\ldots,x^t} q^0(x^0) \prod_{s=1}^{t} q^s(x^s \mid x^{s-1}) \log \left(\frac{q^t(x^t \mid x^{t-1})}{p^t(x^t \mid x^{t-1})} \right).$$

From this expression, we can see that the quantity $q^s(x^s \mid x^{s-1})$ is involved in the last $T - s + 1$ terms of the temporal sum. So computing a solution of the minimization problem would require to mobilize backward algorithms for continuous optimization. In the following, instead of globally minimizing the Kullback-Leibler divergence between q and p, we perform several minimizations of local Kullback-Leibler divergences, at successive time steps, in order to approximate the transition probabilities, using the fact that $KL(q||p)$ can be rewritten as:

$$\sum_{t=1}^{T} KL\left(q^{t-1}(.) q^t(.|.) || q^{t-1}(.) p(.|.)\right).$$

5.2 ITERATIVE MEAN FIELD

The iterative mean field procedure is derived from the expression of the global Kullback-Leibler divergence $KL(q||p)$ by successive minimizations of the first terms involving $q^t(x^t \mid x^{t-1})$ in the temporal sum, and then replacing this transition probability with the result of the minimization in the following terms of the sum:

- $t = 0$: set $q^0 = p^0$

- $t = 1$: compute $\hat{q}^1(x^1 \mid x^0)$ solution of

$$\arg \min_{q^1(x^1|x^0)} \sum_{x^0,x^1} q^0(x^0) q^1(x^1 \mid x^0) \log \left(\frac{q^1(x^1 \mid x^0)}{p(x^1 \mid x^0)} \right).$$

- $t = 2$: compute $\hat{q}^2(x^2 \mid x^1)$ solution of

$$\arg \min_{q^2(x^2|x^2)} \sum_{x^1,x^2} \hat{q}^1(x^1) q^2(x^2 \mid x^1) \log \left(\frac{q^2(x^2 \mid x^1)}{p(x^2 \mid x^1)} \right),$$

  with $\hat{q}^{(}x^1) = \sum_{x^0} q^0(x^0) \hat{q}^1(x^1 \mid x^0)$.

- and so on until $t = T$.

This iterative mean field procedure has already been proposed in Peyrard and Sabbadin (2006), in the context of controlled processes, in order to approximate a complex spatio-temporal distribution on a graph . In our case, the simplicity of the SIR model (several transitions have probability 0 or 1 and there is no control) allows to obtain a simple expression of the solution of the iterative mean field procedure. The solution $\hat{q}^t(x^t \mid x^{t-1})$ which minimizes

$$\sum_{x^{t-1},x^t} \hat{q}^{(}x^{t-1}) q^t(x^t \mid x^{t-1}) \log \left(\frac{q^t(x^t \mid x^{t-1})}{p^t(x^t \mid x^{t-1})} \right)$$

is (see Supplementary Material and Peyrard and Sabbadin, 2006): $\forall x_i^t, x_i^{t-1} \in \mathcal{X}$,

$$\hat{q}_i^t(x_i^t \mid x_i^{t-1}) \propto \exp \left(E_{\hat{q}^{t-1}} [\log p_i(x_i^t \mid x_i^{t-1}, X_{N(i)}^{t-1})] \right), \qquad (4)$$

where $E_{\hat{q}^{t-1}}[.]$ stands for the expectation over the distribution $\hat{q}^{t-1}$ of $X_{N(i)}^{t-1}$. This solution must be normalized. In practice it is classical to switch exponential and expectation operators in (4), to get the approximation:

$$\hat{q}_i^t(x_i^t \mid x_i^{t-1}) = E_{\hat{q}^{t-1}}[p_i(x_i^t \mid x_i^{t-1}, X_{N(i)}^t)] \tag{5}$$

This is all the more relevant for the SIR model as expression (4) leads to $\hat{q}_i^t(I \mid S) = 0$ as soon as $\hat{q}(x_j^{t-1} = S) > 0, \forall j \in N(i)$ (since $\log p_i(x_i^t = I \mid x_i^{t-1} = S, X_j \neq I, \forall j \in N(i)) = -\infty$). For the SIR model, equation (5) leads to (see Supplementary Material):

$$\begin{aligned}
&\hat{q}_i^t(X_i^t = I \mid X_i^{t-1} = S)\\
&= 1 - \sum_{x_{N(i)}} \prod_{j \in N(i)} [(1-\rho_{ji})^{\delta_I(x_j)} \hat{q}_j^{t-1}(X_j^{t-1} = x_j)]\\
&= 1 - \prod_{j \in N(i)} (1-\rho_{ji})\; \hat{q}_j^{t-1}(X_j^{t-1} = I))
\end{aligned} \tag{6}$$

with $\delta_I(x_j)$ equals 1 if $x_j = I$ and zero otherwise. The other transition probabilities are given in Table 1. Then $\hat{q}_i^t$ for $i \in \mathcal{S}$ is obtained as:

$$\begin{aligned}
\hat{q}_i^t(I) &= \hat{q}_i^t(I \mid S)\hat{q}_i^{t-1}(S),\\
\hat{q}_i^t(S) &= \hat{q}_i^t(S \mid S)\hat{q}_i^{t-1}(S),\\
\hat{q}_i^t(R) &= 1 - \hat{q}_i^t(I) - \hat{q}_i^t(S).
\end{aligned}$$

For a vertex $i \in \mathcal{I}$, $\hat{q}_i^t(X_i^t = R) = 1,\ \forall t > 1$.

Algorithm 2, SIR_EEC_MF, is one way to compute the iterative mean field approximation of EEC. Its time complexity is in $\mathcal{O}(n^2 \times \max_{i \in V} |N(i)|)$.

6 EMPIRICAL EVALUATION OF MEAN-FIELD APPROXIMATION

In order to evaluate the quality of the iterative mean field approximation of EES, we compare its evaluation, first with the evaluation obtained using the exact procedure of Section 4 on small graphs (up to 16 vertices), and then with an evaluation obtained by Monte Carlo simulations for larger graphs (up to 5000 vertices).

6.1 SMALL GRAPHS

We considered two families of graphs: random graphs (Newman, 2003) and stochastic block structures (SBS) graphs (Nowicki and Snijders, 2001). To generate a random graph, each potential edge between two vertices is considered in turn and is created with a constant probability p. In a SBS graph, vertices are grouped into classes (clusters) and the edge creation probability p_{ab} varies with the classes a and b of the two vertices involved. We ran experiments on graphs of size 8, 12 and 16, with $p = 0.3$ in the case of the random graph model, and $\{p_{aa} = 0.6, p_{ab} = 0.2, \forall a \neq b\}$ in the case of the SBS model, with 3 classes of vertices of identical size. The proportion $pI0$ of infected vertices at the beginning of the epidemics was set to 0.3 and the epidemic parameters ρ_{ij} were all identical, equal to 0.2. For each graph model and graph size, we generated 10 pairs $(G, \mathcal{I})$ for which the exact and mean field evaluations were computed. Relative and absolute errors are plotted on Figure **??**.

In almost all examples the relative error remains below 15%. Higher relative errors correspond to small size epidemics and must be put into perspective with the corresponding absolute error (expected epidemics size over estimated by at most 1.2 individuals). The gain in computational time with the SIR_EEC_MF algorithm is significant, even for the small graph sizes we considered (see Table 2). Even though a Monte Carlo evaluation (5000 trajectories) is much faster than the exact algorithm, it requires much more time than SIR_EEC_MF.

Cluster. and pour EES

n	8	12	16
Exact	0.0	3.5	3234.9
Monte Carlo	0.4438	0.6061	0.7624
Mean field	0.1729	0.0002587	0.0003596

n	8	12	16
Exact	0.0	39.4045	5006.5
Monte Carlo	0.5373	0.5996	0.7698
Mean field	0.0009	0.0010	0.0011

Table 2: Computational times (in second) for exact, Monte Carlo and mean field EES evaluations, on small graphs. Top: random graphs, bottom: SBS graphs.

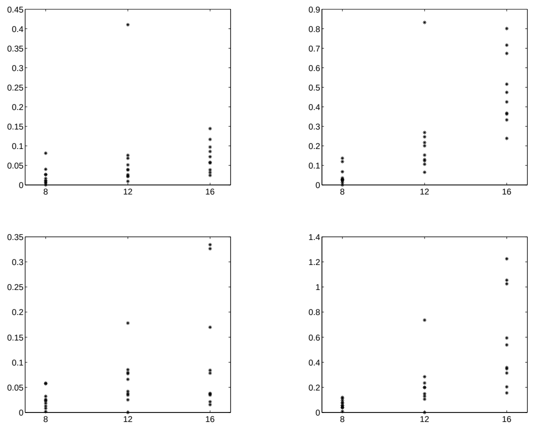

Figure 2: Relative and absolute error between exact and mean field EES evaluations, on graphs of size 8, 12 and 16. Top: random graphs, bottom: SBS graphs. Left: relative error, right: global error.

6.2 LARGE GRAPHS

To evaluate the quality of the mean field approximation on graphs of larger size (up to 5000 vertices), we considered more realistic graphs, namely scale-free graphs. The scale-free distribution of degrees is a feature shared by many real-world graphs of different domains and in particular by those relevant when studying epidemics spreading (social networks, transportation networks, Newman, 2003). We followed the algorithm proposed by Klemm and Eguiluz (2002) to generate these graphs. We set $\rho_{ij} = \rho$ for all vertices and we generated 10 instances $(G, \mathcal{I})$ for different triples of parameters $(\rho, pI0, n)$, for which the Monte Carlo (2500 trajectories) and the mean field evaluations were computed. Exact evaluation is out of reach for the graphs considered. Results are only reported for the largest tested graphs (5000 vertices).

We observed that for a wide range of values of ρ, from low transmission probability and small size epidemics to high transmission probability almost complete infection of the population, the mean field estimations of EES are very close to the Monte Carlo ones (Figure 3 left). As for small size graphs, we observe large relative errors for low transmission probabilities (Figure 3 right), leading to a small over estimation of the mean field EES estimation.

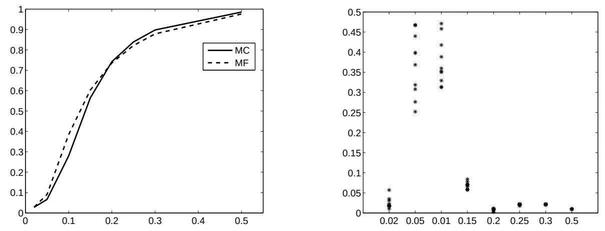

Figure 3: Left: Monte Carlo and mean field EES estimations (as a fraction of the total population of 5000 vertices) for increasing values of ρ. Right: relative error between Monte Carlo and mean field EES estimation for increasing values of ρ.

Running the SIR_EEC_MF algorithm is much faster than performing Monte Carlo simulations. For a fixed number of nodes, running time of the former is independent of ρ, while running time of the latter is longer for intermediate size epidemics (Figure 4 left). In these situations the equilibrium state takes more time steps to be reached. Let us also note that the computational times shown for SIR_EEC_MF could be significantly reduced. Indeed, we recall that SIR_EEC_MF takes as input variable T, an upper bound on the maximum number of steps before the epidemics ends. In practice we set T to an arbitrary constant 50. We observed that even for large graphs, after 50 steps the quantity $\sum_{1\leq t<s} \hat{q}_i^t(X_i^t = I) = \hat{q}_i^s(X_i^s = R)$ has converged (Figure 4 right). Furthermore, convergence is always reached very quickly, after less that 20 iterations.

7 CONCLUSION

In this article, we proved the #P-completeness of the problem of evaluation of the expected cost of an epidemics, spreading on a finite graph according to discrete-time SIR dynamics. We provided an exact solution algorithm as well as an approximation algorithm, based on the mean field principle. In the case of the evaluation of the expected epidemics size, this algorithm was empirically shown to provide estimates close to the exact or Monte Carlo values. However, as is classical with mean field approximations, we observed an overestimation of the epidemics size for low values of transmission proba-

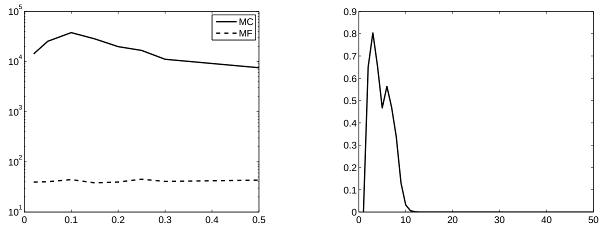

Figure 4: Left: computational time (in second, logarithm scale) for Monte Carlo and mean field EES evaluations, for increasing values of ρ (graph of size 5000). Right: illustration of the evolution of the maximum, over all vertices, of the absolute difference between $\hat{q}_i^t(X_i^t = R)$ and $\hat{q}_i^{t+1}(X_i^{t+1} = R)$ (graph of size 5000).

bility. This behavior could be improved by using higher order approximation schemes, like Bethe approximation and its Machine Learning implementation as message passing algorithms (Yedidia et al., 2005). A potential application for this work is to exploit the fast mean field approximation to design approximate *epidemics control* algorithms, in the spirit of Peyrard and Sabbadin (2006) for general Markov decision processes on graphs, but taking advantage of the simplicity of the SIR model.

8 APPENDIX

Time complexity of the divide and conquer algorithm for particular graphs

We first consider the case where the graph is a tree $\mathcal{T}$, with a single initially infected vertex i_0. In the algorithm SIR_EEC_Exact, Clusters is the set of subtrees $\{\mathcal{T}_i, i \in Children(\mathcal{T}, i_0)\}$, which roots are the children i of i_0. Then,

$$\text{SIR\_EEC\_Exact}(\mathcal{T}, i_0) = c(i_0) + \sum_{i \in Children(\mathcal{T}, i_0)} \rho_{i_0 i} \text{SIR\_EEC\_Exact}(\mathcal{T}_i, i).$$

By a simple inductive reasoning on the depth of the tree $\mathcal{T}$, it is easy to show that the time complexity $T(n) = O(n)$. Indeed, epidemics on trees of depth 0 (corresponding to networks of size 1) can have their EEC computed in constant time. The expected cost of an epidemics on a tree of depth 1, with root vertex infected, can be computed by summing the EEC (weighted by the $\rho_{i_0 i}$) of the children vertices, which can be done in $O(n)$ time. For a tree $\mathcal{T}$ of depth d, each of its subtrees $\mathcal{T}_i$, of size n_i, has at most depth $d-1$, thus by the induction hypothesis, can be evaluated in time $O(n_i)$. The EEC of $\mathcal{T}$ can be computed by summing these values, and so can be computed in $O(1 + \sum_i n_i) = O(n)$.

Now, consider the most unfavorable case for the divide and conquer algorithm, which is the case where the graph is a clique. Let us consider a clique $\mathcal{C}_n$ of size $n+1$, with a single initial infected vertex i_0. The running time $T(n)$ of the algorithm SIR_EEC_Exact, satisfies the recursive equation

$$\begin{aligned} T(n) &= T_C(n) + T_P(n) + \sum_{I \subseteq \mathcal{C}_n, I \neq \emptyset} T(n - |I|), \\ T(n) &= T_C(n) + T_P(n) + \sum_{k=0}^{n-1} C_n^k T(k)., \end{aligned}$$

where C_n^k counts the number of subsets of $\mathcal{C}_n \setminus i_0$ of size k. Noting that the two first terms $T_C(n)$ and $T_P(n)$, which are the times needed to compute Clusters and PotInf respectively, are negligible in front of the third one, we get

$$T(n) = O\left(\sum_{k=0}^{n-1} C_n^k T(k) \right).$$

And since $T(n)$ is obviously increasing,

$$T(n) = O\left(\sum_{k=0}^{n-1} C_n^k T(n-1) \right) = O\left(2^n T(n-1)\right).$$

Finally, by an easy induction,

$$T(n) = O\left(2^n \times 2^{n-1} \times \ldots \times 2^1 \times T(0)\right) = O\left(2^{\frac{n(n+1)}{2}}\right).$$

Derivation of equation (4):

$$
\begin{aligned}
KL(\hat{q}^{t-1}(.)q^t(.|.)) \mid \hat{q}^{t-1}(.)p(.|.)) &= \textstyle\sum_{x^{t-1},x^t} \hat{q}^{t-1}(x^{t-1})q^t(x^t|x^{t-1}) \log \frac{q^t(x^t|x^{t-1})}{p(x^t|x^{t-1})} \\
= \quad & \textstyle\sum_{x^{t-1},x^t} \hat{q}^{t-1}(x^{t-1})q^t(x^t|x^{t-1}) \left(\sum_{i=1}^n \log q_i^t(x_i^t|x_i^{t-1}) - \log p_i(x_i^t|x_i^{t-1}, x_{N(i)}^{t-1}) \right) \\
= \quad & \textstyle\sum_{i=1}^n \Big[\sum_{x_{N(i)}^{t-1},x_i^t} \hat{q}^{t-1}(x_i^{t-1}, x_{N(i)}^{t-1}) q_i^t(x_i^t|x_i^{t-1}) \Big(\log q_i^t(x_i^t|x_i^{t-1}) - \log p_i(x_i^t|x_i^{t-1}, x_{N(i)}^{t-1}) \Big) \Big].
\end{aligned}
$$

We are looking for the minimum of KL, with respect to the variables $q_i^t(x_i^t|x_i^{t-1})$, by solving

$$\frac{\partial KL}{\partial q_i^t(x_i^t|x_i^{t-1})} = 0.$$

By derivation, we get

$$\frac{\partial KL}{\partial q_i^t(x_i^t|x_i^{t-1})} = \sum_{x_{N(i)}^{t-1}} \hat{q}^{t-1}(x_i^{t-1}, x_{N(i)}^{t-1}) \Big(\log q_i^t(x_i^t|x_i^{t-1}) - \log p_i(x_i^t \mid x_i^{t-1}, x_{N(i)}^{t-1}) + 1 \Big).$$

Since $\hat{q}^{t-1}(x_i^{t-1}, x_{N(i)}^{t-1}) = \hat{q}^{t-1}(x_i^{t-1})\hat{q}^{t-1}(x_{N(i)}^{t-1})$, under the independance property of distributions in $\mathcal{Q}$ (independent Markov chains),

$$\frac{\partial KL}{\partial q_i^t(x_i^t|x_i^{t-1})} = \hat{q}_i^{t-1}(x_i^{t-1}) \left[\sum_{x_{N(i)}^{t-1}} \hat{q}^{t-1}(x_{N(i)}^{t-1}) \Big(\log q_i^t(x_i^t|x_i^{t-1}) - \log p_i(x_i^t|x_i^{t-1}, x_{N(i)}^{t-1}) + 1 \Big) \right].$$

And finally

$$\frac{\partial KL}{\partial q_i^t(x_i^t|x_i^{t-1})} = 0 \Leftrightarrow \log q_i^t(x_i^t|x_i^{t-1}) = E_{\hat{q}^{t-1}} \Big[\log p_i(x_i^t \mid x_i^{t-1}, X_{N(i)}^{t-1}) \Big] - 1,$$

from which we get an iterative mean field solution of the form of equation (4).

Derivation of equation (6):

$$
\begin{aligned}
& \hat{q}_i^t(X_i^t = I \mid X_i^{t-1} = S) \\
= \; & 1 - \sum_{x_{N(i)}} \hat{q}_{N(i)}^{t-1}(X_{N(i)}^{t-1} = x_{N(i)})(1-\rho)^{NI(x_{N(i)})} \\
= \; & 1 - \sum_{x_{N(i)}} \prod_{j \in N(i)} [(1-\rho)^{\delta_I(x_j)} \hat{q}_j^{t-1}(X_j^{t-1} = x_j)] \\
= \; & 1 - \prod_{j \in N(i)} \sum_{x_j} [(1-\rho)^{NI(x_j)} \hat{q}_j^{t-1}(X_j^{t-1} = x_j)] \\
= \; & 1 - \prod_{j \in N(i)} [(1-\rho)\hat{q}_j^{t-1}(X_j^{t-1} = I) \\
& \qquad\qquad + \hat{q}_j^{t-1}(X_j^{t-1} \neq I)] \\
= \; & 1 - \prod_{j \in N(i)} (1 - \rho \, \hat{q}_j^{t-1}(X_j^{t-1} = I))
\end{aligned}
$$

where $NI(x_{N(i)}) = \sum_{j \in N(i)} \mathbb{1}_{\{x_j = I\}}$ is the number of vertices infected in state $x_{N(i)}$.

References

Ball, M. O. (1980). Complexity of network reliability computation. *Networks 10*, 153–165.

Ball, M. O., C. J. Colbourn, and J. S. Provan (1992). Network reliability. Technical Report TR 92-74, Harvard University.

Chandler, D. (1987). *Introduction to Modern Statistical Mechanics*. Oxford University Press.

Ferrari, M. J., S. Bansal, L. A. Meyers, and O. N. Bjørnstad (2006). Network frailty and the geometry of herd immunity. *Proceedings of the Royal Society of London Series B—Biological Sciences 273*(1602), 2743–2748.

House, T. and M. J. Keeling (2011). Epidemic prediction and control in clustered populations. *Journal of Theoretical Biology 272*(1), 1 – 7.

Klemm, K. and V. M. Eguiluz (2002). Highly clustered scale-free networks. *Phyical Review E 65*.

Neal, P. (2003). SIR epidemics on a Bernoulli random graph. *Journal of Applied Probability 40*(3), 779–782.

Newman, M. (2002). Spread of epidemic disease on networks. *Physical Review E 66*.

Newman, M. (2003). The structure and function of complex networks. *SIAM Review 45*, 167–256.

Nowicki, K. and T. Snijders (2001). Estimation and prediction for stochastic blockstructures. *Journal of the American Statistical Association 96*(455), 1077–1087.

Peyrard, N., U. Dieckmann, and A. Franc (2008). Long-range correlations improve understanding the influence of network structure on per contact dynamics. *Theoretical Population Biology* (3), 383–394.

Peyrard, N. and R. Sabbadin (2006). Mean field approximation of the Policy Iteration algorithm for graph-based Markov decision processes. In *17th European Conference on Artificial Intelligence (ECAI'06)*, Riva del Garda, Italy, pp. 595–599.

Salathé, M. and J. Jones (2010). Dynamics and control of diseases in networks with community structure. *PLoS Computational Biology 6*(4).

Yedidia, J. S., W. T. Freeman, and Y. Weiss (2005). Constructing free-energy approximations and generalized belief propagation algorithms. *Information Theory, IEEE Transactions on 51*(7), 2282–2312.

```
Algorithm: SIR_EEC_MF
Data  : $\{G = (V, E), \mathcal{I}, \mathcal{S}, \rho, c, T\}$
Result: $\{EEC\}$
% Initialization;
for $i \in \mathcal{S}$ do
    $Q_i^1(I) \leftarrow [1 - \prod_{j \in N(i)}(1 - \rho_{ji})]$;
    $Q_i^t(S) \leftarrow [\prod_{j \in N(i)}(1 - \rho_{ji})]$;
    $EEC_i \leftarrow Q_i^1(I)$;
end
for $i \in \mathcal{I}$ do
    $Q_i^1(I) \leftarrow 0$;
end
% Main loop;
for $t = 2$ $to$ $t = T$ do
    for $i \in \mathcal{S}$ do
        $Q_i^t(I) \leftarrow [1 - \prod_{j \in N(i)}(1 - \rho_{ji} Q_j^{t-1}(I))] Q_i^{t-1}(S)$;
        $Q_i^t(S) \leftarrow [\prod_{j \in N(i)}(1 - \rho_{ji} Q_j^{t-1}(I))] Q_i^{t-1}(S)$;
        $EEC_i \leftarrow EEC_i + Q_i^t(I)$;
    end
    for $i \in \mathcal{I}$ do
        $Q_i^t(I) \leftarrow 0$;
    end
end
$EEC \leftarrow c\mathcal{I} + \sum_{i \in \mathcal{S}} c_i EEC_i$ ;
```

Algorithm 2: SIR_EEC_MF