Containing Epidemic Outbreaks by Message-Passing Techniques

Containing Epidemic Outbreaks by Message-Passing Techniques

F. Altarelli,[1,2,*] A. Braunstein,[1,3,2,†] L. Dall'Asta,[1,2,‡] J. R. Wakeling,[1,§] and R. Zecchina[1,3,2,¶]

[1]*Department of Applied Science and Technology, Politecnico di Torino, Corso Duca degli Abruzzi 24, 10129 Torino, Italy*

[2]*Collegio Carlo Alberto, Via Real Collegio 30, 10024 Moncalieri, Italy*

[3]*Human Genetics Foundation, Via Nizza 52, 10126 Torino, Italy*

(Received 11 September 2013; revised manuscript received 16 January 2014; published 8 May 2014)

The problem of targeted network immunization can be defined as the one of finding a subset of nodes in a network to immunize or vaccinate in order to minimize a tradeoff between the cost of vaccination and the final (stationary) expected infection under a given epidemic model. Although computing the expected infection is a hard computational problem, simple and efficient mean-field approximations have been put forward in the literature in recent years. The optimization problem can be recast into a constrained one in which the constraints enforce local mean-field equations describing the average stationary state of the epidemic process. For a wide class of epidemic models, including the susceptible-infected-removed and the susceptible-infected-susceptible models, we define a message-passing approach to network immunization that allows us to study the statistical properties of epidemic outbreaks in the presence of immunized nodes as well as to find (nearly) optimal immunization sets for a given choice of parameters and costs. The algorithm scales linearly with the size of the graph, and it can be made efficient even on large networks. We compare its performance with topologically based heuristics, greedy methods, and simulated annealing on both random graphs and real-world networks.

DOI: 10.1103/PhysRevX.4.021024

Subject Areas: Complex Systems, Interdisciplinary Physics, Statistical Physics

I. INTRODUCTION

One of the key questions of computational epidemiology is how best to distribute limited resources of treatment and vaccination so that they will be most effective in suppressing or reducing outbreaks of disease. This problem is heightened by the entangled networks of interactions via which diseases can spread: In a large complex network, the existence of high-degree hubs can help a virus spread rapidly throughout the population even if the probability of transmission from an individual contact is low. In particular, it is known that the epidemic spread undergoes a critical threshold phenomenon similar to percolation, with the threshold approaching zero for very heterogeneous (scale-free) networks [1,2]. Early works on network immunization drew attention to the differences between random immunization and targeted immunization strategies [3–6].

*fabrizio.altarelli@polito.it
†alfredo.braunstein@polito.it
‡luca.dallasta@polito.it
§joseph.wakeling@polito.it
¶riccardo.zecchina@polito.it

A simple random immunization strategy can consist in fixing a fraction or a density of immunized nodes and averaging the outcome of the epidemic process over all possible realizations of the immunization set. On the contrary, targeted immunization strategies correlate the choice of immunized nodes with some topological feature, such as the degree or other centrality measures. This can be experimentally shown to have some positive effect in reducing the spread of diseases [3–6]. Most topologically based algorithms for immunization follow an incremental procedure, in which the set of immunized nodes is initially empty and then it is progressively filled, adding, one by one, the nodes that are most relevant with respect to a particular topological metric. Despite the computational cost, recalculating the topological metric after each immunization step (i.e., after removing the immunized node from the graph) usually provides much better results than computing it only once on the original graph [4]. Further improvements were obtained by means of more complex immunization strategies, based on graph partitioning [7,8] and on the optimization of the susceptible size [9]. Besides the heterogeneity of contacts, clustering, community structure, and modularity also have a major impact on disease dynamics [10,11]; therefore, the same immunization strategy can produce contrasting results on networks with different topological features. This is a consequence of the fact that topological heuristic methods neglect important features of the spreading rule, and most

common metrics used to measure their effectiveness, such as the largest connected component or the largest nonimmune cluster size, are proxies that may not reflect the true susceptibility to an epidemic. These techniques also neglect the cost of vaccination, which may vary widely depending on the chosen target.

To overcome these limitations, several authors tried to quantify more explicitly the effects of immunization strategies on the outbreak dynamics [12–17], and network immunization was mathematically formulated as a proper optimization problem that can be proven to be NP hard in a plethora of different variants [16,18–25]. Standard optimization techniques such as Monte Carlo (MC) methods or integer/linear programming are computationally very expensive and may take a prohibitive amount of time to reach reasonably good results even on relatively small networks. On the other hand, the *greedy* optimization strategies usually proposed are guaranteed to approximate the optimal result by a constant factor only in some fortunate case [16,18,25].

Recent progress in combinatorial optimization have shown that algorithms based on the message-passing principle, and developed using methods from the statistical physics of disordered systems, outperform, in many cases, both greedy algorithms and simulated annealing, even in a complex optimization problem involving stochastic parameters [26,27] and dynamical rules [28,29]. In many cases in which MC algorithms get trapped in local minima of the (free-)energy function, message-passing algorithms can find considerably better results. The remarkable performances of these algorithms are combined with considerably good computation time scaling properties. While on a wide variety of optimization problems the computational complexity of simulated annealing scales exponentially with the system size, message-passing algorithms typically require a time that scales roughly linearly with the number of messages (i.e., the number of edges).

In this paper, we show that, under some approximations, network immunization can be written as a constrained optimization problem, in which the constraints are fixed-point equations for some local (node or edge) variables describing the average stationary state of the dynamics. These constraints and a suitably defined objective (energy) function are then used to derive a message-passing approach to the optimization problem and to design efficient algorithms on large networks. We apply this method to find optimal immunization strategies for both susceptible-infected-recovered (SIR) and susceptible-infected-susceptible (SIS) models.

In Sec. II, we recall the main ideas and formulas of mean-field methods in epidemic models, which are usually used to estimate the average stationary properties of an epidemic outbreak. The optimal immunization problem is introduced in Sec. III, opportunely defined in terms of mean-field quantities. Section IV is devoted to the definition of the message-passing approach and the derivation of the corresponding belief-propagation (BP) and max-sum (MS) equations. In Sec. V, we use BP to understand in detail the immunization properties on the prototypical case of random regular graphs. The comparison with other optimization methods on more general graphs, including random graphs and real-world networks, is discussed in Sec. VI.

II. MEAN-FIELD METHODS IN EPIDEMIC MODELS

Over the years, a large number of stochastic epidemic models have been introduced, with the aim of addressing some specific features of different diseases [30,31]. In the most simple model, the epidemic spreading induces, in the nodes, irreversible stochastic transitions from a susceptible state to an infected one. Infected individuals can recover by either returning to the susceptible state or becoming permanently resistant to the disease. One can then increase the complexity of the stochastic model by introducing other intermediate states, or compartments, such as exposure and latency. In the following, we discuss the most basic models of epidemic spreading, providing for each of them a set of approximated equations of mean-field type that are valid on very general graph structures. Their solution describes the statistical properties of the stationary state corresponding to a given set of initial conditions and external parameters. In addition, such equations allow us to measure the level of infection once a configuration of initially immunized nodes is chosen.

A. Irreversible epidemic processes

The susceptible-infected-recovered (SIR) model was formulated by Kermack and McKendrick [32] to describe the irreversible propagation through a population of individuals of an infectious disease, such as measles, mumps, or cholera. The SIR stochastic dynamics is defined over a graph $G = (V, E)$, representing the contact network of a set V of individuals. At any given time step t (e.g., a day), a node i can be in one of three states: susceptible ($\mathcal{S}$), infected ($\mathcal{I}$), and recovered/removed ($\mathcal{R}$). The state of node i at time t is represented by a variable $x_i^t \in \{\mathcal{S}, \mathcal{I}, \mathcal{R}\}$. We assume that each node i is initially infected with probability $q_i \in [0, 1]$ (independent of the other nodes). At each time step, an infected node i can first spread the disease to each susceptible neighbor j with given probability $T_{ij} \in (0, 1]$, and then recover with probability r_i. Once recovered, individuals do not get sick anymore (they are effectively removed from the graph). The probability that an infected node i directly transmits the disease to j before i recovers is given by

$$p_{ij} = 1 - \sum_{t=1}^{\infty} (1 - T_{ij})^t r_i (1 - r_i)^{t-1} = \frac{T_{ij}}{T_{ij} + (1 - T_{ij}) r_i}. \quad (1)$$

It is thus possible to construct a completely static representation of the process that maps the final state onto the outcome of a bond percolation process [33]. This relationship can be made mathematically clear as follows. Let us consider a treelike graph and define m_{ij} to be the probability that node i is eventually infected when considering the graph obtained in the absence of the neighboring node j. Exploiting the factorization of probabilities on the sub-branches of the tree emerging from i, the quantity m_{ij} satisfies the equation

$$m_{ij} = q_i + (1 - q_i)\left[1 - \prod_{k \in \partial i \setminus j} (1 - p_{ki} m_{ki})\right], \quad (2)$$

where ∂i denotes the set of neighbors of i. Since infected nodes eventually recover, in the final state nodes can only be either in the susceptible state or in the recovered one. From the knowledge of the conditional marginals m_{ij}, one can compute the probability m_i that a node i is eventually infected, i.e., the probability that i is recovered in the final state,

$$m_i = q_i + (1 - q_i)\left[1 - \prod_{k \in \partial i} (1 - p_{ki} m_{ki})\right]. \quad (3)$$

Although Eqs. (2) and (3) are exact only on trees, they have also been successfully applied to study the SIR model on general random graphs [33–35]. A comparison between the solutions of these equations and the results of simulations of the SIR stochastic process is shown in Fig. 1 for a random regular graph (RRG) of $N = 10^3$ nodes and degree $K = 4$. For simplicity, we considered uniform self-infection probabilities $q_i = q$, $\forall\ i \in V$ and uniform transmission probabilities $p_{ij} = p$, $\forall\ (i, j) \in E$. In the SIR stochastic process, we defined a measure of the "outbreak" size as the average fraction f of nodes that have been infected during the epidemic spreading. Since all infected nodes eventually recover, this metric can also be defined as

$$f = \frac{1}{N} \sum_i \Pr\left[x_i^{\infty} = \mathcal{R}\right], \quad (4)$$

where $\Pr\left[x_i^{\infty} = \mathcal{R}\right]$ is the probability that node i is eventually recovered. In Fig. 1, f is plotted as a function of the transmission probability p for $q = 0.1,\ 0.01,\ 0.001$. The results, obtained by averaging over 10^4 realizations of the stochastic process, are reported as black circles, while the symmetric bars indicate the fluctuations around the average behavior. The average behavior can also be computed from the solutions of Eqs. (2) and (3), exploiting the fact that, in the treelike approximation, $\Pr\left[x_i^{\infty} = \mathcal{R}\right] \simeq m_i$. The results are reported as a red solid line in Fig. 1. The agreement between the mean-field theory represented by Eqs. (2) and (3) and the simulations is very good for sufficiently large values of q, and then it deteriorates for q of the order of $1/N$ and large values of p. The reason for such a discrepancy is that Eqs. (2) and (3) are correct on treelike structures, i.e., when the disease transmission events to one node coming from two neighbors are not correlated. The "decorrelation" assumption is not correct when the actual number of sources of spontaneous infection is very small. This is obvious in the case of a unique source of infection: The contagion path has the same source; hence, the infection of a node due to disease transmission from her neighbors is a highly correlated process that is not well captured by Eqs. (2) and (3). More precisely, the solution to Eqs. (2) and (3) gives an upper bound for the real probability to be infected [35]. In the limit of infinitely large networks, this approach is expected to provide a correct description of the average final state of the system for any finite value of q and p.

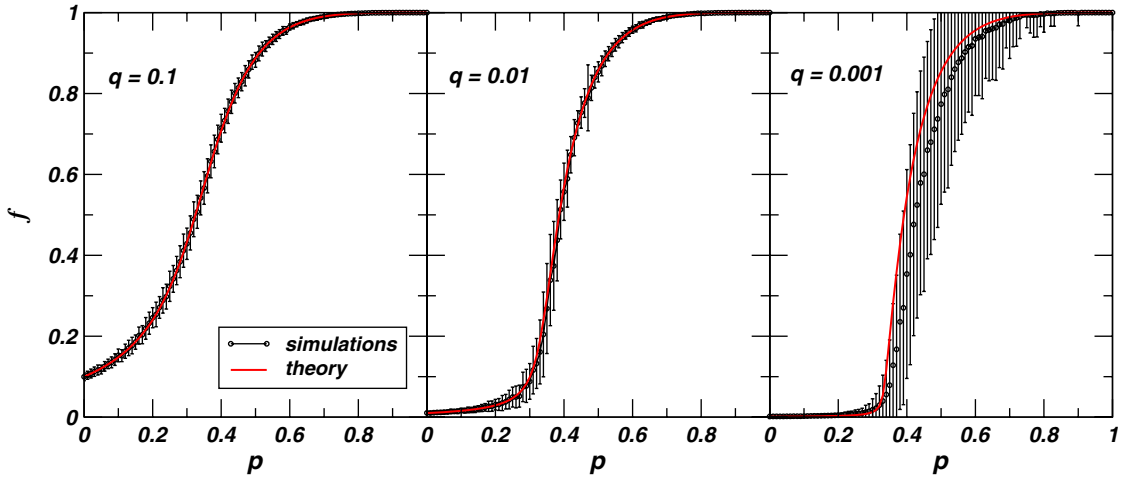

FIG. 1. Black circles represent the average density f of nodes that have been infected by the final (infinite) time in the SIR dynamics on a random regular graph of $N = 10^3$ nodes and degree $K = 4$, as a function of the transmission probability p for $q = 0.1, 0.01, 0.001$ (from left to right). The symmetric bars indicate the fluctuations around the average value computed on 10^4 realizations of the stochastic process. The red solid line is computed from the solution of Eqs. (2) and (3).

Equations (2) and (3) can be easily modified to include immunization of nodes. By considering a set of binary variables $s_i \in \{0,1\}$, in which $s_i = 1$ if node i is immune to the disease, we get $\forall\ i \in V$ and $\forall\ (i,j) \in E$,

$$m_{ij} = (1-s_i)\left\{q_i + (1-q_i)\left[1 - \prod_{k\in\partial i\setminus j}(1-p_{ki}m_{ki})\right]\right\}, \tag{5a}$$

$$m_i = (1-s_i)\left\{q_i + (1-q_i)\left[1 - \prod_{k\in\partial i}(1-p_{ki}m_{ki})\right]\right\}. \tag{5b}$$

Given a configuration $\mathbf{s} = \{s_1, \ldots, s_N\}$ of immune nodes, which we call the immunization set, the solution of Eqs. (5a) and (5b) provides a measure of the corresponding epidemic outbreak. It is possible to show that for a given set of parameters $\{q_i\}$ and $\{p_{ij}\}$, the solution of Eqs. (2) and (3) is unique; therefore, each configuration of immune nodes $\mathbf{s}$ corresponds to a unique solution of Eqs. (5a) and (5b). This property will be crucial for the validity of the optimization method developed in this work.

B. Reversible epidemic processes

The susceptible-infected-susceptible (SIS) model is the prototype of reversible models of epidemic spreading, in which after recovery a node is again susceptible to being infected [30,31]. The state of node i at time t is now represented by a binary variable $x_i^t \in \{\mathcal{S}, \mathcal{I}\}$. At each time step, an infected node i can transmit the disease to each of its susceptible neighbors j with probability p_{ij}, while it recovers with rate r_i (becoming susceptible again). The stochastic process admits an absorbing state in which all nodes are susceptible and the disease has disappeared from the population. When the transmission probabilities are sufficiently large, an active stationary state also exists that is metastable and attractive for the dynamical process. Although in any finite population a fluctuation will eventually bring the system into the absorbing state, the lifetime of the metastable endemic state scales with the size of the graph in such a way that an absorbing phase transition as a function of the transmission probabilities is expected to occur in the thermodynamic limit [36–40]. The critical threshold usually depends on the parameters of the dynamical process as well as on the topological structure of the underlying interaction graph. A variant of the SIS model with spontaneous self-infection was recently introduced [41,42] in order to simplify the numerical and mathematical analysis of the model. The presence of spontaneous self-infection destroys the absorbing state, whereas the metastable state becomes the (unique) stationary state of the dynamics. On the other hand, for a given small self-infection probability, the dynamics shows a clear boundary between the low-infection region and a region of global spreading as a function of the transmission probabilities. By scaling down self-infection, one can extrapolate information on the epidemic phase transition occurring for zero self-infection, avoiding the problems associated with the existence of an absorbing state.

The SIS model on a given graph with N nodes is a Markov chain with 2^N states, whose stationary probability distribution cannot be explicitly computed for large systems. A simple mean-field approximation that provides a good qualitative and quantitative description of the stationary state of the SIS model is obtained by replacing the exact probability distribution by a product measure over the nodes of the graph [38–40,43–45]. This factorization, also known as the N-intertwined model, leads to a set of "quenched" mean-field equations for the evolution of single-node infection probabilities m_i with $i \in V$. Notice the difference between the SIR and SIS cases: While in the SIR model m_i indicates the (mean-field) probability that node i is eventually infected before the final state, in the SIS model it represents the (mean-field) probability that i is infected in the stationary state of the dynamics. In the quenched mean-field approximation, the infection probability of node i at time t satisfies the equation

$$m_i^{t+1} = (1-r_i)m_i^t + (1-m_i^t)\left\{q_i + (1-q_i)\left[1 - \prod_{j\in\partial i}(1-p_{ji}m_j^t)\right]\right\}, \tag{6}$$

where p_{ji} is the transmission probability from j to i, and q_i is the spontaneous self-infection probability. In the stationary state, the mean-field variables $\{m_i\}_{i\in V}$ are given by the solution of the fixed-point equations

$$m_i = (1-s_i)\frac{q_i + (1-q_i)[1-\prod_{j\in\partial i}(1-p_{ji}m_j)]}{r_i + q_i + (1-q_i)[1-\prod_{j\in\partial i}(1-p_{ji}m_j)]}, \tag{7}$$

where $s_i \in \{0,1\}$ says whether node i is immune or not. As for the SIR model, it is possible to show that the mean-field quantity m_i gives an upper bound for the real value of the infection probability of node i in the stationary state [38–40]. The approximation can be improved by considering second-order quantities, i.e., deriving closed equations for single-point marginals and pair correlations, but the actual form of these equations is not unique and depends on the moment closure approximation adopted [38–40]. Nevertheless, in most cases, Eq. (7) already provides a very good description of the stationary state of the SIS stochastic process.

A measure of the outbreak size of the epidemics is given by the average fraction f of infected nodes in the active stationary state. If the stationary state is infinitely long-lived, f can be operatively defined as

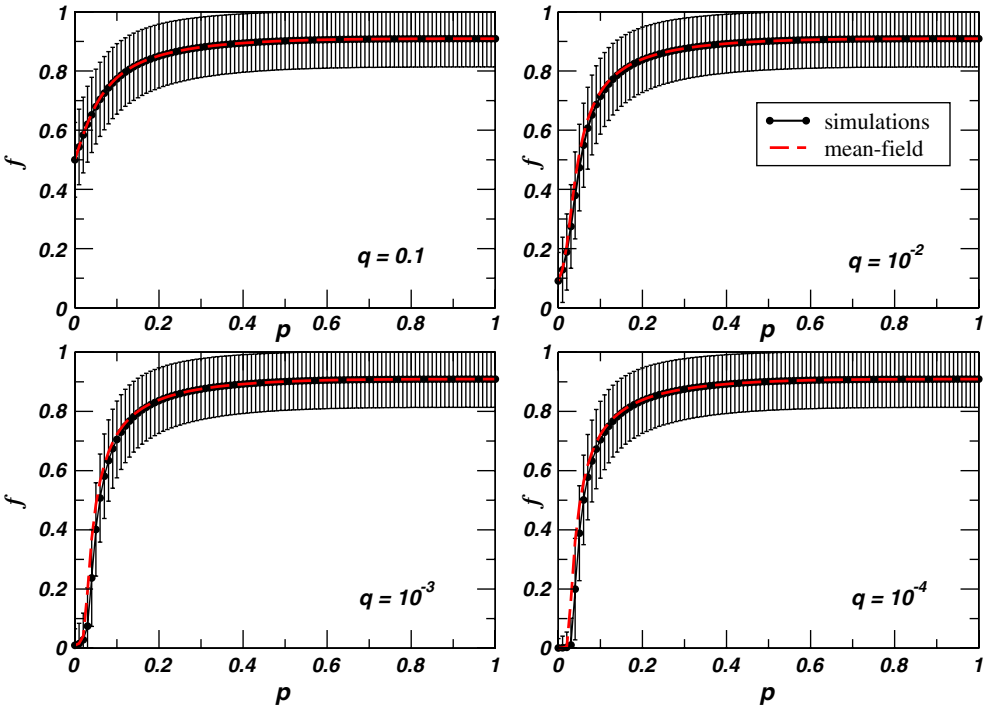

FIG. 2. Black circles represent the average density f of infected nodes in the stationary state ($\tau = 10^4$ steps) of the SIS dynamics on a random regular graph of $N = 10^3$ nodes and degree $K = 4$, as a function of the transmission probability p for q between 10^{-1} and 10^{-4}. The symmetric bars indicate the fluctuations around the average value computed on 10^4 realizations of the stochastic process. The red dashed line is computed from the solution of Eq. (7).

$$f = \frac{1}{N}\sum_i \left\{ \lim_{\tau\to\infty} \frac{1}{\tau}\sum_{t=0}^{\tau} \mathbb{1}[x_i^t = \mathcal{I}] \right\}, \tag{8}$$

where $\mathbb{1}[\cdot]$ is equal to 1 when the argument is verified and 0 otherwise. In practice, simulations are performed for a finite time that is chosen to be much longer than the time necessary to converge to the stationary state. Figure 2 displays the behavior of f as a function of the transmission probabilities p, for different values of the self-infection probability q, on a random regular graph of $N = 10^3$ nodes and degree $K = 4$. The dashed line is the same quantity computed as $f \simeq \frac{1}{N}\sum_{i\in V} m_i$, i.e., approximating the probability that node i is infected in the stationary state by the mean-field one m_i obtained by solving Eq. (7). The agreement between mean-field predictions and results of the simulations is very good in almost all regimes of the parameters.

C. Continuous-time processes

Both SIR and SIS epidemic models are often formulated in continuous time, where spontaneous self-infection probabilities and transmission probabilities are replaced by rates. While the time-dependent evolution of the dynamical processes can be quite different from that of their discrete-time counterparts, the stationary behavior at long times can be similarly described in terms of percolation-like equations. In continuous time, transmission events can be modeled as independent Poisson processes; therefore, the probability that, in a sufficiently small interval of time, a node is infected by the neighbors is proportional to the sum of the independent probabilities of the individual transmission events. Once $\{q_i\}_{i\in V}$ and $\{p_{ij}\}_{(i,j)\in E}$ are intended as probability rates, this reduces to replacing the term $1 - \prod_j (1 - p_{ji} m_j)$ with $\sum_j p_{ji} m_j$ in both Eqs. (5a) and (5b) and Eq. (7).

III. THE OPTIMAL IMMUNIZATION PROBLEM

Preventing or eradicating diseases entails a tradeoff between the costs of treating and hospitalizing infected individuals and the cost of distributing vaccines or drugs. When the contact network is known and these costs can be estimated, one can devise an optimization problem, in which the optimal immunization set is the configuration of immunized nodes that minimizes a properly defined energy function. In the SIR stochastic process, we consider the following energy function:

$$\mathcal{E}_{\mathrm{SIR}}(\mathbf{s}) = \mu \sum_{i\in V} s_i c_i + \epsilon \sum_{i\in V} \ell_i \Pr[x_i^\infty = \mathcal{R}|\mathbf{s}], \tag{9}$$

in which $c_i \in \mathbb{R}$ is the cost of immunization of node i, ℓ_i is a loss (i.e., the cost associated with the infection of node i), and $\Pr[x_i^\infty = \mathcal{R}|\mathbf{s}]$ is the probability that node i eventually recovers (i.e., the probability that the node has been

infected during the epidemic spreading), given the configuration $\mathbf{s}$ of immunized nodes. Estimating the probability $\Pr[x_i^\infty = \mathcal{R}|\mathbf{s}]$ from the simulations of the stochastic SIR process is very cumbersome, making the optimization problem practically unsolvable for sufficiently large graphs. In fact, it is known that finding the probability that a node becomes infected in the course of an epidemic in SIR and related models is an NP-hard problem [46]. The energy function for the optimal immunization problem for the SIS model can be defined in a similar way as

$$\mathcal{E}_{\mathrm{SIS}}(\mathbf{s}) = \mu \sum_{i\in V} s_i c_i + \epsilon \sum_{i\in V} \ell_i \lim_{\tau\to\infty} \frac{1}{\tau} \sum_{t=0}^{\tau} \mathbb{1}[x_i^t = \mathcal{I}]. \quad (10)$$

Computing $\mathcal{E}_{\mathrm{SIS}}(\mathbf{s})$ from direct simulations of the stochastic dynamics is also a very demanding task.

The problem of finding the configuration of seeds that minimizes an energy function associated with a given dynamical model has been investigated by several authors in the recent past. Even for simple stochastic propagation models, such as the independent cascade model [19], several versions of this optimization problem have been introduced and shown to be NP hard. To the best of our knowledge, the methods proposed to solve this class of problems are mostly based on greedy approaches that scale reasonably well with the system size but give solutions that only provide good approximations of the real optima in very peculiar cases.

In this work, we employ a simplified approach, based on the mean-field representation of the dynamical process. We replace the energy functions $\mathcal{E}_{\mathrm{SIR}}$ and $\mathcal{E}_{\mathrm{SIS}}$ with a unique energy function defined as

$$\mathcal{E}(\mathbf{s},\mathbf{m}) = \mu \sum_{i\in V} s_i c_i + \epsilon \sum_{i\in V} m_i, \quad (11)$$

in which $\forall\ i \in V$, m_i is the solution of Eqs. (5a) and (5b) and Eq. (7) for the SIR and SIS models respectively.

This energy can now be computed easily for any configuration $\mathbf{s}$ of immunized nodes, solving the corresponding set of self-consistent equations for $\{m_i\}$. Once we know how to compute the energy function, the optimization problem can be studied with different methods. The simplest optimization method is a greedy algorithm that resembles the incremental procedure used in topologically based heuristics. Starting from an empty set of immunized nodes, the greedy principle imposes selecting the node whose immunization causes the largest drop in the energy function. Once the node is included in the set of immunized nodes, a second node is chosen following the same greedy principle. One by one, all nodes are added to the immunization set. Under this criterion, the best configuration of immunized nodes is the one that realizes the lowest energy. The greedy algorithm is fast, but it presents several drawbacks: first, there is no guarantee that the best solution found is the optimal one; moreover, we cannot stop the greedy procedure before almost all nodes are added to the immunization set because the energy function can display several local minima during this process. Another famous optimization method is simulated annealing, in which a MC algorithm is used to find the configuration $\mathbf{s}$ minimizing the energy $\mathcal{E}(\mathbf{s})$. For a sufficiently slow annealing schedule, the algorithm is guaranteed to find the minimum, i.e., the optimal immunization set. However, this can be computationally unfeasible in large graphs, since MC algorithms get trapped in local minima of the (free-)energy function and the convergence to the global one requires a time that can scale exponentially with the system size. In the next section, we will develop a message-passing approach to the optimal immunization problem. Message-passing algorithms can find the global optimum or, in general, considerably better results than simulated annealing in a running time that scales only linearly with the number of messages (i.e., the number of edges). The implementation details of the greedy algorithm and simulated annealing are described in Appendix A.

IV. BELIEF-PROPAGATION APPROACH TO THE OPTIMAL IMMUNIZATION PROBLEM

Since each configuration of immunized nodes corresponds to a unique value of the energy function, we can define a probability function $Q(\mathbf{s})$ by associating to each configuration $\mathbf{s}$ a Boltzmann weight $e^{-\beta\mathcal{E}(\mathbf{s},\mathbf{m})}$ and tracing over the variables $\mathbf{m}$ with the constraint that they are a solution of the corresponding percolation-like equations. It follows that Eqs. (5a) and (7) become a set of hard constraints for the auxiliary variables $\mathbf{m}$ that must be included in the optimization problem in addition to the energetic terms. As for standard constraint-satisfaction problems over discrete variables (i.e., $\mathbf{s}$) defined on the vertices of a graph, it is possible to apply the cavity method [47,48] and to develop efficient message-passing algorithms, such as BP and MS.

In the following, we present the derivation of the BP and MS equations both for the SIR and SIS models; we discuss the approximation methods employed and the convolution technique used to solve them efficiently.

A. Derivation for the SIR model

In the SIR model, Eq. (5a) defines the values assumed by the auxiliary variables $\mathbf{m}$ for each configuration of immunized nodes $\mathbf{s}$. The relevant variables on which we must trace to compute the probability weight associated with configuration $\mathbf{s}$ are $\{m_{ij}\}_{(i,j)\in E}$, with the condition that they satisfy Eq. (5a). Let us consider the partition function of the problem,

$$Z = \sum_{\mathbf{s}} \int d\mathbf{m} e^{-\beta \mathcal{E}(\mathbf{s},\mathbf{m})} \prod_{(i,j)\in E} \Psi_{ij}(s_i, m_{ij}, \{m_{ki}\}_{k\in\partial i\setminus j}), \tag{12}$$

where the integration over $d\mathbf{m}$ selects (for each $\mathbf{s}$) the unique set of values of $\mathbf{m}$ that satisfies the constraints

$$\Psi_{ij}(s_i, m_{ij}, \{m_k\}_{k\in\partial i\setminus j}) = \delta\left(m_{ij} - (1-s_i)\left\{1 - (1-q_i)\prod_{k\in\partial i\setminus j}(1-p_{ki}m_{ki})\right\}\right) \tag{13}$$

and where the energy is

$$\begin{aligned}\mathcal{E}(\mathbf{s},\mathbf{m}) &= \sum_i \mathcal{E}_i(s_i, m_i) \\ &= \mu \sum_i s_i c_i + \epsilon \sum_i \ell_i m_i \\ &= \mu \sum_i s_i c_i + \epsilon \sum_i \ell_i (1-s_i)\left\{q_i + (1-q_i)\left[1 - \prod_{k\in\partial i}(1-p_{ki}m_{ki})\right]\right\}.\end{aligned}$$

The probability associated with a configuration of immunized nodes $\mathbf{s}$ is

$$Q(\mathbf{s}) = \frac{1}{Z}\int d\mathbf{m} e^{-\beta \mathcal{E}(\mathbf{s},\mathbf{m})} \prod_{(i,j)\in E} \Psi_{ij}(s_i, m_{ij}, \{m_{ki}\}_{k\in\partial i\setminus j}). \tag{14}$$

In order to derive the BP equations, we consider an infinite tree, and we marginalize over all variables but j. In this way, we obtain the probability $Q_j(s_j)$ for the variable s_j,

$$Q_j(s_j) \propto \prod_{i\in\partial j}\int_0^1 dm_{ij}\int_0^1 dm_{ji} e^{-\beta \mathcal{E}_j(s_j,m_j)} \prod_{i\in\partial j}\Psi_{ji}(s_j, m_{ji}, \{m_{kj}\}_{k\in\partial j\setminus i}) \prod_{i\in\partial j} P_{ij}(m_{ij}, m_{ji}), \tag{15}$$

in which the factorization over the neighbors i of j comes from the fact that in a tree, each sub-branch is independent of the others, and the partial partition function of the sub-branch is represented by the "cavity marginal" or BP message $P_{ij}(m_{ij}, m_{ji})$. The latter denotes the joint probability for the auxiliary variables m_{ij} and m_{ji} in the absence of the hard constraints over node j (i.e., $\prod_{i\in\partial j}\Psi_{ji}$) and of the energetic term $\mathcal{E}_j$; they satisfy the BP equations

$$P_{ij}(m_{ij}, m_{ji}) \propto \sum_{s_i} \prod_{k\in\partial i\setminus j}\int_0^1 dm_{ki}\int_0^1 dm_{ik} e^{-\beta \mathcal{E}_i(s_i,m_i)} \prod_{\ell\in\partial i}\Psi_{i\ell}(s_i, m_{i\ell}, \{m_{\ell' i}\}_{\ell'\in\partial i\setminus\ell}) \prod_{k\in\partial i\setminus j} P_{ki}(m_{ki}, m_{ik}). \tag{16}$$

Unlike most common cases, the BP message P_{ij} depends on both auxiliary variables defined in the edge (i, j). This happens because both of them enter in the definition of the energetic term (in fact, they are both contributing to m_j). On the contrary, the set of hard constraints $\prod_{i\in\partial j}\Psi_{ji}$ does not depend on the variable m_{ij}. It follows that in cases in which the energetic term $\mathcal{E}$ does not depend on m_j [e.g., for $\epsilon = 0$ in Eq. (14)], it can be shown that P_{ij} only depends on m_{ij}.

From the BP marginals, we can compute the single-site total probability marginal $P_i(m_i)$,

$$P_i(m_i) = \sum_{s_i} \prod_{k\in\partial i}\int_0^1 dm_{ki}\int_0^1 dm_{ik} e^{-\beta \mathcal{E}_i(s_i,\{m_{ki}\})} \prod_{k\in\partial i}\Psi_{ik}(s_i, m_{ik}, \{m_{k' i}\}_{k'\in\partial i\setminus k}) \prod_{k\in\partial i} P_{ki}(m_{ki}, m_{ik}), \tag{17}$$

that represents the probability distribution of the values assumed by the variable m_i on that node. The update rule of the BP equations (16) can be represented graphically on a factor graph as shown in Fig. 3.

The cavity message $P_{ij}(m_{ij}, m_{ji})$ is a real function defined on the square $[0, 1] \times [0, 1]$. Since there is no information on the shape of this function, we proceed by discretizing the interval [0, 1] in a number N_B of bins and solving Eq. (16) numerically by assuming that $P_{ij}(m_{ij}, m_{ji})$ is defined on a two-dimensional mesh of $N_B \times N_B$ identical cells. Because of discretization, the hard constraints in Eq. (16) are only approximately satisfied. In

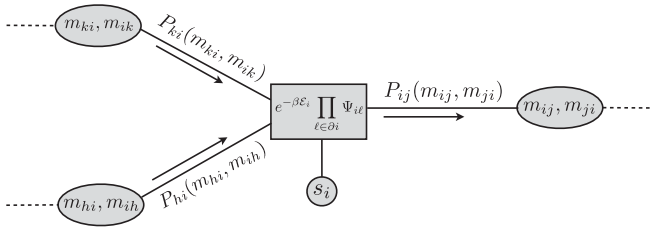

FIG. 3. Factor-graph representation of the BP equations for the optimal immunization problem in the SIR model. Variable nodes are of two types: One type includes pairs of auxiliary variables $\{(m_{ij}, m_{ji})\}$; the other includes the immunization variables $\{s_i\}$. There is a factor node for each node i of the original graph, including the energetic term $\mathcal{E}_i$ and all hard constraints $\{\Psi_{i\ell}\}$ involving node i as a first label.

the following, the integrals over $\mathbf{m}$ variables will be replaced by sums whenever we consider discretized versions of the equations. With a naive discretization, it is not easy to keep the propagation of the error under control during the iteration of the BP update rule. We also considered a relaxed version of the constraints for which we can bound the error due to the discretization; however, we checked numerically that the two versions of the algorithm give practically the same results. In practice, we observed that a discretization scheme that is linear in the logarithm of the single factors in Eq. (2) helps to conserve the precision when the number of factors is large (although other schemes give similar results).

The computation time of the update rule in the BP equations as written in Eq. (16) scales as $N_B^{k_i-1}$, where k_i is the degree of node i, making the trace over all configurations of the neighbors practically unfeasible on most graphs. The computational complexity of the BP update rule can be considerably reduced by exploiting the properties of convolutions of messages. For an arbitrary set $D\subseteq\partial i\backslash j$, we define the convolution function

$$M_D(S,T) = \sum_{\substack{m_{ki},\, k\in D \text{ s.t.} \\ S=\prod_{k\in D}[1-m_{ki}p_{ki}]}} \prod_{k\in D} P_{ki}\left(m_{ki}, q_i + (1-q_i)\left[1-\frac{T}{1-m_{ki}p_{ki}}\right]\right), \tag{18}$$

where $T=\prod_{k\in\partial i}[1-m_{ki}p_{ki}]$ so that $0\le T\le S\le 1$. Then, for any disjoint sets D_1 and D_2, we have

$$M_{D_1\cup D_2}(S,T) = \sum_{\substack{S_1,S_2 \text{ s.t.} \\ S=S_1S_2}} M_{D_1}(S_1,T)M_{D_2}(S_2,T). \tag{19}$$

One can start from the empty set and add the elements of $\partial i\backslash j$ one by one, using the convolution rules. Finally, the convolution function over the complete set $\partial i\backslash j$ can be used to compute the outgoing message as

$$P_{ij}(m_{ij},m_{ji}) \propto \mathcal{S}_{ij}(m_{ij},m_{ji}) + \mathcal{N}_{ij}(m_{ij},m_{ji}), \tag{20}$$

where the two terms, defined as

$$\mathcal{S}_{ij}(m_{ij},m_{ji}) = e^{-\beta\mu c_i} \prod_{k\in\partial i\backslash j}\sum_{m_{ki}} P_{ki}(m_{ki},0), \tag{21a}$$

$$\mathcal{N}_{ij}(m_{ij},m_{ji}) = e^{-\beta\epsilon\{1-(1-m_{ji}p_{ji})(1-m_{ij})\}} M_{\partial i\backslash j}\left(1-\frac{m_{ij}-q_i}{1-q_i}, (1-m_{ji}p_{ji})\left(1-\frac{m_{ij}-q_i}{1-q_i}\right)\right), \tag{21b}$$

refer to node i being immunized or not (the proportionality symbol, as usual, means that the message has to be properly normalized). Using the convolution trick, the computational complexity of the update rule on a node of degree k reduces to $O(kN_B^3)$. The factor N_B^3 comes from the computation of the trace over the auxiliary variables $\{m_{ki}\}$ by means of a convolution function: For each value of the N_B values taken by T, we have to sum over all N_B values taken by S_1 and by S_2.

Finally, in order to directly explore the optimal immunization assignments, we can define the MS messages as

$$\hat{P}_{ij}(m_{ij},m_{ji}) = \lim_{\beta\to\infty}\frac{1}{\beta}\log P_{ij}(m_{ij},m_{ji}).$$

For an arbitrary set $D\subseteq\partial i\backslash j$, we define the convolution function as

$$\hat{M}_D(S,T) = \max_{\substack{m_{ki},\, k\in D \text{ s.t.} \\ S=\prod_{k\in D}[1-m_{ki}p_{ki}]}} \sum_{k\in D} \hat{P}_{ki}\left(m_{ki}, q_i + (1-q_i)\left[1 - \frac{T}{1-m_{ki}p_{ki}}\right]\right), \tag{22}$$

where $T = \prod_{k\in\partial i}[1-m_{ki}p_{ki}]$ so that $0 \le T \le S \le 1$. Then, for any disjoint sets D_1 and D_2, we have

$$\hat{M}_{D_1\cup D_2}(S,T) = \max_{\substack{S_1,S_2 \text{ s.t.} \\ S=S_1S_2}} \{\hat{M}_{D_1}(S_1,T) + \hat{M}_{D_2}(S_2,T)\}. \tag{23}$$

The MS equations (in their discretized and efficient versions) read

$$\hat{P}_{ij}(m_{ij},m_{ji}) = \max\{\hat{\mathcal{S}}_{ij}(m_{ij},m_{ji}), \hat{\mathcal{N}}_{ij}(m_{ij},m_{ji})\} + C_{ij}, \tag{24}$$

where C_{ij} is an (irrelevant) additive constant, and

$$\hat{\mathcal{S}}_{ij}(m_{ij},m_{ji}) = \mu c_i + \sum_{k\in\partial i\setminus j} \max_{m_{ki}} \hat{P}_{ki}(m_{ki},0), \tag{25a}$$

$$\hat{\mathcal{N}}_{ij}(m_{ij},m_{ji}) = \epsilon\{1-(1-m_{ji}p_{ji})(1-m_{ij})\} + \hat{M}_{\partial i\setminus j}\left(1-\frac{m_{ij}-q_i}{1-q_i}, (1-m_{ji}p_{ji})\left(1-\frac{m_{ij}-q_i}{1-q_i}\right)\right). \tag{25b}$$

B. Derivation for the SIS model

In the SIS model, the relevant auxiliary variables by means of which we can connect the configuration of immunized nodes with the corresponding energy value are the mean-field variables $\{m_i\}_{i\in V}$ satisfying Eq. (7). As was done for the SIR model, we introduce Eq. (7) as a set of hard constraints in a statistical-mechanics model by means of the partition function

$$Z = \sum_{\mathbf{s}} \int d\mathbf{m} e^{-\beta\mathcal{E}(\mathbf{s},\mathbf{m})} \prod_{i\in V} \Psi_i(s_i, m_i, \{m_j\}_{j\in\partial i}), \tag{26}$$

where now we have

$$\Psi_i(s_i,m_i,\{m_j\}_{j\in\partial i}) = \delta\left(m_i - (1-s_i)\frac{q_i + (1-q_i)[1-\prod_{j\in\partial i}(1-p_{ji}m_j)]}{r_i + q_i + (1-q_i)[1-\prod_{j\in\partial i}(1-p_{ji}m_j)]}\right) \tag{27}$$

and the usual energy term

$$\mathcal{E}(\mathbf{s},\mathbf{m}) = \sum_i \mathcal{E}_i(s_i,m_i) = \mu\sum_i s_i c_i + \epsilon\sum_i \ell_i m_i. \tag{28}$$

The probability associated with a configuration of immunized nodes $\mathbf{s}$ is

$$Q(\mathbf{s}) = \frac{1}{Z}\int d\mathbf{m} e^{-\beta\mathcal{E}(\mathbf{s},\mathbf{m})} \prod_{i\in V} \Psi_i(s_i,m_i,\{m_j\}_{j\in\partial i}). \tag{29}$$

We repeat the same derivation as before, assuming that the graph is an infinite tree. Marginalizing over all variables but j, we compute the probability $Q_j(s_j)$ for the variable s_j as

$$Q_j(s_j) \propto \int_0^1 dm_j \prod_{i\in\partial j}\int_0^1 dm_i e^{-\beta\mathcal{E}_j(s_j,m_j)} \Psi_j(s_j,m_j,\{m_i\}_{i\in\partial j}) \prod_{i\in\partial j} P_{ij}(m_i,m_j), \tag{30}$$

where the BP messages $P_{ij}(m_i,m_j)$ satisfy the BP equations

$$P_{ij}(m_i, m_j) \propto \sum_{s_i} \prod_{k\in\partial i\setminus j} \int_0^1 dm_k e^{-\beta\mathcal{E}_i(s_i,m_i)} \Psi_i(s_i, m_i, \{m_k\}_{k\in\partial i}) \prod_{k\in\partial i\setminus j} P_{ki}(m_k, m_i), \tag{31}$$

and they represents the joint probability that the mean-field variables on i and j assume values m_i and m_j in the absence of the constraint Ψ_j and of the energetic term $\mathcal{E}_j$. Figure 4 displays a graphical representation of the BP equations (31) on the corresponding factor graph.

The single-site total probability marginal $P_i(m_i)$ is given by

$$P_i(m_i) = \sum_{s_i} \prod_{k\in\partial i} \int_0^1 dm_k e^{-\beta\mathcal{E}_i(s_i,m_i)} \Psi_i(s_i, m_i, \{m_k\}_{k\in\partial i}) \prod_{k\in\partial i} P_{ki}(m_k, m_i). \tag{32}$$

As already described for the SIR model, the BP equations can be solved numerically using a discretized version of the messages, in which the [0, 1] interval is divided into N_B bins. In this way, however, the number of operations required to compute the trace in Eq. (31) scales exponentially with the degree of the node; therefore, we employ the convolution method again. For an arbitrary set $D\subseteq\partial i\setminus j$, we define the quantity

$$M_D(S, m_i) = \sum_{\substack{\{m_k\},\, k\in D \text{ s.t.} \\ S = \prod_{k\in D}[1 - m_k p_{ki}]}} \prod_{k\in D} P_{ki}(m_k, m_i). \tag{33}$$

Then, for any disjoint sets D_1 and D_2, we have

$$M_{D_1\cup D_2}(S, m_i) = \sum_{\substack{S_1, S_2 \text{ s.t.} \\ S = S_1 S_2}} M_{D_1}(S_1, m_i) M_{D_2}(S_2, m_i), \tag{34}$$

and the BP equations become

$$\begin{aligned} P_{ij}(m_i, m_j) &\propto \sum_{s_i} \prod_{k\in\partial i\setminus j} \int dm_k e^{-\beta\mathcal{E}_i(s_i,m_i)} \Psi_i(s_i, m_i, \{m_k\}_{k\in\partial i}) \prod_{k\in\partial i\setminus j} P_{ki}(m_k, m_i) \\ &= \sum_{s_i} \sum_{S} M_{\partial i\setminus j}(S, m_i) \delta\left(m_i - (1 - s_i)\frac{q_i + (1 - q_i)[1 - (1 - p_{ji}m_j)S]}{r_i + q_i + (1 - q_i)[1 - (1 - p_{ji}m_j)S]}\right) e^{-\beta\mathcal{E}_i(s_i,m_i)} \\ &= e^{-\beta\mu c_i} + M_{\partial i\setminus j}\left(\frac{1 - m_i - r_i m_i}{(1 - q_i)(1 - m_i)(1 - p_{ji}m_j)}, m_i\right) e^{-\beta\epsilon\ell_i m_i}. \end{aligned} \tag{35}$$

Again, we can derive MS equations

$$\hat{P}_{ij}(m_i, m_j) = \max\{\hat{\mathcal{S}}_{ij}(m_i, m_j), \hat{\mathcal{N}}_{ij}(m_i, m_j)\} + C_{ij}, \tag{36}$$

where

$$\hat{\mathcal{S}}_{ij}(m_i, m_j) = \mu c_i + \sum_{k\in\partial i\setminus j} \max_{m_k} \hat{P}_{ki}(m_k, 0), \tag{37a}$$

$$\hat{\mathcal{N}}_{ij}(m_i, m_j) = \epsilon\ell_i m_i + \hat{M}_{\partial i\setminus j}\left(\frac{1 - m_i - r_i m_i}{(1 - q_i)(1 - m_i)(1 - p_{ji}m_j)}, m_i\right), \tag{37b}$$

and C_{ij} is an (irrelevant) additive constant.

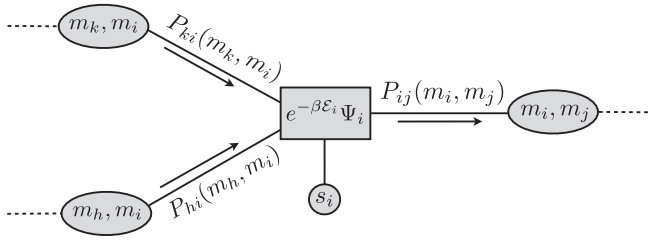

FIG. 4. Factor-graph representation of the BP equations for the optimal immunization problem in the SIS model. Variable nodes are of two types: One type includes pairs of auxiliary variables $\{(m_i, m_j)\}$; the other includes the immunization variables $\{s_i\}$. There is a factor node for each node i of the original graph, including the energetic term $\mathcal{E}_i$ and the hard constraint Ψ_i.

V. BP RESULTS ON ENSEMBLES OF RANDOM GRAPHS

On a general graph, the BP equations (16) and (31) are valid under the hypothesis of fast decay of correlations with the distance or replica-symmetric (RS) assumption [47,48]. Under this assumption, the statistical properties of the system are described by a unique Gibbs state (i.e., replica symmetry), and the BP equations admit a unique solution. Random graphs are natural benchmark structures for evaluating the quality of the results obtained by numerically solving the BP equations with histograms and the performances of the corresponding optimization method. In order to isolate and study the effects of immunization on the statistical properties of epidemic spreading, we consider a completely homogeneous setup: a RRG with uniform values of both spontaneous self-infection and disease transmission along the edges, i.e., $q_i = q$, $\forall\ i \in V$ and $p_{ij} = p$, $\forall\ (i,j) \in E$. For the sake of simplicity, we also consider uniform loss parameters and uniform immunization costs, i.e., $c_i = \ell_i = 1$, $\forall\ i \in V$. In the BP approach explained in Sec. IV, we can give a larger statistical weight to allocations of the immunized nodes that correspond to lower values of the energy. Increasing β for $\epsilon > 0$ [see Eq. (11)], the distribution $Q(\mathbf{s})$ becomes biased towards immunization sets that generate a smaller expected number of infected nodes compared to random immunizations of the same density. In the limit of $\beta \to \infty$, the weight is concentrated on the minima of the energy function, i.e., on the optimal immunization sets. In this framework, an interesting global observable is the generalization of the quantity f defined in Sec. III when we perform an average over all possible immunization sets $\mathbf{s}$ with the corresponding weight $Q(\mathbf{s})$. We call this quantity $\langle f \rangle$. We can exploit the definition of f in terms of the variables $\mathbf{m}$ and use BP to obtain an estimate of $\langle f \rangle$,

$$\langle f \rangle = \frac{1}{N}\sum_{i\in V}\langle m_i \rangle = \frac{1}{N}\sum_{i\in V}\int_0^1 dm_i P_i(m_i) m_i, \tag{38}$$

where $P_i(m_i)$ is given by Eqs. (17) and (32) for the SIR and SIS models, respectively. The chemical potential μ can be used to control the average fraction of immunized nodes, denoted by $\langle v \rangle$, that is computed from the solution of the BP equations as

$$\langle v \rangle = \frac{1}{N}\sum_{i\in V}\langle s_i \rangle = \frac{1}{N}\sum_{i\in V}\sum_{s_i=0,1} Q_i(s_i) s_i, \tag{39}$$

and for $Q_i(s_i)$ we use Eqs. (15) and (30) for the SIR and SIS models, respectively. It is thus possible to compute $\langle f \rangle$

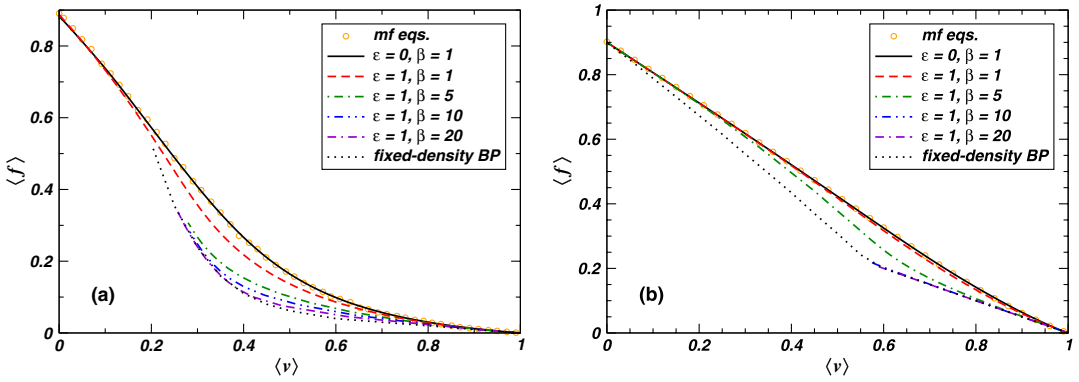

FIG. 5. Results for the analysis of the SIR and SIS models on random regular graphs of degree $K = 4$, with uniform self-infection probability $q = 0.1$ and uniform transmission probability $p = 0.5$. (a) For the SIR model, we plot the average density of nodes that got infected during the epidemic spreading $\langle f \rangle$ as a function of the average density $\langle v \rangle$ of immunized nodes. BP results on infinitely large graphs are reported for $\epsilon = 0$ and $\beta = 1$ (black solid line) and for $\epsilon = 1$ and $\beta = 1$ (red dashed line), 5 (green dot-dashed line), 10 (blue double-dot-dashed line), and 20 (violet dot-double-dashed line). Results of sampling over Eqs. (5a) and (5b) (orange circles), corresponding to random immunization, are also displayed. (b) For the SIS model, we plot the average density $\langle f \rangle$ of infected nodes in the stationary state as a function of the average density $\langle v \rangle$ of immunized nodes. BP results on infinitely large graphs are reported for $\epsilon = 0$ and $\beta = 1$ (black solid line) and for $\epsilon = 1$ and $\beta = 1$ (red dashed line), 5 (green dot-dashed line), 10 (blue double-dot-dashed line), and 20 (violet dot-double-dashed line). Results of sampling over Eq. (7) (orange circles), corresponding to random immunization, are also displayed.

as a function of $\langle v \rangle$ for a fixed choice of the other parameters. We present here results obtained for infinitely large regular random graphs, obtained using the BP equations in the single-link approximation, i.e., when we self-consistently solve the BP equations assuming that all nodes have essentially the same statistical properties. For both SIR and SIS models, the results of the BP equations in the single-link approximation are shown in Fig. 5, with the choice of parameters $q = 0.1$ and $p = 0.5$ and different values of ϵ and β. In the case of random immunization ($\epsilon = 0$), the results obtained using BP on infinitely large RRG are compared with the average behavior observed by sampling the solutions of Eqs. (5a) and (5b) [and Eq. (7), respectively] and by simulating the SIR (and SIS) stochastic process on finite RRG of $N = 10^3$ nodes. The latter are obtained by sampling over 10^3 configurations of immunized nodes for each value of $\langle v \rangle$. The agreement is very good for both models.

Increasing β for $\epsilon = 1$, the solutions of the BP equations show a monotonic decrease in the density of infected nodes. The reduction of the infection level is particularly visible for intermediate values of $\langle v \rangle$, while it becomes almost negligible at small and large densities of immunized nodes. The fact that the density $\langle v \rangle$ is not directly fixed (as for microcanonical systems) but that it is implicitly varied as an effect of tuning the chemical potential μ and then evaluated from the outcome of the BP equations is the cause of an undesirable issue at large values of β. For large β, the free energy of the statistical mechanics problem is not convex over the whole interval [0, 1] of values assumed by $\langle v \rangle$. The Legendre transform, implicitly used in the cavity method [48], spontaneously selects the convex envelope of the free energy, limiting the interval of the possible values assumed by the density of immunized nodes. This is visible in Fig. 5; as β increases for $\epsilon = 1$, the curves get interrupted at some (nonzero) value of $\langle v \rangle$. This means that, for this choice of the parameters, smaller nonzero immunization sets are, in energetic terms, less convenient than no immunization at all (i.e., of the point at $\langle v \rangle = 0$). To eliminate this phenomenon, one can fix the value of $\langle v \rangle$ by introducing an adaptive external field that is self-consistently adjusted during the iterations of the BP equations [49]. Using this technique, we were able to explore all values of the density of immunized nodes, although the procedure is not guaranteed to converge at all values of β. More precisely, we computed the set of points indicated with black dots in Fig. 5, which correspond to the best results (lowest value of $\langle f \rangle$) obtained by varying β at fixed values of $\langle v \rangle$. We expect these results to be very close to the optimal ones for a large graph.

Upon decreasing the spontaneous self-infection probability q, the effect of the optimization becomes more pronounced. This is shown in Fig. 6, where the results of optimization are compared with random immunization for both $q = 0.1$ and $q = 0.01$. In the same figure, we also plot the entropy $\mathcal{S}$ of the immunization sets as a function of the density $\langle v \rangle$ (computed from the solution of the BP equations as in Ref. [48]). The red dashed lines, referring to random immunization, show the standard entropy curve given by $\mathcal{S}(\langle v \rangle) = -\langle v \rangle \log\langle v \rangle - (1 - \langle v \rangle) \log(1 - \langle v \rangle)$ and derived from a binomial distribution of $\langle v \rangle N$

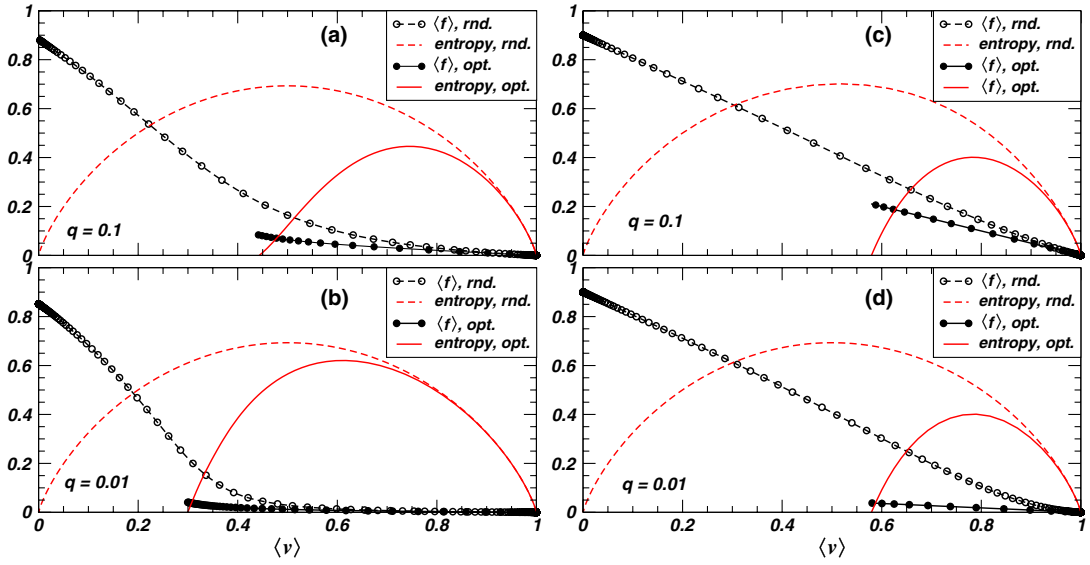

FIG. 6. Results for the analysis of the SIR and SIS models on random regular graphs of degree $K = 4$, with uniform self-infection probability $q = 0.1$. 0.01 and uniform transmission probability $p = 0.5$. For the SIR model (a,b), we plot the average density of nodes that got infected during the epidemic spreading $\langle f \rangle$ as a function of the average density $\langle v \rangle$ of immunized nodes, both in the random case (open circles) and under optimization (full circles). For the SIS model (c,d), we plot the average density $\langle f \rangle$ of infected nodes in the stationary state as a function of the average density $\langle v \rangle$ of immunized nodes, both in the random case (open circles) and under optimization (full circles).

immunized nodes among the N possible ones. Red solid lines instead represent the entropy curves under strong optimization, i.e., corresponding to low-energy states. In this case, the entropy becomes negative for sufficiently small values of $\langle v \rangle$, meaning that no configurations of immunized nodes can be found in the Gibbs state defined at that temperature (this is related to the nonconvexity argument explained before). In other words, in that region it is (energetically) more convenient to let the disease spread naturally (with no immunity) than performing immunization.

With the BP analysis, we can go beyond the study of the average macroscopic behavior and focus on more informative quantities, such as the distribution $P_i(m_i)$. In the absence of immunization, $P_i(m_i)$ is a delta function because Eq. (5a) [and Eq. (7)] has a unique solution. After averaging over a set of configurations of immunized nodes, each node i is represented by a distribution $P_i(m_i)$ instead of a single value of infection probability m_i. In a completely homogeneous setup (i.e., random regular graph and uniform parameters), all nodes have the same statistical properties; therefore, we can ignore the node label in the distribution and assume $P_i(m_i) = P(m)$, $\forall\ i \in V$. When the set of immunized nodes is drawn from a uniform distribution (i.e., $\epsilon = 0$), we can directly compare the shape of $P(m)$ obtained by solving the BP equations with that obtained either by explicitly simulating the stochastic process or by repeatedly solving the mean-field equations with a large sample of configurations of immunized nodes. Examples of the results obtained for the SIR model on RRG of degree $K = 4$, $q = 0.1$, and $p = 0.5$ are displayed in Fig. 7. The shape of the distribution $P(m)$ varies considerably for different values of the average immunization density $\langle v \rangle$. The comparison between the results for $\epsilon = 0$ and the empirical distributions obtained by sampling solutions of Eqs. (5a) and (5b) and by means of simulations of the stochastic dynamics is very favorable. The agreement between BP and sampling results does not seem to be affected by the discretization of the [0, 1] interval employed in order to solve the BP equations (here, $N_B = 200$). Also, the agreement with the results of simulations is very good, and it improves by increasing the number of configurations of immunized nodes employed in the simulations (here, we performed an average over 10^4 realizations at fixed density).

The distribution is always very heterogeneous, and the average value (reported in Fig. 5) is not at all representative of the behavior of the system. In general, $P(m)$ is characterized by a series of isolated peaks at low values of m and by a continuous distribution in the bulk. The origin of the delta peaks is strictly related to the local structure of the graph around a node after immunization. For instance, the first peak for $m > 0$ corresponds to isolated nodes, i.e., nodes that are completely surrounded by immunized ones and thus disconnected from the rest of the graph. Such nodes are infected with probability q, which is exactly the position of the first delta peak. A second peak occurs at $m = 0.145$, corresponding to the probability of a node being part of an isolated infected dimer. For infected trimers (chains of length 3), the central node corresponds to $m = 0.187$, while the external nodes have $m = 0.165$. These peaks are visible in Fig. 7. Then, one can continue identifying other local clusters, such as starlike structures and small chains, each one corresponding to an isolated peak in the distribution. The continuous bulk of the distribution should instead be identified with a superposition of values due to large-scale clusters of infected nodes.

In Fig. 8, we report a similar plot for the SIS model. The structure of $P(m)$ is generally different, with few isolated peaks in which the weight of the distribution concentrates. This effect could be due to the naive mean-field approximation applied at the level of the equations used as hard constraints, which is less accurate than that applied in the SIR model.

There is a remarkable difference in the shape of $P(m)$ when a nonuniform weight is associated with immunization sets that generate different levels of infection. In the BP formalism, this is done by increasing ϵ from zero. Figure 9(a) displays the distribution $P(m)$ for the SIR model at a fixed value of $\langle v \rangle = 0.6$, $\epsilon = 1$, and different values of β. Increasing β, the distribution concentrates on a narrower interval of values of m. The information in these plots is important because it can be used to compute, for a given node in a graph, the probability that such a node is infected for a given immunization strategy. One can also

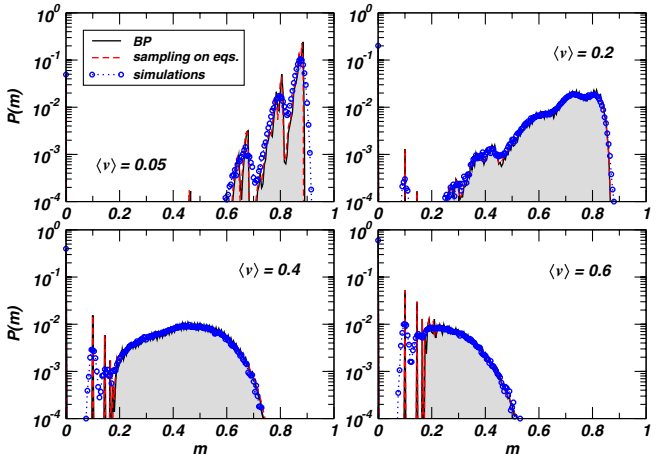

FIG. 7. Distribution $P(m)$ of the probability m that a node got infected during the SIR process on random regular graphs of degree $K = 4$ for $\langle v \rangle = 0.05, 0.2, 0.4, 0.6$. Results are computed using BP by means of Eq. (17) (black solid line) on infinite networks, by sampling solutions of Eqs. (5a) and (5b) (red dashed lines) and by means of simulations of the stochastic dynamics, both on a finite network of $N = 10^3$ nodes and with a sample of 10^4 immunization sets.

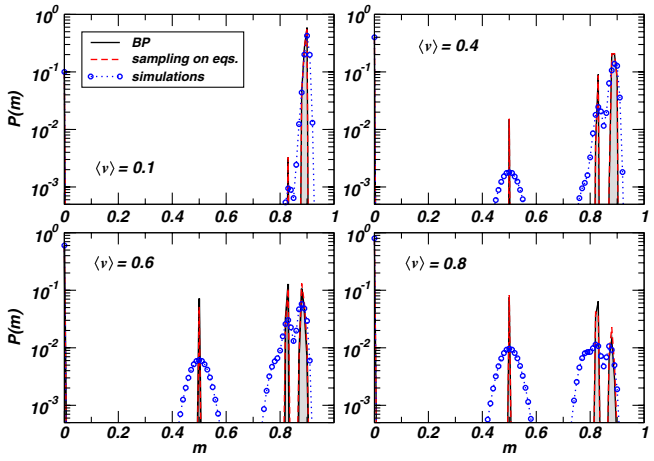

FIG. 8. Distribution $P(m)$ of the probability m that a node is infected in the stationary state of the SIS process on random regular graphs of degree $K = 4$ for $\langle v \rangle = 0.1$, 0.4, 0.6, 0.8. Results are computed using BP by means of Eq. (32) (black solid line) on infinite networks, by sampling solutions of Eq. (7) (red dashed lines), and by means of simulations of the stochastic dynamics, both on a finite network of $N = 10^3$ nodes and with a sample of 10^4 immunization sets.

ask how many immunization sets exist at a given density $\langle v \rangle$ that generate an average infection level of $\langle f \rangle$. For a RRG with degree $K = 4$ and $\langle v \rangle = 0.6$, this is shown in Fig. 9(b), where we plot the entropy of immunization sets of fixed density as a function of $\langle f \rangle$. These results are obtained from the solutions of the BP equations by increasing β from 1 (for $\epsilon = 1$). In Fig. 10, we report analogue plots for the SIS model with $\langle v \rangle = 0.6$. For a random allocation of immunized nodes [black circles in Fig. 10(a)], $P(m)$ exhibits (in addition to a delta in zero) a series of narrow peaks at positive m. Increasing β, those at larger m slowly disappear, leaving only a delta peak at $m \approx 0.5$. Figure 10(b) shows the behavior of the entropy curve at $\langle v \rangle = 0.6$ for the SIS model. In both models, the entropy at $\langle v \rangle = 0.6$ does not vanish continuously when the minimum value of infection $\langle f \rangle$ is reached, but it remains considerably large. This is in accordance with the behavior of the entropy curves displayed in Fig. 6. A continuously vanishing behavior can be observed instead if we select a density value that falls in the region in which, at large β, the entropy curve becomes negative (see later discussion about MS results).

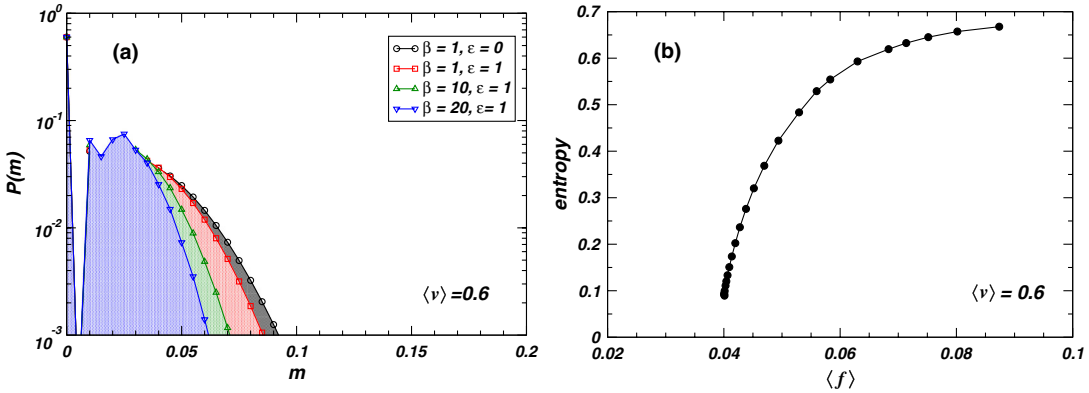

FIG. 9. (a) Distribution $P(m)$ of the probability m that a node got infected during the SIR process on random regular graphs of degree $K = 4$, $q = 0.1$, $p = 0.5$, $\langle v \rangle = 0.6$, and different values of β and ϵ. (b) Entropy of immunization sets of density $\langle v \rangle = 0.6$ as a function of $\langle f \rangle$ (obtained by increasing β for $\epsilon = 1$).

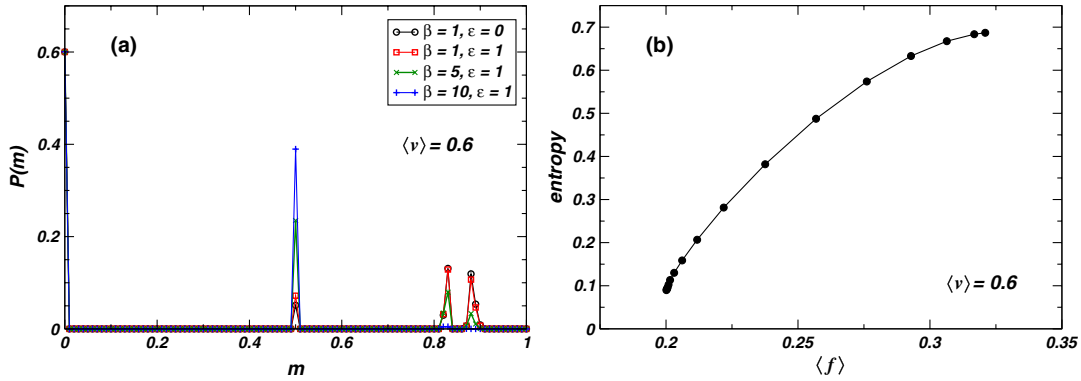

FIG. 10. (a) Distribution $P(m)$ of the probability m that a node is infected in the stationary state of the SIS process on random regular graphs of degree $K = 4$, $q = 0.1$, $p = 0.5$, $\langle v \rangle = 0.6$, and different values of β and ϵ. (b) Entropy of immunization sets of density $\langle v \rangle = 0.6$ as a function of $\langle f \rangle$ (obtained by increasing β for $\epsilon = 1$).

For $q = 0.1$, the BP results have been obtained using a discretization of the interval [0, 1] in $N_B = 200$ bins. In general, the smallest possible nonzero value assumed by m is $m = q$; therefore, in order to have a sufficiently good resolution, one should always use $N_B > 1/q$. Figure 11 shows that, while a continuous percolation-like phase transition is expected at $\langle v \rangle \approx 1/3$ for random immunization, the solution of the BP equations obtained with the discretization method does not reproduce this feature for small values of N_B. The smooth crossover converges to the expected sharp transition when the number of bins is increased, in a way that is reminiscent of finite-size scaling. A similar phenomenology is observed in the presence of optimization as long as q is nonzero (when $q = 0$, the optimization pushes the system into the trivial solution with no infected nodes). The accurate solution of BP and MS equations by discretization of the marginals ($N_B > 1/q$) may become unfeasible for very small values of q because the time complexity of the algorithms scales as N_B^3. A possible solution is that of adopting a mixed representation, such as

$$P(m) = \sum_{\ell=1}^{L} a_\ell \delta(m - m_\ell) + \sum_{\ell'=1}^{N_B-1} b_{\ell'} \mathbb{1}[m_{\ell'} < m \leq m_{\ell'+1}], \tag{40}$$

that exploits the knowledge of the position of the first L delta peaks of the distribution $P(m)$ computed by solving the underlying equations for $\{m_i\}$ or $\{m_{ij}\}$ on small structures. This approach will be developed in a future work.

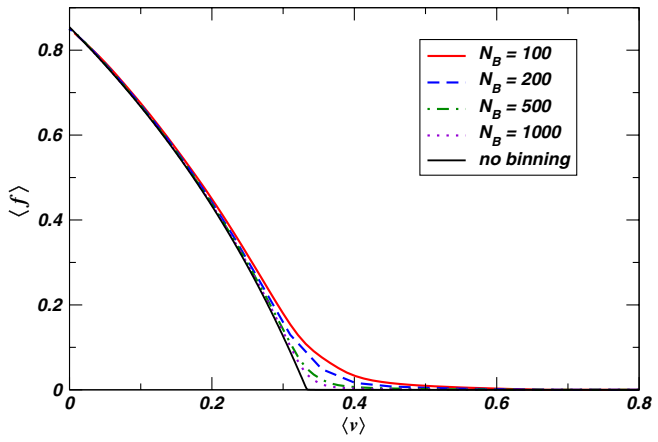

FIG. 11. Effects of the discretization of the BP marginals in the limit $q \to 0$. The curves represent the average density of nodes that got infected during the epidemic spreading $\langle f \rangle$ as a function of the average density $\langle v \rangle$ of immunized nodes for random immunization ($\epsilon = 0$) in the SIR model on random regular graphs of degree $K = 4$ and uniform transmission probability $p = 0.5$. For $q = 0$, a percolation-like phase transition is expected at $\langle v \rangle \approx 1/3$. For low values of N_B, the BP equations (in the single-link approximation, in order to mimic an infinite graph) give a smooth crossover that approaches the correct threshold effect by increasing the number of bins.

VI. COMPARISON BETWEEN DIFFERENT OPTIMIZATION METHODS

A comparison between different optimization methods should take into account both the performances of the optimization and the efficiency of the algorithms. Random regular graphs with uniform parameters are not a particularly good setup to compare different methods because all nodes are equally important. This is visible in Fig. 12(a), which displays, for the SIR model, the curve $\langle f \rangle$ as a function of $\langle v \rangle$ obtained by several different immunization algorithms: a recalculated degree centrality (green dot-dashed line), recalculated eigenvector centrality (blue line), greedy algorithm (maroon dashed line), and density-constrained simulated annealing (red squares). The results obtained with these algorithms on a RRG with $N = 10^3$ nodes are compared with the best prediction computed using density-constrained BP equations on infinitely large RRG (violet crosses). Each point is computed at a different temperature, which is the lowest temperature at which the convergence of BP equations at that density $\langle v \rangle$ is reached. For the same optimization problem, the energy as a function of the density of immunized nodes is shown in Fig. 12(b) (we set $\mu = \epsilon = 1.0$). Greedy and topologically based algorithms perform very similarly, and approximately the same result is obtained with fixed-density simulated annealing, even in the case of rather slow annealing schedules with up to 10^5 steps between $\beta = 0.1$ and $\beta = 10^3$. The BP prediction suggests that (at least for infinitely large graphs) slightly lower energy values could be reached for intermediate values of $\langle v \rangle$. Remarkably, using just 100 bins, the MS algorithm was able to find an immunization set at such lower energies (black vertical cross), which is expected to be the optimal one. We note that the plain MS equations do not always converge. To overcome this difficulty, we employed the reinforcement technique [26,27,56,57]. We also ran simulated annealing with $> 10^6$ steps and without density constraints, and we found the same optimal point. Notice that, although the computational time of MC-based methods is comparable with that of MS for graphs of few thousands of nodes, simulated annealing becomes unfeasible for large-scale networks.

We also analyzed some epidemic properties associated with the immunization set found by MS and compared them to those expected by solving the BP equations in the single-link approximation at the same density $\langle v \rangle = 0.363$. Figure 13(a) displays the behavior of $P(m)$ computed using BP for different values of ϵ and β, and the same quantity obtained by solving Eqs. (5a) and (5b) for the configuration of immunized nodes obtained using the MS algorithm (crosses). The accord between the latter and the BP results for large β values is very good [compare the blue

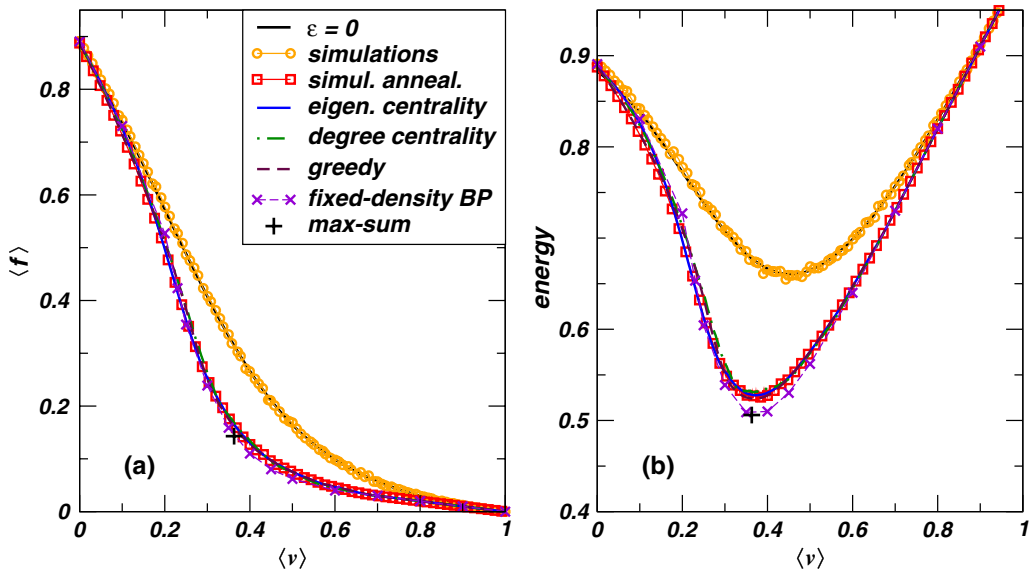

FIG. 12. Comparison between several immunization methods on a RRG with $N = 10^3$ nodes and degree $K = 4$ for the SIR model (with $q = 0.1$, $p = 0.5$): (a) average fraction of infected nodes $\langle f \rangle$ vs average fraction of immunized nodes $\langle v \rangle$; (b) energy (for $\mu = 1.0$) vs average fraction of immunized nodes $\langle v \rangle$. Reported results include the following: Simulated annealing (red squares), eigenvectors centrality (blue solid line), degree centrality (green dot-dashed line), greedy (maroon dashed line), and the best obtained solving fixed-density BP equations (violet crosses) and MS (black vertical cross). Random immunization is also reported (black line for the BP results with $\epsilon = 0$ and yellow circles for the results of direct stochastic simulations).

distribution with the violet one in Fig. 13(a)]. For the same choice of parameters, Fig. 13(b) shows the entropy as a function of the density of infected nodes $\langle f \rangle$ (obtained by performing BP for increasing values of β). The entropy decreases monotonically and vanishes at $\langle f \rangle \approx 0.143$, which indicates the minimum level of infection attainable at that density of immunized nodes. The vertical line marking the density of immunized nodes found in the solution obtained using the MS algorithm on a graph of $N = 10^3$ nodes perfectly matches this lower bound.

While on random regular graphs all optimization methods give similar results, the performances are starkly different when considering networks with nonhomogeneous degree sequence, clustering, and other topological properties typical of real-world systems. We first consider two computer-generated uncorrelated random graphs of the same size, $N = 1000$, and different degree distributions: an Erdös-Rényi random graph of average degree $z = 10$ and a random graph with a degree sequence sampled from the power-law degree distribution $P(k) \propto k^{-\gamma}$ with exponent $\gamma = 2.2$. The plots of the energy as a function of the density of immunized nodes in Fig. 14 show the different performances of the algorithms for some choice of the parameters q, p, μ, and ϵ. Remarkably, in both cases the MS algorithm finds a point that is, energetically, at least as good as the best result of the

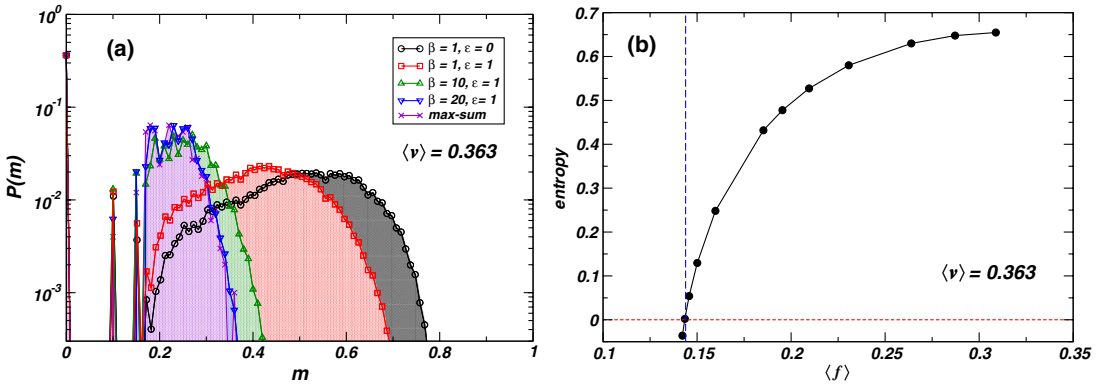

FIG. 13. (a) Distribution $P(m)$ of the probability m that a node got infected during the SIR process on random regular graphs of degree $K = 4$, $q = 0.1$, $p = 0.5$, $\langle v \rangle = 0.6$, and different values of β and ϵ. (b) Entropy of immunization sets of density $\langle v \rangle = 0.363$ as a function of $\langle f \rangle$ (obtained by increasing β for $\epsilon = 1$).

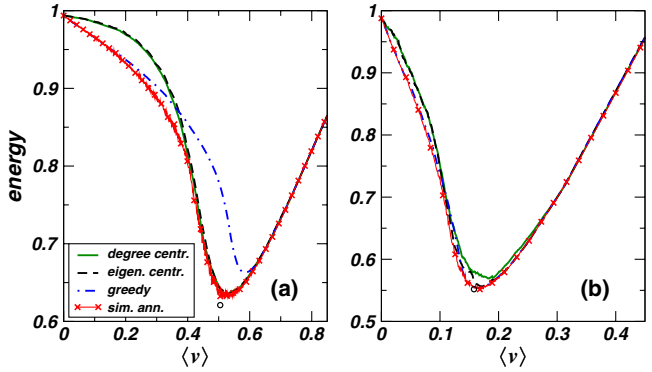

FIG. 14. Comparison between several immunization methods for the SIR model on two synthetic networks of $N = 1000$ nodes: (a) Erdös-Rényi random graph of average degree $z = 10$ and parameters $q = 0.1$, $p = 0.5$, and $\mu = \epsilon = 1.0$; (b) heterogeneous random graph with power-law degree distribution $P(k) \propto k^{-\gamma}$ with exponent $\gamma = 2.2$ and parameters $q = 0.1$, $p = 0.9$, and $\epsilon = 1$, $\mu = 2$. The data represent the average energy as a function of the average fraction of immunized nodes $\langle v \rangle$ for degree centrality (green line), eigenvector centrality (black dashed line), the greedy algorithm (blue dot-dashed line), and simulated annealing (red crosses). The black circles (with error bars) are the results obtained using the MS algorithm.

other methods (black circle). Since for both networks the density of triangles is small (see Table I), the assumption of a local treelike structure should be considered approximately correct. However, it is known that message-passing algorithms also perform very well in cases in which the standard hypothesis of treelike structure is not fully satisfied. To check this crucial point, we performed the same comparative experiment on a series of real-world (undirected) networks with non-negligible clustering and community structure (see Table I): (a) a social network of frequent associations between 62 dolphins in a community living in New Zealand [52], (b) coappearance network of characters in the novel Les Miserables [53], (c) a time-independent projection of the Sociopatterns contact network, describing the face-to-face behavior of people during an exhibition at the Science Gallery in Dublin [54], and (d) a fully symmetric version of a friendship network among 685 students at a school in the U.S. [55]. All networks are of moderate size in order to make possible a direct comparison of the different methods with simulated annealing with a sufficiently slow annealing schedule. Figure 15 shows that the algorithms under study can give rather different results: Interestingly, topologically based methods reach much lower energy values than the greedy one. Despite the large density of triangles and the existence of community structure, message-passing methods work very well: In all cases, the optimal point found using the MS algorithm (black cross) is energetically equally good or even better than those obtained with the other methods.

TABLE I. Summary of the main characteristics of the networks considered for the comparison between different optimization methods: number of nodes N, number of undirected edges $|E|$, and the transitivity ratio, defined as the average ratio between the actual number of closed triplets and the number of connected triplets of nodes. More details on the structure of the networks are provided in the text.

Network type	N	$\lvert E\rvert$	Transitivity ratio
Zachary's [50,51]	34	77	0.248
Dolphins [50,52]	62	159	0.309
Les Miserables [50,53]	77	254	0.499
Sociopatterns [54]	413	2767	0.436
School friendship [55]	685	3441	0.344
Random regular	1000	2000	5×10^{-4}
Erdös-Rényi (ER)	1000	5037	0.01
Scale-free (SF)	1000	2102	0.028

When inhomogeneous costs are added, the algorithms based on topological metrics are ruled out. Although they might work very well for some values of μ (the parameter governing the tradeoff between terms in the energy), they do not take into account the contribution to the energy associated with the additional parameters. Energy-based methods are more flexible, as they adapt to variations of all parameters appearing in the energy. The greedy algorithm is, however, limited by the incremental procedure that finds local minima of the energy with a rather small density of immunized nodes. Adding further nodes to the immunization set, the results tend to worsen. On the contrary,

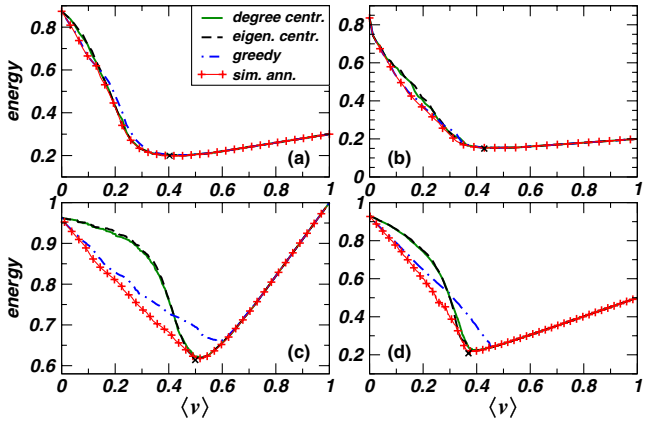

FIG. 15. Comparison between several immunization methods for the SIR model on four real-world networks: (a) social network in a community of dolphins (parameters $q = 0.1$, $p = 0.5$, $\mu = 0.3$, $\epsilon = 1$), (b) network of characters of Les Misérables (parameters $q = 0.1$, $p = 0.5$, $\mu = 0.2$, $\epsilon = 1$), (c) the Sociopatterns contact network ($q = 0.1$, $p = 0.5$, $\mu = \epsilon = 1$), (d) friendship network in a U.S. high school ($q = 0.1$, $p = 0.5$, $\mu = 0.5$, $\epsilon = 1$). The energy as a function of $\langle v \rangle$ is shown, for immunization sets obtained using degree centrality (green line), eigenvector centrality (black dashed line), the greedy algorithm (blue dot-dashed line), and fixed-density simulated annealing (red vertical crosses). The optimal points found by Max-Sum are indicated by black crosses.

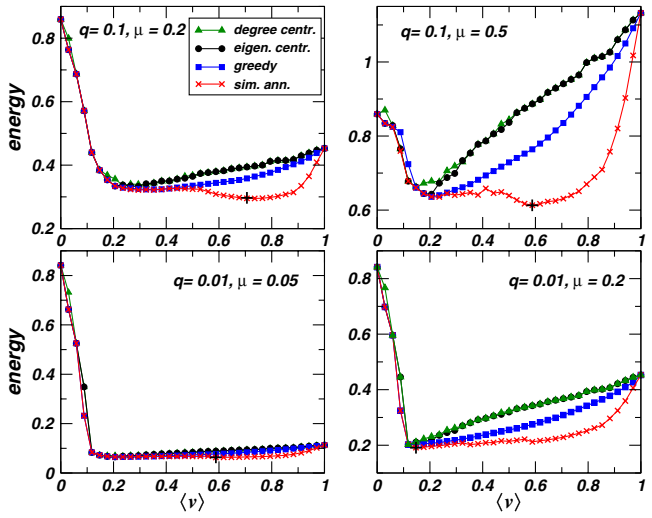

FIG. 16. Energy as a function of $\langle v \rangle$ for immunization sets obtained using degree (green triangles) and eigenvector centrality (black circles), the greedy algorithm (blue squares), and fixed-density simulated annealing (red crosses), for the SIR model ($p = 0.5$) on Zachary's karate club network of $N = 34$ individuals. The MS results are indicated by black vertical crosses. The immunization cost was assumed to be half of the degree of a node.

simulated annealing and the MS algorithm are able to find immunization sets with a larger number of immunized nodes but that are, overall, less expensive. This phenomenon is observed in the data reported in Fig. 16, where we analyzed immunization strategies for the SIR model on the famous Zachary's social network of friendships between 34 members of a karate club at a U.S. university in the 1970 s [51]. On such a small network, the immunization sets obtained by simulated annealing (performed with 10^5 MC steps between $\beta = 0.1$ and $\beta = 10^4$) are likely to correspond to the global energy minima (for each value of $\langle v \rangle$). In order to generate a rough energy landscape, we considered immunization costs that are correlated with the degree of the nodes ($c_i = k_i/2$, where k_i is the degree of i). Figure 16 shows that for both $q = 0.1$ and $q = 0.01$, changing the tradeoff between the terms of the energy (i.e., increasing μ), the optimal immunization set is found at very different (larger) values of the density $\langle v \rangle$. Despite the heterogeneous costs and the fact that Zachary's network is highly clustered, MS correctly finds the optimal immunization set for all cases under study (black vertical crosses in Fig. 16).

VII. DISCUSSION

In this work, we considered the optimization problem of targeted network immunization against epidemic spreads. The corresponding energy function to be minimized is a tradeoff between the costs of immunizing nodes and the expected extent of the infection. This optimization problem turns out to be computationally worst-case intractable; it was proven to be NP hard even for very simple stochastic propagation models. For two prototypical models of stochastic epidemic spreading, the SIR and SIS models, we considered mean-field equations that can be used to evaluate the level of infection in the stationary state associated with every configuration of immunized nodes. The original energy function and the problem of finding the optimal immunization set can be (approximately) recast in terms of these mean-field variables in a mixed representation in which the selection of nodes to be immunized is represented by binary variables and the local level of infection (in the stationary state) is expressed by continuous variables in [0, 1]. In this formulation, constraints and energy terms are local, allowing the application of the cavity method and the development of efficient message-passing algorithms such as BP. Our results obtained using BP equations on random regular graphs shed light on the statistical properties of immunization sets, uncovering in which regions of the parameter space, and to what extent, targeted immunization is actually more effective than random immunization. The zero-temperature limit of these equations gives the MS algorithm, which can be used to find a solution to the optimization problem. We showed, both on synthetic and real-world networks with nontrivial topological structure, that the MS algorithm outperforms several popular immunization methods based on topological metrics and greedy strategies. The solution found using MS is not guaranteed to be optimal; therefore, we also performed simulated annealing, which is able to reach the optimum, at least for a sufficiently slow (maybe exponential) annealing schedule. For networks of moderate size, we could compare the lowest-energy immunization set found by MC methods with the solution found by MS. The latter was always at least as good as the former, providing experimental evidence of the validity of the optimization technique. Moreover, unlike MC-based methods, the MS algorithm scales only linearly with the network size. As a drawback, the algorithm scales as N_B^3, where N_B is the number of bins necessary to represent the distribution of real values as histograms. We emphasize that the discretization method used in the current implementation is a very naive and straightforward one, and there are several ways to considerably reduce N_B by adopting more efficient representation of the messages (as explained in Sec. V). For this reason, we expect that message-passing algorithms such as the one proposed here could be used to study the immunization problem even on very large networks. This will be the scope of future research.

The results of the comparison between different optimization methods on a variety of networks show that, as long as the network and the parameters are sufficiently homogeneous, all methods give approximately the same results. In this case, heuristic strategies based on topological metrics, such as degree centrality or eigenvector centrality, could be preferred, as they are very simple

and fast. The MS algorithm also performs very well in network structures that are not treelike and that contain many triangles and dense substructures, such as the four real-world networks considered. When the energy function includes inhomogeneous costs, simpler heuristics turned out to be suboptimal in our experiments. Moreover, greedy-based methods that take into account the correct energy function are based on a progressive scheme (as immunized nodes are added incrementally) and often fail to reach the ground state whenever the energy landscape is rugged. This is well demonstrated by results obtained on the small, but representative, Zachary's karate club network.

The method presented here could be applied to a number of other optimization problems including, but not restricted to, other epidemic models in discrete or continuous time, provided that similar fixed-point equations are defined for node or edge variables. The control variables in that case could be a set of node-dependent external parameters. Using message-passing techniques, it could be possible to select the configuration of external parameters that corresponds to some desired outcome for the global state of the system. This general formulation can be used in a wide spectrum of applications in problems involving the control of network dynamics.

ACKNOWLEDGMENTS

The authors acknowledge the European Grants FET Open No. 265496 and ERC No. 267915 and Italian FIRB Project No. RBFR10QUW4.

APPENDIX: ALGORITHMS

1. Incremental heuristic algorithms

The heuristic algorithms considered in this paper are based on an incremental procedure, by means of which the algorithm adds, one by one, nodes to the immunization set, each time choosing the node that minimizes some score function. Hence, given a graph $G = (V, E)$ and one of the scoring strategies described below, the immunization algorithm is based on the following iteration. For $k = 1, \ldots, |V|$: (1) [score] compute the score vector $\mathbf{z}$; (2) [immunization] choose node index i_k^* corresponding to the highest component of score $\mathbf{z}$ and remove the graph from it.

This procedure gives an immunization set $\{i_1^*, \ldots, i_k^*\}$ for any desired number k of immunized nodes.

The actual definition of the score vector depends on the heuristic strategy considered. For degree centrality, the score of a node is just the number of nonimmunized neighbors. For the eigenvector centrality, the score of a node is, recursively, a function of the scores of neighbors,

$$z_i = \frac{1}{\lambda}\sum_{j\in\partial i} z_j, \tag{A1}$$

where λ is a properly defined constant. The score vector $\mathbf{z}$ satisfies the eigenvector equation $A\mathbf{z} = \lambda\mathbf{z}$, where A is the adjacency matrix, such that $a_{ij} = 1$ if there is an edge between nodes i and j and $a_{ij} = 0$ otherwise. Under the condition that $\mathbf{z} > 0$ and the graph is connected, the constant λ corresponds to the greatest eigenvalue of the adjacency matrix (Perron-Frobenius theorem). Hence, the score vector $\mathbf{z}$ can be computed by iteration from a homogeneous initial condition $z_i^0 = 1,\ \forall\ i \in V$ using the power method, i.e., defining

$$z_i^{t+1} = \frac{1}{\lambda_t}\sum_{j=1}^{N} a_{ij} z_j^t, \tag{A2}$$

where λ_t is an appropriately defined normalization constant (recomputed at each time step). If the iteration converges, it gives the eigenvector centrality of the nodes.

In the greedy algorithm implemented in the present paper, the score of a node i is equal to the variation in energy, computed from Eq. (11), associated with the addition of i to the immunization set. Given $\mathbf{s} = (s_1, \ldots, s_{i-1}, 0, s_{i+1}, \ldots, s_N)$ and $\mathbf{s}' = (s_1, \ldots, s_{i-1}, 1, s_{i+1}, \ldots, s_N)$, we define

$$z_i = \mathcal{E}(\mathbf{s}) - \mathcal{E}(\mathbf{s}'). \tag{A3}$$

2. Simulated annealing

The space of 2^N binary configurations corresponding to immunization sets can be sampled using a Monte Carlo algorithm that, at large inverse temperatures, converges towards the minima of the energy function $\mathcal{E}$. In practice, starting from a randomly selected binary configuration, the convergence to a global minimum can be achieved only by using an annealing schedule that guarantees a sufficiently slow decrease in temperature. Given a randomly chosen initial condition $\mathbf{s} = (s_1, \ldots, s_N)$, and an initial value β_I for the inverse temperature β, the adopted annealing schedule reaches a final value β_F in a number M of proposed single-spin flip. Unfortunately, the number M of steps required to reach the minimum of the energy at large β often scales exponentially with the system size N.

In summary, the algorithm works as follows: (0) choose an initial condition $\mathbf{s}$, set $\beta \leftarrow \beta_I$; (1) randomly select a node i to be flipped; (2) given $\mathbf{s} = (s_1, \ldots, s_{i-1}, s_i, s_{i+1}, \ldots, s_N)$ and $\mathbf{s}' = (s_1, \ldots, s_{i-1}, 1 - s_i, s_{i+1}, \ldots, s_N)$, compute the variation of energy $\Delta\mathcal{E} = \mathcal{E}(\mathbf{s}') - \mathcal{E}(\mathbf{s})$; (3) accept the move $\mathbf{s} \leftarrow \mathbf{s}'$ with probability $e^{-\beta\Delta\mathcal{E}}$; (4) set $\beta \leftarrow \beta + \delta\beta$; (5) if $\beta < \beta_F$, then go to point (1).

In our simulations, we tested different experimental setups, using both a linear schedule with $\delta\beta = (\beta_F - \beta_I)/M$ and a faster exponential one, in which $\delta\beta = \beta[(\beta_F/\beta_I)^{1/M} - 1]$. We usually considered $\beta_I < 1$ and β_F between 10^3 and 10^4, with a number of single-spin proposed moves M between 10^5 and 10^6.

In the fixed-density simulated annealing algorithm, we considered only moves that do no change the number of

immunized nodes, such as immunization exchange moves. In this case, the overall algorithm consists in repeating the annealing schedule for any possible value of the number of immunized nodes from 0 to N.

[1] R. Cohen, K. Erez, D. ben Avraham, and S. Havlin, *Resilience of the Internet to Random Breakdowns*, Phys. Rev. Lett. **85**, 4626 (2000).

[2] R. Pastor-Satorras and A. Vespignani, *Epidemic Spreading in Scale-Free Networks*, Phys. Rev. Lett. **86**, 3200 (2001); *Epidemic Dynamics and Endemic States in Complex Networks*, Phys. Rev. E **63**, 066117 (2001).

[3] R. Pastor-Satorras and A. Vespignani, *Immunization of Complex Networks*, Phys. Rev. E **65**, 036104 (2002).

[4] P. Holme, B. J. Kim, C. N. Yoon, and S. K. Han, *Attack Vulnerability of Complex Networks*, Phys. Rev. E **65**, 056109 (2002).

[5] R. Cohen, S. Havlin, and D. ben Avraham, *Efficient Immunization Strategies for Computer Networks and Populations*, Phys. Rev. Lett. **91**, 247901 (2003).

[6] P. Holme, *Efficient Local Strategies for Vaccination and Network Attack*, Europhys. Lett. **68**, 908 (2004).

[7] Y. Chen, G. Paul, S. Havlin, F. Liljeros, and H. E. Stanley, *Finding a Better Immunization Strategy*, Phys. Rev. Lett. **101**, 058701 (2008).

[8] J. Hadidjojo and S. A. Cheong, *Equal Graph Partitioning on Estimated Infection Network as an Effective Epidemic Mitigation Measure*, PLoS One **6**, e22124 (2011).

[9] C. M. Schneider, T. Mihaljev, S. Havlin, and H. J. Herrmann, *Suppressing Epidemics with a Limited Amount of Immunization Units*, Phys. Rev. E **84**, 061911 (2011).

[10] N. Masuda, *Immunization of Networks with Community Structure*, New J. Phys. **11**, 123018 (2009).

[11] M. Salathé and J. H. Jones, *Dynamics and Control of Diseases in Networks with Community Structure*, PLoS Comput. Biol. **6**, e1000736 (2010).

[12] C. Borgs, J. Chayes, A. Ganesh, and A. Saberi, *How to Distribute Antidote to Control Epidemics*, Random Struct. Algorithms **37**, 204 (2010).

[13] F. Chung, P. Horn, and A. Tsiatas, *Distributing Antidote Using PageRank Vectors*, Internet Math. **6**, 237 (2009).

[14] E. Gourdin, J. Omic, and P. Van Mieghem, in *Proceedings of the 8th International Workshop on the Design of Reliable Communication Networks (DRCN)* (IEEE, Krakow, 2011), pp. 86–93.

[15] F. Bauer and J. T. Lizier, *Identifying Influential Spreaders and Efficiently Estimating Infection Numbers in Epidemic Models: A Walk Counting Approach*, Europhys. Lett. **99**, 68007 (2012).

[16] V. M. Preciado, M. Zargham, C. Enyioha, A. Jadbabaie, and G. Pappas, *Optimal Vaccine Allocation to Control Epidemic Outbreaks in Arbitrary Networks*, arXiv:1303.3984.

[17] V. Kashirin and L. Dijkstra, *A Heuristic Optimization Method for Mitigating the Impact of a Virus Attack*, Procedia Computer Science **18**, 2619 (2013).

[18] D. Kempe, J. Kleinberg, and É. Tardos, in *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '03* (ACM, New York, NY, 2003), pp. 137–146.

[19] G. Giakkoupis, A. Gionis, E. Terzi, and P. Tsaparas, Technical Report No. C-2005-75, Department of Computer Science, University of Helsinki, 2005.

[20] J. Aspnes, K. Chang, and A. Yampolskiy, *Inoculation Strategies for Victims of Viruses and the Sum-of-Squares Partition Problem*, J. Comput. Syst. Sci. **72**, 1077 (2006).

[21] P.-A. Chen, M. David, and D. Kempe, in *Proceedings of the 11th ACM Conference on Electronic Commerce, EC '10* (ACM, New York, NY, 2010), pp. 179–188.

[22] C. Budak, D. Agrawal, and A. E. Abbadi, in *Proceedings of the 20th International Conference on World Wide Web, WWW '11* (ACM, New York, NY, 2011), pp. 665–674.

[23] U. H. Karkada, L. A. Adamic, J. M. Kahn, and T. J. Iwashyna, *Limiting the Spread of Highly Resistant Hospital-Acquired Microorganisms via Critical Care Transfers: A Simulation Study*, Intensive Care Medicine **37**, 1633 (2011).

[24] N. P. Nguyen, G. Yan, M. T. Thai, and S. Eidenbenz, in *Proceedings of the 3rd Annual ACM Web Science Conference, WebSci '12* (ACM, New York, NY, 2012), pp. 213–222.

[25] B. A. Prakash, L. Adamic, H. Iwashyna, H. Tong, and C. Faloutsos, *Fractional Immunization on Networks* (Austin, Texas, 2013).

[26] F. Altarelli, A. Braunstein, A. Ramezanpour, and R. Zecchina, *Stochastic Optimization by Message Passing*, J. Stat. Mech. (2011) P11009.

[27] F. Altarelli, A. Braunstein, A. Ramezanpour, and R. Zecchina, *Stochastic Matching Problem*, Phys. Rev. Lett. **106**, 190601 (2011).

[28] F. Altarelli, A. Braunstein, L. Dall'Asta, and R. Zecchina, *Large Deviations of Cascade Processes on Graphs*, Phys. Rev. E **87**, 062115 (2013).

[29] F. Altarelli, A. Braunstein, L. Dall'Asta, and R. Zecchina, *Optimizing Spread Dynamics on Graphs by Message Passing*, J. Stat. Mech. (2013) P09011.

[30] R. M. Anderson and R. M. May, *Infectious Diseases of Humans: Dynamics and Control* (Oxford University Press, Oxford, New York, 1991).

[31] H. W. Hethcote, *The Mathematics of Infectious Diseases*, SIAM Rev. **42**, 599 (2000).

[32] W. O. Kermack and A. G. McKendrick, *A Contribution to the Mathematical Theory of Epidemics*, Proc. R. Soc. A **115**, 700 (1927).

[33] M. E. J. Newman, *Spread of Epidemic Disease on Networks*, Phys. Rev. E **66**, 016128 (2002).

[34] L. Dall'Asta, *Inhomogeneous Percolation Models for Spreading Phenomena in Random Graphs*, J. Stat. Mech. (2005) P08011.

[35] B. Karrer and M. E. J. Newman, *Message Passing Approach for General Epidemic Models*, Phys. Rev. E **82**, 016101 (2010).

[36] D. Chakrabarti, Y. Wang, C. Wang, J. Leskovec, and C. Faloutsos, *Epidemic Thresholds in Real Networks*, ACM Trans. Inform. Syst. Secur. **10**, 1 (2008).

[37] C. Castellano and R. Pastor-Satorras, *Thresholds for Epidemic Spreading in Networks*, Phys. Rev. Lett. **105**, 218701 (2010).

[38] P. Van Mieghem, *The N-Intertwined SIS Epidemic Network Model*, Computing **93**, 147 (2011).
[39] P. Van Mieghem, *The Viral Conductance of a Network*, Comput. Commun. **35**, 1494 (2012).
[40] P. Van Mieghem, *Epidemic Phase Transition of the SIS Type in Networks*, Europhys. Lett. **97**, 48004 (2012).
[41] P. Van Mieghem and E. Cator, *Epidemics in Networks with Nodal Self-Infection and the Epidemic Threshold*, Phys. Rev. E **86**, 016116 (2012).
[42] E. Cator and P. Van Mieghem, *Susceptible-Infected-Susceptible Epidemics on the Complete Graph and the Star Graph: Exact Analysis*, Phys. Rev. E **87**, 012811 (2013).
[43] C. Li, R. van de Bovenkamp, and P. Van Mieghem, *Susceptible-Infected-Susceptible Model: A Comparison of N-Intertwined and Heterogeneous Mean-Field Approximations*, Phys. Rev. E **86**, 026116 (2012).
[44] B. Guerra and J. Gómez-Gardeñes, *Annealed and Mean-Field Formulations of Disease Dynamics on Static and Adaptive Networks*, Phys. Rev. E **82**, 035101 (2010).
[45] A. V. Goltsev, S. N. Dorogovtsev, J. G. Oliveira, and J. F. F. Mendes, *Localization and Spreading of Diseases in Complex Networks*, Phys. Rev. Lett. **109**, 128702 (2012).
[46] M. Shapiro and E. Delgado-Eckert, *Finding the Probability of Infection in an SIR Network is NP-Hard*, Math. Biosci. **240**, 77 (2012).
[47] M. Mézard, G. Parisi, and M. A. Virasoro, *Spin Glass Theory and Beyond* (World Scientific, Singapore, 1987), Vol. 9.
[48] M. Mézard and A. Montanari, *Information, Physics, and Computation* (Oxford University Press, Oxford, 2009).
[49] F. Krzakala, F. Ricci-Tersenghi, and L. Zdeborová, *Elusive Spin-Glass Phase in the Random Field Ising Model*, Phys. Rev. Lett. **104**, 207208 (2010).
[50] M. E. J. Newman, Network Dataset Collection, http://www-personal.umich.edu/~mejn/netdata/.
[51] W. W. Zachary, *An Information Flow Model for Conflict and Fission in Small Groups*, J. Anthropol. Res. **33**, 452 (1977).
[52] D. Lusseau, K. Schneider, O. J. Boisseau, P. Haase, E. Slooten, and S. M. Dawson, *The Bottlenose Dolphin Community of Doubtful Sound Features a Large Proportion of Long-Lasting Associations*, Behav. Ecol. Sociobiol. **54**, 396 (2003).
[53] D. Knuth, *The Stanford GraphBase: A Platform for Combinatorial Computing* (Addison-Wesley, Reading, MA, 1993).
[54] L. Isella, J. Stehlé, A. Barrat, C. Cattuto, J.-F. Pinton, and W. Van den Broeck, *What's in a Crowd? Analysis of Face-to-Face Behavioral Networks*, J. Theor. Biol. **271**, 166 (2011).
[55] J. Moody, *Peer Influence Groups: Identifying Dense Clusters in Large Networks*, Soc. Networks **23**, 261 (2001).
[56] M. Bayati, C. Borgs, A. Braunstein, J. Chayes, A. Ramezanpour, and R. Zecchina, *Statistical Mechanics of Steiner Trees*, Phys. Rev. Lett. **101**, 037208 (2008).
[57] M. Bailly-Bechet, C. Borgs, A. Braunstein, J. Chayes, A. Dagkessamanskaia, J.-M. François, and R. Zecchina, *Finding Undetected Protein Associations in Cell Signaling by Belief Propagation*, Proc. Natl. Acad. Sci. U.S.A. **108**, 882 (2011).