Dissemination strategy for immunizing scale-free networks

Alexandre O. Stauffer and Valmir C. Barbosa
*Programa de Engenharia de Sistemas e Computação, COPPE, Universidade Federal do Rio de Janeiro,
Caixa Postal 68511, 21941-972 Rio de Janeiro RJ, Brazil*

(Received 27 April 2006; published 6 November 2006)

We consider the problem of distributing a vaccine for immunizing a scale-free network against a given virus or worm. We introduce a method, based on vaccine dissemination, that seems to reflect more accurately what is expected to occur in real-world networks. Also, since the dissemination is performed using only local information, the method can be easily employed in practice. Using a random-graph framework, we analyze our method both mathematically and by means of simulations. We demonstrate its efficacy regarding the trade-off between the expected number of nodes that receive the vaccine and the network's resulting vulnerability to develop an epidemic as the virus or worm attempts to infect one of its nodes. For some scenarios, the method is seen to render the network practically invulnerable to attacks while requiring only a small fraction of the nodes to receive the vaccine.

DOI: 10.1103/PhysRevE.74.056105 PACS number(s): 89.75.Da, 89.20.Ff, 89.20.Hh, 89.75.Fb

I. INTRODUCTION

The term "scale-free" is widely used to designate the class of large networks that have node degrees distributed as a power law [1,2], according to which the probability that a randomly chosen node has degree a is proportional to $a^{-\tau}$ for some parameter $\tau > 1$. There has been a recent surge of interest in scale-free networks, as a great variety of large real-world networks, such as the Internet, the WWW, social networks, and scientific-collaboration networks, have been empirically observed to have node-degree distributions that approximately follow a power law [3,4]. In contrast with the classical random-graph model introduced by Erdős and Rényi, whose node-degree distribution is the Poisson distribution and is therefore sharply concentrated around its mean value [5,6], scale-free networks normally contain nodes with a wide range of degrees, typically with a few nodes of extremely high degrees coexisting with a plethora of low-degree nodes.

In this paper, we consider the problem of preventing viruses or worms from spreading on scale-free computer networks. The fact that node degrees are in this case distributed according to a power law has a profound impact on the way the network operates. In particular, it makes the problem of fighting the proliferation of viruses and other infections much more challenging, since the presence of high-degree nodes dramatically increases the rate at which a virus may propagate [7,8]. For this reason, instead of combating the proliferation of a virus in an already infected network, we consider a preventive immunization strategy, which consists of distributing the appropriate vaccine to a small subset of the network's nodes, striving to immunize those nodes that can more efficiently block the spread of a future infection, should it occur. The goal of this approach is to distribute the vaccine to as few nodes as possible while making the network invulnerable to an epidemic, that is, to the occurrence of a state in which a relatively large number of nodes is infected.

We note that, even though our focus is on preventing infections by viruses or worms in computer networks, most of our results and conclusions may also find application in other domains amenable to modeling by scale-free networks, provided the notions of an infection and of the transmittal of a vaccine from one node to another make sense. One example is the spreading of epidemics in human populations. Another is the proliferation of errors in networks of bibliographic citations.

We can measure the efficacy of an immunization strategy by two indicators: the expected spread, which is the expected fraction of the network's nodes that receive the vaccine, and the expected vulnerability, which is the expected fraction of the network's nodes that may become infected when the virus attempts to infect a randomly chosen node of the immunized network. Clearly, these two indicators are strongly influenced by how we select the nodes to receive the vaccine. A simple rule for choosing these nodes is to randomly select a given fraction of the network's nodes [8–10]. When applied to scale-free networks, we know that this rule normally gives unsatisfactory results, as it only achieves a reasonably small expected vulnerability for prohibitively high expected spreads. An alternative rule consists of distributing the vaccine to all the nodes that have degrees greater than a given value [8,9,11]. Despite being more efficient for scale-free networks than the previous strategy, as it achieves quite a small expected vulnerability with only a modest expected spread, applying this rule to real-world networks is known to be usually difficult [12]. The use of this rule demands global knowledge regarding the location of the nodes having the highest degrees, while the nodes of many real-world networks may only be assumed to have information that can be directly inferred from their immediate neighborhoods. Yet another alternative is to randomly choose some of the network's nodes and, for each of them, to immunize a randomly chosen fraction of its neighbors [12]. This rule, however, and in fact the previous two as well, seem hard to implement in practice on computer networks, since apparently they require that the vaccine be somehow transmitted to a given fraction of the network's nodes by means other than the network's own.

In this paper, we assume that the vaccine enters the network at a single node, called the originator. We assign to this node the responsibility of starting the dissemination of the vaccine by initiating the method called heuristic flooding for disseminating information in networks [13]. Let u be the originator. For each neighbor v of u, this method prescribes that u forward the vaccine to v with probability given by a heuristic function $h(a,b)$, where a and b are, respectively, the degrees of u and v. Each of the nodes that receive the vaccine, when receiving it for the first time, proceeds likewise and probabilistically forwards the vaccine to its own neighbors. By not requiring that the nodes of the network have information beyond what can be inferred from their immediate neighborhoods, this strategy can be easily used in practice. Furthermore, it represents more accurately what occurs in real scenarios, since it does not rely on the prior selection of nodes that characterizes all the three immunization strategies mentioned above, but rather assumes that the vaccine spreads out of a single node (say, the very site of its development or the site responsible for its distribution) via a heuristically controlled form of flooding.

Our immunization strategy shares with the chaining strategies of [14] the characteristic that the vaccine may enter the network at any node, from which it is then passed on to some of the other nodes of the network. Each of such chaining strategies embodies a different local policy whereby a node, having received the vaccine, selects one other node for being forwarded the vaccine if the fraction of immunized nodes in the network is still less than some preestablished value f. But the strategy that we introduce also differs from these chaining strategies in important aspects. One of them is that, if the forwarding policy at each node is deterministic (e.g., send the vaccine to the highest-degree neighbor), then cycling is a possibility and the fraction f of immunized nodes may never be achieved. Another, and perhaps the most important one, is that employing the desired fraction f of immunized nodes to control the progress of the forward chaining carries with it the inherent assumption that the number of nodes in the network is known. So, even though the chaining strategies of [14] are based on local decisions given the network's size, requiring that such size be known bespeaks a dependency on global properties, just as for one of the strategies we discussed above.

Our strategy has neither of these drawbacks and we think this is to be credited to a fundamental difference in how the immunization problem is approached. Instead of aiming at immunizing some given fraction of the network's nodes, what it seeks is to provide a heuristic function that, for a given class of networks (in the case of this paper, scale-free networks), can be expected to immunize as small a fraction of the network's nodes as possible while providing a significant level of invulnerability. Moreover, it does so independently of any network-wide properties, and may then be regarded, to the best of our knowledge, as the first of a kind.

It is also worth mentioning that our strategy need not assume that the network is completely uninfected to begin with. In fact, in many practical scenarios it is the case that both immunizing and healing rely on the exact same vaccine. This is the case, for example, of computer viruses. In such cases, what our strategy prescribes is that some nodes be immunized/healed as the vaccine propagates from its node of entry in the network; as for the remaining nodes, should any of them be already infected, the guarantee exists that the infection will be contained. In a similar vein, our strategy may also lend itself gracefully to the dissemination of new protection measures that are specific to no infecting agent in particular. While such a generic approach is arguably full of difficulties, it seems to be favored by many researchers (cf. [15] and references therein).

We organize the remainder of the paper as follows. In Sec. II, we use a random-graph framework and the formalism introduced in [16–18], whose details are discussed as they are needed, to obtain mathematical results for the aforementioned efficacy indicators. We utilize our analytical results in Sec. III to discover the properties that an ideal heuristic function should have to be efficient. We then introduce a heuristic function that seeks to approximate this ideal and therefore can be used to disseminate the vaccine. In Sec. IV, we discuss simulation results on random graphs having node degrees distributed according to a power law. Our results reveal that this heuristic function performs very attractively for the ranges of τ (the distribution's parameter) that typically are thought to hold for networks like the Internet (i.e., τ below roughly 2.5). They also agree satisfactorily with our analytical predictions. We conclude in Sec. V.

II. MATHEMATICAL ANALYSIS

Let G be a random graph having n nodes, whose degrees are distributed independently from one another and identically to a random variable K_G. We assume that the nodes of G are interconnected in an independent way given their degrees, which therefore remain independent. We base our mathematical analysis of this section on the formalism introduced in [18] and target the case in which G has a formally infinite number of nodes. [We also adopt the usual notation for the asymptotic behavior of functions: $f(x)$ is $\Theta(g(x))$ if and only if there exist positive constants C_1, C_2, and x_0 such that $C_1 g(x) \leqslant f(x) \leqslant C_2 g(x)$ for all $x \geqslant x_0$; $f(x)$ is $o(g(x))$ if and only if $\lim_{x\to\infty} f(x)/g(x)=0$.]

Let $P_G(a)$ be the probability that a randomly chosen node of G has degree a, i.e., the probability that $K_G=a$. The average degree in G, denoted by Z_G, is clearly

$$Z_G = \sum_{a=0}^{n-1} a P_G(a). \tag{1}$$

Given that the degrees of two adjacent nodes are independent from each other, the probability that some node's neighbor has degree b is identical to the expected fraction of edges incident to degree-b nodes, which is given by

$$\frac{bP_G(b)}{\sum_{a=0}^{n-1} aP_G(a)} = \frac{bP_G(b)}{Z_G}. \tag{2}$$

From [16,18], a necessary and sufficient condition for a size-$\Theta(n)$ connected component to almost surely exist in G (i.e., to exist with a probability that approaches 1 as $n\to\infty$) is that

$$\sum_{b=1}^{n-1}(b-1)\frac{bP_G(b)}{Z_G} > 1, \tag{3}$$

which intuitively means that, given a randomly chosen node u of G, a size-$\Theta(n)$ connected component exists almost surely if and only if a neighbor of u is expected to have more than one neighbor besides u. We denote such a size-$\Theta(n)$ connected component of G (its giant connected component) by GCC_G. When (3) is satisfied, with high probability all the other connected components of G are small, comprising only $o(n)$ nodes, and G is said to be above the phase transition that gives rise to GCC_G. On the other hand, when (3) is not satisfied, with high probability all the connected components of G are small, each consisting of $o(n)$ nodes, and G is said to be below the phase transition that gives rise to GCC_G.

Given a randomly chosen node u of G and a neighbor v of u, we define the reach of u through v as the set of nodes that can be reached by a path starting at u and whose first edge is (u,v). A node belongs to GCC_G if and only if it has at least one neighbor through which its reach contains a large, size-$\Theta(n)$ number of nodes. Let q be the probability that a node has a small, size-$o(n)$ reach through a given neighbor. The probability that a degree-a node belongs to GCC_G is then $1-q^a$, and the probability that a randomly chosen node of G belongs to GCC_G, which we denote by θ_G, is

$$\theta_G = 1 - \sum_{a=0}^{n-1} q^a P_G(a). \tag{4}$$

The probability q that u has a small, size-$o(n)$ reach through v can be obtained from the probability that v itself has a small, size-$o(n)$ reach through each of its other neighbors (i.e., excluding u). Since the probability that two neighbors of u have another common neighbor (i.e., besides u) varies with n proportionally to n^{-1} [18], which for large n is negligible, the probability that v has a small, size-$o(n)$ reach through a given neighbor is also q, thus leading to

$$q = \sum_{b=1}^{n-1} q^{b-1}\frac{bP_G(b)}{Z_G}. \tag{5}$$

This equation can be solved numerically and then used in Eq. (4) to obtain θ_G.

From now on, we assume that G is above the phase transition and, therefore, GCC_G exists. Furthermore, since G can be unconnected and real-world computer networks are normally connected, we assume that it is the graph induced by GCC_G, rather than G itself, that models the network, and also condition the remainder of our analysis accordingly.

A. Expected spread

In this section, we calculate the expected spread in GCC_G, which is denoted by P_{s} and consists of the expected fraction of the nodes of GCC_G that are immunized when a vaccine is distributed using the heuristic flooding described in Sec. I. We resort to the same method of analysis developed in [13]. Let S be a directed subgraph of G that spans all the nodes of G. For a degree-a node u and a degree-b neighbor v of u in G, the probability that the directed edge $(u\rightarrow v)$ exists in S is given by $h(a,b)$, the heuristic function employed during the vaccine dissemination. Before proceeding to the calculation of P_{s}, we pause for a brief study of S.

The neighbors of a node u in S can be classified into two different types: the in-neighbors, those from which an edge exists directed toward u; and the out-neighbors, those toward which an edge exists directed from u. If a directed path exists starting at some node u and ending at another node v, then we say that u reaches v in S or that v is in the reach of u in S. Note that if u receives the vaccine, then the reach of u in S is part of the set of nodes that become immunized.

The connected components of a directed graph can also be of two basic types. First, there are the weakly connected components, which are constituted by the nodes that can reach one another by undirected paths, i.e., paths for which the directions of the edges are disregarded. The other type is that of the strongly connected components, each comprising a maximal set of nodes that can both reach and be reached from one another.

Similarly to the case of the undirected graph G, there is a criterion for deciding whether S almost surely has a size-$\Theta(n)$ weakly connected component, commonly known as the giant weakly connected component of S, denoted by GWCC_S. Likewise, there is another criterion according to which S almost surely has a size-$\Theta(n)$ strongly connected component, commonly referred to as the giant strongly connected component, denoted by GSCC_S. Clearly, when both GWCC_S and GSCC_S exist, as we henceforth assume, all the nodes of GSCC_S belong also to GWCC_S, and all the nodes of GWCC_S belong also to GCC_G.

Since GSCC_S exists by assumption, we can define two other size-$\Theta(n)$ connected components of S, which we refer to as the giant in-component (GIN_S), formed by the nodes that can reach GSCC_S, and the giant out-component (GOUT_S), formed by the nodes reachable from GSCC_S. Note that, by definition, the nodes of GSCC_S belong also to both GIN_S and GOUT_S. We denote by θ_S^{in} and θ_S^{out} the expected fraction of the nodes of G that belong to, respectively, GIN_S and GOUT_S. Figure 1 illustrates an instance of graph G [Fig. 1(a)] and a possible instance of its directed subgraph S [Fig. 1(b)].

Assuming that the originator is randomly chosen among the nodes of GCC_G, the vaccine is guaranteed to be distributed to a size-$\Theta(n)$ set of nodes if the originator belongs to GIN_S, which happens with probability $\theta_S^{\mathrm{in}}/\theta_G$. When this is the case, the nodes that receive the vaccine either belong to GOUT_S, corresponding to a fraction $\theta_S^{\mathrm{out}}/\theta_G$ of the nodes of GCC_G, or are not in GOUT_S despite being reachable from the originator, and then amount to a small, size-$o(n)$ number of nodes. Neglecting the latter nodes is equivalent to assuming that nodes receive the vaccine only if the originator is in GIN_S. In this case, only the nodes in GOUT_S receive the vaccine and we have

$$P_{\mathrm{s}} = \frac{\theta_S^{\mathrm{in}}\theta_S^{\mathrm{out}}}{\theta_G^2}. \tag{6}$$

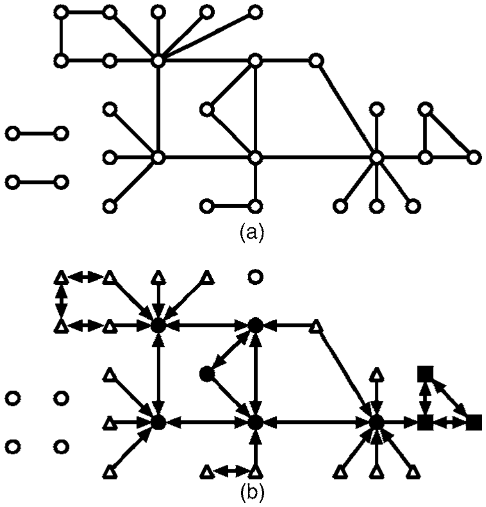

FIG. 1. A G instance (a) and one possible instance of the directed subgraph S of the G instance (b). Part (b) also shows the nodes belonging to GSCC_S (filled circles), GIN_S (filled circles and triangles), and GOUT_S (filled circles and filled squares).

In order to obtain θ_S^{in}, recall that the nodes of GIN_S are the only ones that have a non-negligible reach. Considering a degree-a node u of G and a degree-b neighbor v of u in G, we say that v is a dead end with respect to u in S if either $(u \to v)$ is not an edge of S, or it is but the reach of u through v in S is negligible, consisting of only $o(n)$ nodes. Denoting by q_b^{in} the conditional probability that the reach of u through v in S is negligible given that u is an in-neighbor of v in S, we obtain the probability that v is a dead end with respect to u in S, which is

$$1 - h(a,b) + h(a,b)q_b^{\mathrm{in}}. \tag{7}$$

And since the probability that v has degree b is given by Eq. (2), the probability that a given neighbor of a degree-a node is a dead end with respect to it in S, which we denote by w_a^{in}, is

$$w_a^{\mathrm{in}} = \sum_{b=1}^{n-1} [1 - h(a,b) + h(a,b)q_b^{\mathrm{in}}]\frac{bP_G(b)}{Z_G}. \tag{8}$$

Because a node belongs to GIN_S if and only if at least one of its neighbors in G is not a dead end with respect to it in S, we arrive at

$$\theta_S^{\mathrm{in}} = 1 - \sum_{a=0}^{n-1} (w_a^{\mathrm{in}})^a P_G(a). \tag{9}$$

As a means to calculate q_b^{in}, let us consider a degree-b node v of G reached by following a directed edge $(u \to v)$ of S. The reach of u through v in S is negligible, which happens with probability q_b^{in}, if and only if all of the other $b-1$ neighbors of v in G (i.e., excluding u) are themselves dead ends with respect to v in S. This clearly leads to

$$q_b^{\mathrm{in}} = (w_b^{\mathrm{in}})^{b-1}. \tag{10}$$

Equations (8) and (10) can be put together to yield another equation where w_a^{in} is a function of all the other w^{in}'s. This equation can then be solved numerically to obtain θ_S^{in} via Eq. (9).

We can follow a completely analogous derivation and obtain θ_S^{out} by noting that a node belongs to GOUT_S if and only if it can be reached from a size-$\Theta(n)$ set of nodes. Let u be a degree-a node of G and v a neighbor of u in G. We denote by w_a^{out} the probability that either u is not an out-neighbor of v in S or is but the number of nodes that can reach u through v in S is small, consisting of only $o(n)$ nodes. Also, we denote by q_b^{out} the conditional probability that the number of nodes that can reach u through v in S is small, given that the degree of v in G is b and u is an out-neighbor of v. In a way analogous to the one that led to Eqs. (8)–(10), we obtain

$$w_a^{\mathrm{out}} = \sum_{b=1}^{n-1} [1 - h(b,a) + h(b,a)q_b^{\mathrm{out}}]\frac{bP_G(b)}{Z_G}, \tag{11}$$

$$\theta_S^{\mathrm{out}} = 1 - \sum_{a=0}^{n-1} (w_a^{\mathrm{out}})^a P_G(a), \tag{12}$$

and

$$q_b^{\mathrm{out}} = (w_b^{\mathrm{out}})^{b-1}. \tag{13}$$

Also, and identically to the derivation of θ_S^{in}, we can unify Eqs. (11) and (13) and calculate the value of each w_a^{out} numerically to obtain θ_S^{out} via Eq. (12).

B. Expected vulnerability

Consistently with the simplifying assumptions of Sec. II A, we keep assuming that no node is immunized when the originator does not belong to GIN_S. When this happens, all nodes of GCC_G remain vulnerable to the virus, and if the virus infects a node of GCC_G it may propagate until the entire GCC_G is infected. Let us analyze the case in which the originator does belong to GIN_S.

As before, we assume that only the nodes of GOUT_S receive the vaccine. Let V be an undirected subgraph of G that spans all the nodes of G, and let an edge (u,v) of G belong to V if and only if neither u nor v belongs to GOUT_S. That is, given a certain instance of the subgraph S, subgraph V contains all the edges of G that are not incident to nodes of GOUT_S. Clearly, the edges of V represent the edges through which the virus may propagate if it reaches either of an edge's (unimmunized) end nodes. Figure 2 illustrates the subgraph V corresponding to the G and S instances of Fig. 1.

Once again, and similarly to the case of G, a criterion exists for deciding whether a size-$\Theta(n)$ connected component almost surely exists in V. We denote such a component by GCC_V. When it does exist, and since all the other connected components of V contain with high probability only $o(n)$ nodes (which we again neglect), a virus may only proliferate into a large, size-$\Theta(n)$ set of nodes if it first infects a node of GCC_V. This, of course, is predicated upon the originator being in GIN_S and dissemination taking place exclusively inside GOUT_S, the assumptions of Sec. II A.

We define the expected vulnerability of GCC_G, denoted by P_v, as the fraction of the nodes of GCC_G that may become

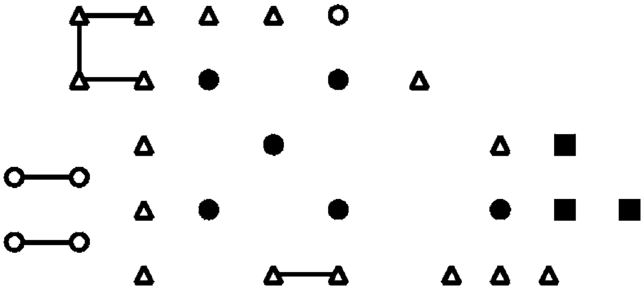

FIG. 2. The graph V that corresponds to the G and S instances of Fig. 1. Nodes represented by filled circles or filled squares belong to GOUT_S.

infected when the virus attempts to infect a randomly chosen node of GCC_G. Let θ_V be the fraction of the nodes of G that belong to GCC_V. If the originator does not belong to GIN_S (which occurs with probability $1-\theta_S^{\mathrm{in}}/\theta_G$), then $P_{\mathrm{v}}=1$; if it does belong to GIN_S (with probability $\theta_S^{\mathrm{in}}/\theta_G$), then $P_{\mathrm{v}}=\theta_V/\theta_G$ if and only if the virus first infects a node of GCC_V, which occurs with probability θ_V/θ_G. We then have

$$P_{\mathrm{v}} = 1 - \frac{\theta_S^{\mathrm{in}}}{\theta_G} + \frac{\theta_S^{\mathrm{in}}}{\theta_G}\left(\frac{\theta_V}{\theta_G}\right)^2. \tag{14}$$

Henceforth in this section we concentrate on calculating θ_V for the case in which GCC_V does exist. Clearly, a node of G belongs to GCC_V only if it does not belong to GOUT_S. Through the remainder of the section, let u be a degree-a node of G that does not belong to GOUT_S and v a neighbor of u in G. Given that v has degree b, we define $h_{b|a}$ as the probability that the edge $(v\to u)$ exists in S. Since u does not belong to GOUT_S, node v must be such that it satisfies one of the following conditions: either edge $(v\to u)$ does not exist in S, which happens with probability $1-h(b,a)$, or $(v\to u)$ exists in S but the number of nodes that can reach u through v is small, which occurs with probability $h(b,a)q_b^{\mathrm{out}}$. We can then express $h_{b|a}$ as the ratio of the probability that the latter condition is satisfied to the probability that either the former or the latter is. This leads to

$$h_{b|a} = \frac{h(b,a)q_b^{\mathrm{out}}}{1-h(b,a)+h(b,a)q_b^{\mathrm{out}}}. \tag{15}$$

Now let $p_{b|a}$ be the probability that v has degree b in G. Clearly, $p_{b|a}$ is proportional to the joint probability that v satisfies one of the above conditions regarding the existence of edge $(v\to u)$ in S and also that a node's neighbor in G has degree b. That is, $p_{b|a}$ is proportional to $[1-h(b,a)+h(b,a)q_b^{\mathrm{out}}]bP_G(b)/Z_G$. Using Eq. (11), we obtain

$$p_{b|a} = \left(\frac{1-h(b,a)+h(b,a)q_b^{\mathrm{out}}}{w_a^{\mathrm{out}}}\right)\frac{bP_G(b)}{Z_G}. \tag{16}$$

Let b be the degree of v in G. Because u does not belong to GOUT_S, nodes u and v are neighbors in V if and only if v does not belong to GOUT_S either. If $(v\to u)$ is an edge of S, which occurs with probability $h_{b|a}$, then v is obviously not in GOUT_S, as it would otherwise make u belong to GOUT_S along with it. On the other hand, if $(v\to u)$ is not an edge of S (with probability $1-h_{b|a}$), then v does not belong to GOUT_S if and only if the number of nodes that can reach it in S is small, which happens with probability q_b^{out}. It follows that the probability that u and v are neighbors in V is given by

$$h_{b|a} + (1-h_{b|a})q_b^{\mathrm{out}}. \tag{17}$$

When u and v are indeed neighbors in V, we define q_b^V as the probability that u has a small reach in V through v. We say that v is a dead end with respect to u in V if either v is not a neighbor of u in V, which occurs with probability $1-[h_{b|a}+(1-h_{b|a})q_b^{\mathrm{out}}]$, or it is but the reach of u through v in V is small, which occurs with probability $[h_{b|a}+(1-h_{b|a})q_b^{\mathrm{out}}]q_b^V$. Thus, the probability that v is a dead end with respect to u in V is

$$\begin{aligned}1 - [h_{b|a} + (1-h_{b|a})q_b^{\mathrm{out}}] + [h_{b|a} + (1-h_{b|a})q_b^{\mathrm{out}}]q_b^V \\ = h_{b|a}q_b^V + (1-h_{b|a})(1-q_b^{\mathrm{out}}+q_b^{\mathrm{out}}q_b^V),\end{aligned} \tag{18}$$

so the probability that a neighbor of u is a dead end with respect to u in V, which we denote by w_a^V, is clearly

$$w_a^V = \sum_{b=1}^{n-1}[h_{b|a}q_b^V + (1-h_{b|a})(1-q_b^{\mathrm{out}}+q_b^{\mathrm{out}}q_b^V)]p_{b|a}. \tag{19}$$

In order to calculate q_b^V, notice that the reach of u through v in V is small if and only if all other $b-1$ neighbors of v in G are themselves dead ends with respect to v in V. Then, assuming that the degrees of a node's neighbors in G remain independent from one another even under the condition that the node does not belong to GOUT_S, we have

$$q_b^V = (w_b^V)^{b-1}. \tag{20}$$

Putting Eqs. (19) and (20) together leads to an equation where w_a^V is a function of all the other w^V's, which can then be solved numerically for $0\leqslant a\leqslant n-1$.

We are finally in a position to calculate the value of θ_V. Let u be a randomly chosen node of G having degree a. In order to belong to GCC_V, node u must not belong to GOUT_S, which occurs with probability $(w_a^{\mathrm{out}})^a$. Furthermore, u belongs to GCC_V only if at least one of its neighbors is not a dead end with respect to it in V, which occurs with probability $1-(w_a^V)^a$. It then follows that

$$\theta_V = \sum_{a=0}^{n-1}(w_a^{\mathrm{out}})^a[1-(w_a^V)^a]P_G(a). \tag{21}$$

III. THE HEURISTIC FUNCTION

The efficiency of heuristic flooding as a means of immunizing a network depends heavily on the choice of the heuristic function $h(a,b)$. Before introducing our heuristic function, we elaborate on the properties of subgraph S that we may expect to lead to good results for P_{s} and P_{v}.

First of all, it is clear that S must be above the phase transition that gives rise to GSCC_S, thereby guaranteeing that GSCC_S, GIN_S, and GOUT_S almost surely exist. When this is the case, the nodes of GIN_S are the most suitable ones for being the originator, as they can immunize a non-negligible number of nodes. But since we cannot assume any prior information on the originator, GIN_S should contain as many nodes as possible in order to make the probability that the originator is chosen from outside it as small as possible. With regard to GOUT_S, we know that it contains the nodes that receive the vaccine when the originator belongs to GIN_S. In order to prevent an excessive number of nodes from receiving the vaccine, the size of GOUT_S should be kept to modest values. Putting these two observations together, we ideally want GIN_S to span all the nodes of the network, GSCC_S to contain only the nodes that can more efficiently block the spreading of an infection, and GOUT_S to be the same as GSCC_S.

Since we know that immunizing the nodes with the highest degrees is an efficient way to prevent epidemics in scale-free networks [8,9,11], we introduce in this section a heuristic function that stimulates the transmission of the vaccine to high-degree nodes. Introducing a parameter $\alpha \geqslant 0$, and considering a degree-a node u that has the vaccine and a degree-b neighbor v of u, our heuristic function $h(a,b)$, which gives the probability that u sends the vaccine to v, is defined as follows:

(i) If $b=1$, that is, v has no neighbor besides u, then $h(a,b)=0$ and u deterministically decides not to send the vaccine to v. In this case, since u is already immune, should v become infected it can transmit the virus to no other node, so we choose not to give v the vaccine.

(ii) If $a \leqslant 2 \leqslant b$, that is, u has degree at most 2 and v has degree at least 2, then $h(a,b)=1$ and u deterministically decides to send the vaccine to v. This is meant to force some low-degree nodes to forward the vaccine, thereby precluding a premature conclusion of heuristic flooding and, as a consequence, leading to a larger GIN_S.

(iii) For all the other positive values of a and b, we let

$$h(a,b) = \tanh\left(\frac{b-1}{(a-2)^{\alpha}}\right). \tag{22}$$

Clearly, for fixed $a > 2$, $h(a,b)$ increases with b, so the vaccine is more likely to be transmitted to high-degree nodes. For fixed $b > 1$, $h(a,b)$ decreases with a, thus reflecting the intuition that, when u is a high-degree node, sending the vaccine to v may be unnecessary even if v is a high-degree node (there are probably other paths through which the vaccine can be transmitted from u to v).

Figure 3 shows two plots illustrating the heuristic function of Eq. (22) for $\alpha=0.7$ [Fig. 3(a)] and $\alpha=1.0$ [Fig. 3(b)].

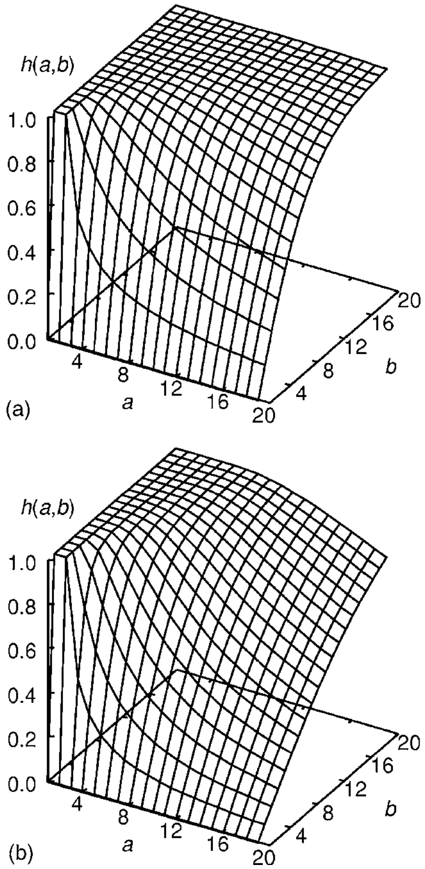

FIG. 3. Plots of the heuristic function given by Eq. (22) for $\alpha=0.7$ (a) and $\alpha=1.0$ (b).

IV. SIMULATION RESULTS

We have conducted extensive simulations on random graphs with node degrees distributed according to a power law. Generating such a graph is achieved in two phases [18]. Let $u_1, u_2, \ldots, u_n$ be the nodes of the random graph we want to generate. In the first phase, for $i=1,\ldots,n$ we sample the degree d_i of each u_i from the power-law distribution, obtaining the so-called degree sequence of the graph. If $\Sigma_{i=1}^{n} d_i$ turns out to be odd, then we discard the entire degree sequence and sample a new one, repeating the process until the sum of the degrees comes out even. In the second phase, we consider an imaginary urn having $\Sigma_{i=1}^{n} d_i$ labeled balls, the labels of d_i of them being u_i. We then successively remove pairs of balls from the urn until it has no more balls. For each pair we remove—say, of labels u_i and u_j—we add edge (u_i, u_j) to the graph. This method can produce graphs having multiple edges (more than one edge joining the same two nodes) or self-loops (an edge joining a node to itself), but it has the advantage of generating graphs whose degrees remain independent even after the edges are added, which is a core assumption of our analysis.

We carried out our simulations for $n=10\,000$ and $2 \leqslant \tau \leqslant 3$. For each value of τ, we generated 500 G instances [each one almost surely above the phase transition that gives rise to GCC_G, since in the scale-free case the condition in (3) becomes $\tau < 3.47$ [13]]. For each G instance, we used the heuristic $h(a,b)$ to both sample 1000 instances of the subgraph S and, in an independent way, conduct 1000 vaccine disseminations by heuristic flooding from an originator selected randomly among the nodes of the largest connected component of the G instance. For each S instance, we selected the largest strongly connected component and calculated the sizes of the corresponding in-component (counting the nodes that can reach the strongly connected component) and out-component (counting the nodes that can be reached

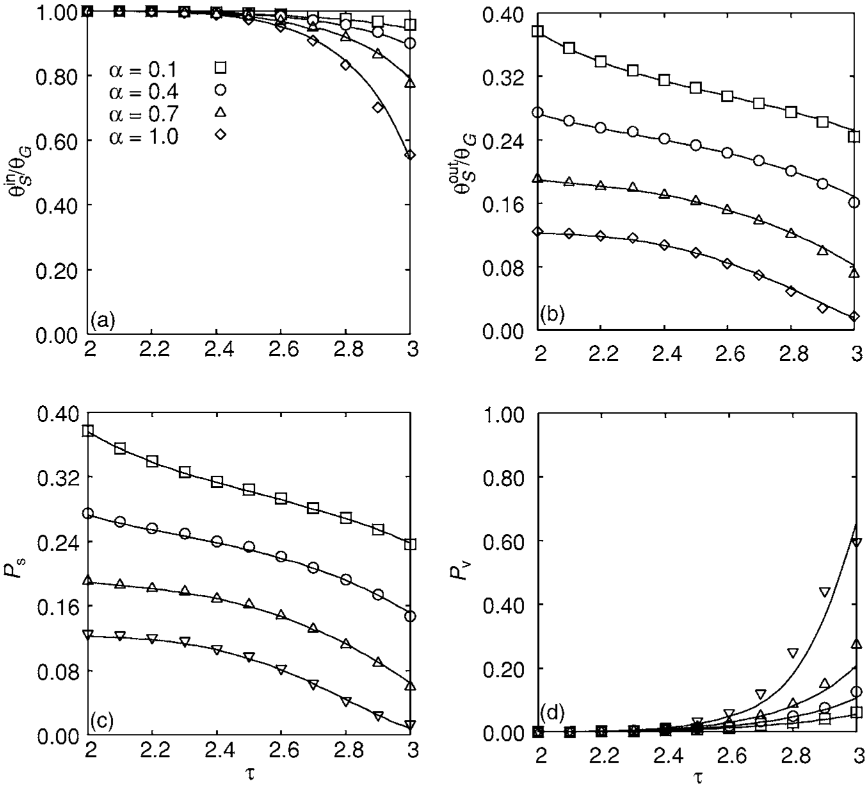

FIG. 4. Simulation results of vaccine dissemination by heuristic flooding. Solid lines give the analytic predictions.

from the strongly connected component). We then obtained the expected sizes of GIN_S and GOUT_S by averaging these quantities over the 500 000 samples. For each vaccine dissemination, we calculated the fraction of nodes that receive the vaccine and the fraction of nodes to which an infection may spread when an attempt at infecting a randomly chosen node inside the largest connected component of G takes place. We then obtained P_s and P_v by averaging these quantities over the 500 000 samples.

Simulation results are shown in Fig. 4 for $\alpha=0.1$, 0.4, 0.7, and 1.0. We note, in general, a satisfactory agreement between analytic and simulation results, with the exception of part (d), in which case the deviation may be attributed to the approximations made during the derivation of θ_V in Sec. II B to yield Eq. (21).

When $\tau \leqslant 2.5$, the plots for θ_S^{in}/θ_G and θ_S^{out}/θ_G [Figs. 4(a) and 4(b)] reveal that the heuristic function introduced in Sec. III results in a GIN_S that spans almost all the nodes of GCC_G, while the size of GOUT_S keeps to a relatively modest fraction of GCC_G. For example, for $\tau \leqslant 2.5$ and $\alpha=1.0$, the relative size of GIN_S is always above 0.97 and the relative size of GOUT_S is always below 0.13. For $\tau > 2.5$, the relative size of GIN_S decreases with τ, thus evidencing that heuristic flooding has more difficulty disseminating the vaccine when the graph is sparser.

Owing to P_s being given by $(\theta_S^{in}/\theta_G)(\theta_S^{out}/\theta_G)$ [cf. Eq. (6)], and to θ_S^{in}/θ_G being relatively close to 1 [Fig. 4(a)], the plots for P_s [Fig. 4(c)] are of course similar to the plots for θ_S^{out}/θ_G [Fig. 4(b)]. Furthermore, given a value of α, P_s decreases with τ, which means that heuristic flooding spreads through a smaller number of nodes when the graph is sparser, as, in this case, there are fewer paths conducting to the high-degree nodes.

As for P_v [Fig. 4(d)], we note that, for $\tau \leqslant 2.5$, P_v is nearly zero. This result is a natural consequence both of the guiding principle of the heuristic introduced in Sec. III, which ascribes more probability for transmitting the vaccine to nodes having higher degrees, and of the result for θ_S^{in}/θ_G [Fig. 4(a)], which indicates that GIN_S spans almost all the nodes of GCC_G. As τ is increased to values greater than 2.5, P_v moves farther away from zero, since the size of GIN_S decreases and, therefore, the probability that heuristic flooding distributes the vaccine to only a small number of nodes increases. Regarding the value of α, we note a clear trade-off between P_s and P_v. If we were to adjust α in such a way as to decrease P_s, we would have an increase in P_v, which shows that the number of immunized nodes has a direct impact on the resulting vulnerability of the network.

V. CONCLUSION

We have considered in this paper the problem of immunizing a scale-free network against a virus or worm. We introduced an immunization strategy that we believe reflects more accurately what happens in real scenarios. In our strategy, we assume that the vaccine enters the network at exactly one node, in general the site of the vaccine's development or the site in charge of its distribution, for example. This node begins the dissemination of the vaccine by heuristic flooding, aiming at immunizing the nodes that have the highest degrees. With this purpose in mind, we introduced a heuristic function that gives more probability to forwarding the vaccine toward nodes with higher degrees.

We obtained analytical and simulation results on random graphs having node degrees distributed according to a power law. Our mathematical analysis has innovative aspects that we expect may shed some light on obtaining analytical results for similar distributed algorithms. Also, we hope our analysis can contribute to the development of new heuristic functions for vaccine dissemination. With regard to our simulation results, they show satisfactory agreement with our mathematical analysis and highlight the expected trade-off between the number of nodes that receive the vaccine and the vulnerability of the network to future infections. Especially for power laws with relatively small value for the parameter τ, our heuristic function achieves very good results, making the network practically invulnerable to an epidemic while requiring the immunization of only roughly 10% of the nodes.

Our strategy, however, is not without its drawbacks. For example, when GIN_S spans only a relatively small fraction of GCC_G, and depending on the values of τ and α, the probability that only very few nodes are immunized is non-negligible and given by the complement of the data in Fig. 4(a). In addition, and notwithstanding the fact that for every value of τ in the range we studied there exists a value of α that yields satisfactory expected invulnerability [cf. Fig. 4(d)], in some cases deciding whether to apply the method depending on the value of τ may be an issue. While in such cases requiring knowledge of the value of τ would bespeak an inherent dependency of the strategy upon a global property of the network, we note that such knowledge is not needed in general. Rather, it would be sufficient in such cases merely to estimate how the value of τ relates to the apparent threshold of roughly 2.5 that seems to divide the very successful scenarios from the others. But such an estimate is intimately related to how connected G is, which seems to be a much simpler property of which the individual nodes may have some knowledge.

We note, finally, that one possible direction in which this paper's research may be extended, in addition to the search for other heuristic functions, is that of allowing for multiple concurrent initiators. While algorithmically (i.e., from the perspective of flooding the network) such an extension is trivial, extending the analysis of Sec. II is expected to be a significantly more complex endeavor.

ACKNOWLEDGMENTS

The authors acknowledge partial support from CNPq, CAPES, and a FAPERJ BBP grant.

[1] A.-L. Barabási and R. Albert, Science **286**, 509 (1999).

[2] M. E. J. Newman, SIAM Rev. **45**, 167 (2003).

[3] M. Faloutsos, P. Faloutsos, and C. Faloutsos, in *Proceedings of ACM SIGCOMM* (ACM Press, New York, 1999), pp. 251–262.

[4] R. Albert and A.-L. Barabási, Rev. Mod. Phys. **74**, 47 (2002).

[5] P. Erdős and A. Rényi, Publ. Math. (Debrecen) **6**, 290 (1959).

[6] B. Bollobás, *Random Graphs*, 2nd ed. (Cambridge University Press, Cambridge, UK, 2001).

[7] R. Pastor-Satorras and A. Vespignani, Phys. Rev. Lett. **86**, 3200 (2001).

[8] R. Pastor-Satorras and A. Vespignani, in *Handbook of Graphs and Networks*, edited by S. Bornholdt and H. G. Schuster (Wiley-VCH, Weinheim, 2003), pp. 111–130.

[9] R. Albert, H. Jeong, and A.-L. Barabási, Nature (London) **406**, 378 (2000).

[10] R. Cohen, K. Erez, D. ben-Avraham, and S. Havlin, Phys. Rev. Lett. **85**, 4626 (2000).

[11] R. Cohen, K. Erez, D. ben-Avraham, and S. Havlin, Phys. Rev. Lett. **86**, 3682 (2001).

[12] R. Cohen, S. Havlin, and D. ben-Avraham, Phys. Rev. Lett. **91**, 247901 (2003).

[13] A. O. Stauffer and V. C. Barbosa, IEEE/ACM Trans. Networking (to be published), URL http://arxiv.org/abs/cs.NI/0409001

[14] P. Holme, Europhys. Lett. **68**, 908 (2004).

[15] M. Costa, J. Crowcroft, M. Castro, A. Rowstron, L. Zhou, L. Zhang, and P. Barham, ACM SIGOPS Operating Syst. Rev. **39**, 133 (2005).

[16] M. Molloy and B. Reed, Random Struct. Algorithms **6**, 161 (1995).

[17] M. Molloy and B. Reed, Combinatorics, Probab. Comput. **7**, 295 (1998).

[18] M. E. J. Newman, S. H. Strogatz, and D. J. Watts, Phys. Rev. E **64**, 026118 (2001).