Contents lists available at ScienceDirect

Physica A

journal homepage: www.elsevier.com/locate/physa

Effects of epidemic threshold definition on disease spread statistics

C. Lagorio [a,*], M.V. Migueles [a], L.A. Braunstein [a,b], E. López [c], P.A. Macri [a]

[a] Instituto de Investigaciones Físicas de Mar del Plata (IFIMAR), Departamento de Física, Facultad de Ciencias Exactas y Naturales, Universidad Nacional de Mar del Plata-CONICET, Funes 3350, (7600) Mar del Plata, Argentina

[b] Center for Polymer Studies, Boston University, Boston, MA 02215, USA

[c] CABDyN Complexity Cluster and Department of Physics, University of Oxford, Park End Street Oxford, OX1 1HP, United Kingdom

ARTICLE INFO

Article history:
Received 25 July 2008
Received in revised form 21 October 2008
Available online 9 November 2008

PACS:
64.60.ah
87.23.Ge
89.75.-k

Keywords:
Epidemic spread on networks
Percolation

ABSTRACT

We study the statistical properties of SIR epidemics in random networks, when an epidemic is defined as only those SIR propagations that reach or exceed a minimum size s_c. Using percolation theory to calculate the average fractional size $\langle M_{\mathrm{SIR}} \rangle$ of an epidemic, we find that the strength of the spanning link percolation cluster P_∞ is an upper bound to $\langle M_{\mathrm{SIR}} \rangle$. For small values of s_c, P_∞ is no longer a good approximation, and the average fractional size has to be computed directly. We find that the choice of s_c is generally (but not always) guided by the network structure and the value of T of the disease in question. If the goal is to always obtain P_∞ as the average epidemic size, one should choose s_c to be the typical size of the largest percolation cluster at the critical percolation threshold for the transmissibility. We also study Q, the probability that an SIR propagation reaches the epidemic mass s_c, and find that it is well characterized by percolation theory. We apply our results to real networks (DIMES and Tracerouter) to measure the consequences of the choice s_c on predictions of average outcome sizes of computer failure epidemics.

The study of disease spread has seen renewed interest recently [1–3] due to the emergence of new infectious lethal diseases such as AIDS and SARS [4,5]. New tools, ranging from powerful computer models [6] to new conceptual developments [1,8–12], have emerged in hopes of understanding and addressing the problem effectively.

Among the new tools that have become available to tackle infectious disease propagation, complex network theory [13,14] has been of considerable interest [5,2], as a way to address the shortcomings of more classic approaches [4] where all individuals in the population of interest are assumed to have an equal probability to infect all other individuals (random-mixing). In contrast to the random-mixing approach, complex networks (heterogeneous mixing) assume that each individual (represented by a node) has a defined set of contacts (represented by links) to other specific individuals (called neighbors), and infections can be propagated only through these contacts. This new technical framework has produced novel insights that are expected to help considerably in the fight against infectious diseases [10,5].

The use of complex network theory requires a few pieces of information in order to be correctly applied. First, it is important to understand the kind of disease being considered, as this will dictate the specifics of the network model that need to be used. For example, the flu virus usually spreads among people that come in contact even briefly, leading to networks with exponential distributions [7] or fat-tailed distributions of connections with large average degree [6]. On the other hand, sexually transmitted diseases are better described by more sparse, and fairly heterogeneous contact networks [4]. Thus, these two examples easily illustrate one of the complications of the problem: the structure of the network to be used. Other aspects involve the life cycle of the pathogen, seasonality, etc. Additionally, social and practical aspects involving public health policy and strategic planning play important roles in the problem.

[*] Corresponding author.
E-mail address: clagorio@mdp.edu.ar (C. Lagorio).

0378-4371/$ – see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.physa.2008.10.045

Regarding the issue of network structure, a few models have been proposed as useful substrates for disease propagation. Among these, truncated scale-free network structures [2] have received considerable interest [9,12]. In these networks, each node has a probability $P(k)$ to have k links (degree k) connecting to it, with $P(k)$ being characterized by the form

$$P(k) = \left[k^{-\lambda} \exp(-k/\kappa)\right] / \left[\mathrm{Li}_{\lambda}(\mathrm{e}^{-1/\kappa})\right], \tag{1}$$

with $k \geq k_{\min}$, where $k_{\min}$ is the lowest degree that a node can have and κ is an arbitrary degree cutoff reflecting the properties of the substrate network for the disease [15]. The reason for including the exponential cutoff is two-fold: first many real-world graphs appear to show this cutoff; second it makes the distribution normalizable for all λ, and not just $\lambda \geq 2$ [16].

Another important issue of propagation relates to the type of disease being considered and its dynamics. In this sense, a general model for a number of diseases (including the ones mentioned at the beginning) is the SIR model, which separates the population into three groups: susceptible, infected and recovered (or removed), approximating well the characteristics of many microparasitic diseases [4]. The solution to the SIR model corresponds to the determination of the number of susceptible, infected, and recovered individuals at a given time. Public health officials are particularly interested in the final outcome of the disease propagation, measured through the number of individuals S_{SIR}, out of a population of N, that became infected at any time. Notice that S_{SIR} should not be confused with the number of susceptible individuals [17]. Another useful way to express the solution of the model is through the average fraction of infected individuals (also known as attack rate) $\langle M_{\mathrm{SIR}}\rangle = \langle S_{\mathrm{SIR}}/N\rangle$, where $\langle\rangle$ denotes averages over realizations.

A number of details related to SIR determine the methods that correctly yield S_{SIR} [9,12]. One common formulation of SIR assumes that on each time step, an infected node has a probability β to infect any of its susceptible neighbors, and once infected the node recovers in exactly t_R time steps. This yields an overall probability T, called the transmissibility, to use any given network link of a node that becomes infected. For this case, when the networks have very simple structure [18], $\langle M_{\mathrm{SIR}}\rangle$ can be determined using a mapping to the link percolation model [3,2] of statistical physics [19] (see below). If the SIR propagation details change, modified forms of percolation may be used [9,12].

From the standpoint of public health policy and strategic planning, an important technical point is how to "define" what is considered to be an epidemic, because such a definition determines the level of reaction that health organizations (e.g., World Health Organization) will apply in dealing with a particular infectious disease event. In real-world disease spread situations, as pointed out in several Refs. [2,9,12], epidemiologists are obliged to define a minimum number of people infected, or threshold s_c to distinguish between a so-called outbreak (a small number of individuals where no large intervention is called for), and an epidemic (a significant number of individuals in the population requiring large scale intervention). In Refs. [2,9,12], for instance, s_c has been used, but its impact on average predictions of SIR has not been systematically addressed, even though it is representative of the sensitivity, or urgency, that epidemiologists assign to the disease in question.

In this paper we address the importance of s_c for SIR in complex networks. Using link percolation, we first concentrate on calculating the average fraction $\langle M_{\mathrm{SIR}}(T, s_c)\rangle$ over SIR model realizations for which $S_{\mathrm{SIR}} \geq s_c$. This quantity is important in the public health community to determine the average expectation value for the epidemic size that can arise given the particular pathogen and society affected, *and the epidemic threshold s_c chosen*. To calculate SIR through link percolation, we find that a reweighting procedure is necessary, that has been previously ignored. Once this reweighting is done, $\langle M_{\mathrm{SIR}}(T, s_c)\rangle$ for large s_c (see below for a detailed discussion) approaches $P_\infty(T) \equiv P_\infty(N, T)$, corresponding to the average fractional size of the largest percolation cluster at T, but for s_c smaller than a value that depends on the topology of the network, we find that $\langle M_{\mathrm{SIR}}(T, s_c)\rangle < P_\infty(T)$, for $T_c < T < 1$, (T_c is the percolation threshold, defined below in detail) indicating that the percolation result for P_∞ is an upper bound. Since the choice of s_c determines what is defined to be an epidemic, we also determine $Q \equiv Q(T, s_c)$, the probability that an SIR realization reaches $S_{\mathrm{SIR}} \geq s_c$. Extending our results to situations such as computer networks, where one should be able to declare an epidemic even if few computers are infected due to the "similarity" of the world population of computers (i.e. sharing the same operating system), and thus have large susceptibility, we find that similar results apply.

The rest of the article is structured as follows. Section 1 introduces details of the network model and where it applies, the link percolation method used to solve the SIR model, and the details of the reweighting procedure necessary to obtain correct averages. Sections 2 and 3 introduce and explain the results of the application of the model to disease propagation events in simulated networks and real-world examples (computer networks). Finally, Section 4 summarizes the results of the paper and presents our conclusions.

1. Models and algorithm

To construct networks of size N we use the Molloy–Reed algorithm or Configurational model [20,21], and apply it to the degree distribution given by Eq. (1). Simulations for this type of network have been performed before in Refs. [2,9] for $N = 10^4$ and 10^5, $\lambda = 2$, $k_{\min} = 1$, $\kappa = 5, 10, 20$ and $s_c = 100$ and 200 [23]. We perform our simulations for many values of κ but we present our results only for $\kappa = 10$. Our main results also hold for other degree distributions. Due to the fact that the lower degree is $k_{\min} = 1$ [24] and κ is small, the network is very fragmented and the size of the initial biggest connected cluster (also know as the giant component abbreviated as GC), labeled here as N_{GC}, is typically 60% of the network (for $\kappa = 10$). In each realization we build a new network and work only on the GC of the original network because we are only concerned with the disease spread on connected communities. Isolated clusters cannot propagate a disease.

To simulate SIR, we chose one node at random on the GC of the substrate network, and infect it. Per time step, this infected node has a probability β to infect its first neighbors. Once a neighbor has been infected, it can infect any of its own susceptible neighbors, but it cannot be infected again nor infect another already infected or recovered node. All infected nodes recover after t_R time steps of becoming infected [25]. The transmissibility T is the overall probability that a node infects one of its susceptible neighbors within the time frame $t = 1$ to t_R, given by $\sum_{t=1}^{t_R} \beta(1-\beta)^{t-1} = 1-(1-\beta)^{t_R}$. For every realization of SIR, the total number of nodes that become infected after the infectious transmission has ended is given by $S_{\rm SIR}$. The values of $S_{\rm SIR}$ satisfy a distribution $\Phi(S_{\rm SIR})$.

As mentioned in the introduction, another way to calculate $S_{\rm SIR}$ is through the use of link percolation. This is a process in which an initial network is modified by removing a fraction $1 - T$ of its links (we use T as the probability for a link to be present because of the mapping between link percolation and our SIR model). The effect of the removal is to generate a multitude of clusters, each being a group of nodes that can be reached from each other by following a sequence of edges connected to those nodes. Link percolation has a threshold value $T = T_c$, characterized by the fact that, for $T < T_c$, the size of the largest cluster typically scales as $\log N$, and for $T > T_c$, a large cluster emerges with a size that scales linearly with N, alongside a number of small clusters. Thus, a so-called percolation transition occurs at $T = T_c$ that takes the network from disconnected to connected. In general terms, a similar situation occurs in SIR, where a high likelihood of transmission of the disease (large T) between neighbors typically leads to a large epidemic, but if this likelihood is low (small T), only small localized outbreaks appear (a detailed description of the relation is developed below).

To perform link percolation, we begin in the GC of the substrate network, and randomly eliminate links with probability $1 - T$. Each realization of this process yields multiple connected clusters of various sizes. Realizations are then repeated multiple times, and a distribution of cluster sizes $\phi(S_p)$ emerges. For the quantity $P_\infty(N, T)$ (which we henceforth abbreviate as $P_\infty(T)$), we average the largest cluster size divided by N_{GC} produced in each realization.

The relation between SIR and link percolation can be concretely explained in the following way: each SIR realization begins with a randomly chosen node of the GC, and the infection propagates to a set of nodes $S_{\rm SIR}$ that can all be traced back to the original infection. The links used in this SIR realization, on average, where used with probability T and not used with probability $1 - T$. To draw the correct connection to link percolation, we first must realize that in a given realization of percolation, only one of the many connected clusters can be chosen to represent the infection of SIR. By analogy with the classic Leath algorithm [26] of cluster creation in percolation, we can conclude that the clusters are randomly picked, with probability proportional to their size S_p. Thus, one expects that the average size of SIR realizations is equivalent to a weighted average of percolation realizations, where the weight is given by S_p.

With the previous arguments in mind, and given the dependence of the problem on both T and s_c, we compute $\langle M_{\rm SIR}(T, s_c)\rangle$ through [27]

$$\langle M_{\rm SIR}(T, s_c)\rangle = \sum_{S_{\rm SIR} \geq s_c} \frac{S_{\rm SIR}}{N_{GC}} \, \Phi(S_{\rm SIR}). \tag{2}$$

In order to compare this to link percolation, we perform a weighted average to obtain $\langle M_p(T, s_c)\rangle$, given by

$$\langle M_p(T, s_c)\rangle = \frac{\sum\limits_{S_p \geq s_c} (S_p^2/N_{GC}) \, \phi(S_p)}{\sum\limits_{S_p \geq s_c} S_p \, \phi(S_p)}. \tag{3}$$

We expect that both averages converge to the same value when enough realizations are performed. Additionally, as s_c is increased, we expect $\langle M_p(T, s_c \gg 1)\rangle \to P_\infty(T)$ for $T > T_c$, because a progressively smaller number of small clusters enters into the averaging, and only the largest clusters are used. This creates an interesting scenario, in which $P_\infty(T)$ is a good approximation of the epidemic size only in the limit of a large threshold $s_c \geq S_p^\times$ (a function of T only, defined below). However, for smaller s_c, which is important in more aggressive diseases, only $\langle M_p(T, s_c)\rangle$ is the correct average, which includes both small and large SIR events.

2. Results on the relative average size of the disease

2.1. Mapping between the average fraction size using SIR simulations and the average fraction size of all percolation cluster

As a first step, we illustrate the equality of $\langle M_{\rm SIR}(T, s_c)\rangle$ and $\langle M_p(T, s_c)\rangle$ [9,12] (Fig. 1) by plotting $\langle M_{\rm SIR}(T, s_c)\rangle$ and $\langle M_p(T, s_c)\rangle$. The two curves overlap indicating that the mapping between the two quantities is correct. In the remainder (unless explicitly stated), we perform our simulations using link percolation as opposed to SIR.

The mapping between the steady state of SIR and link percolation is computationally very convenient for several reasons. First, performing simulations of SIR models is computationally more costlier than link percolation. This is due to the fact that for SIR, only a single propagation occurs per realization, as opposed to multiple clusters that appear for link percolation.

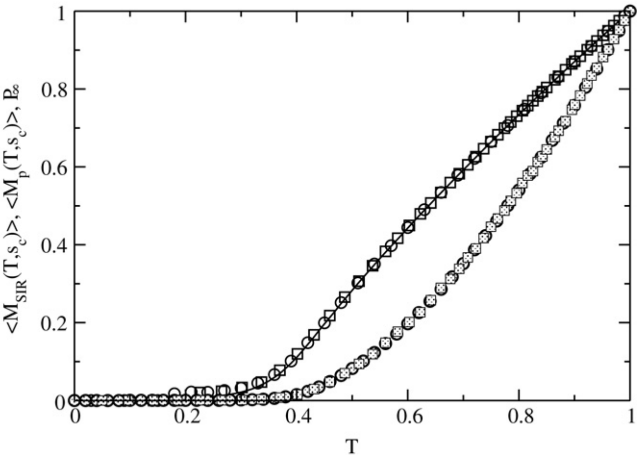

Fig. 1. Comparison between $\langle M_{\rm SIR}(T, s_c)\rangle$ (□), $\langle M_p(T, s_c)\rangle$ (◯), and $P_\infty(T)$ of link percolation (full line). Empty symbols correspond to $s_c = 100$, and dotted symbols to $s_c = 1$. For the transmissibility in the SIR problem, we used $\beta = 0.05$ and a set of values of the recovery t_R to cover a wide range of T. All the simulations were performed on the GC of networks with $\lambda = 2$, $\kappa = 10$, $k_{\min} = 1$, and averaged over 10^4 realizations.

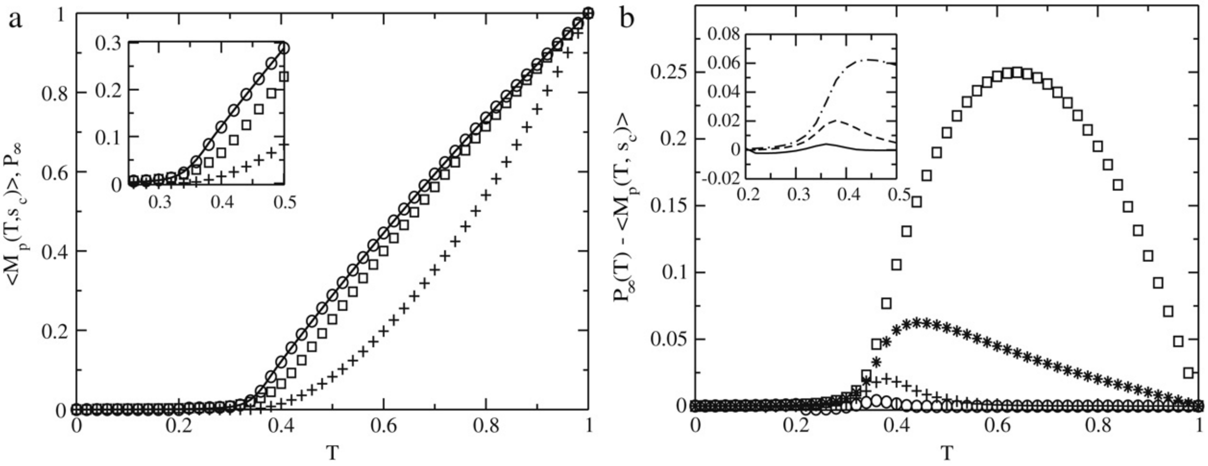

Fig. 2. (a) Plot of $\langle M_p(T, s_c)\rangle$ as a function of T, for $s_c = 200$ (◯), $s_c = 10$ (□) and $s_c = 1$ (+). The full line represents $P_\infty(T)$. The inset shows the details of the main plot close to $T_c \approx 0.32$, i.e, for T near the percolation threshold. We can observe that the departure between $P_\infty(T)$ and $\langle M_{\rm SIR}(T, s_c)\rangle$ is not negligible. (b) $P_\infty - \langle M_p(T, s_c)\rangle$ as a function of T, for $s_c = 1$ (□), $s_c = 10$ (∗), $s_c = 50$ (+) and $s_c = 200$(◯). In the inset we plot the details of the main plot around T_c for $s_c = 10$ (dot dashed line), $s_c = 50$ (dashed line) and $s_c = 200$ (full line). We observe that $P_\infty(T)$ is an upper bound for $\langle M_p(T, s_c)\rangle$ [29]. In all the simulations we used $N = 10^5$, $\lambda = 2$, $\kappa = 10$, $k_{\min} = 1$ and the averages where done over 10^3 realizations on the GC of networks of size $\simeq 0.6N$.

Additionally, SIR propagation has to be performed in a dynamic fashion, which makes it necessary to test over time a given propagation condition, something that does not occur for link percolation, accelerating further the simulations. Finally, this mapping is convenient because it gives another conceptual framework in which to understand the relation between these two problems of disease propagation and percolation models.

A final feature of Fig. 1 is the plot of $P_\infty(T)$. This curve displays good agreement with $\langle M_{\rm SIR}(T, s_c)\rangle$ for the larger s_c. We discuss this issue further in Section 2.2.

2.2. *Effects of s_c on SIR measures*

In Fig. 2(a), we plot $\langle M_p(T, s_c)\rangle$ to explore the effect of s_c on this average. We can see from the plot that only for larger s_c (for our simulation parameters $\approx$ 200) the curves of $P_\infty(T)$ and $\langle M_p(T, s_c)\rangle$ coincide for $1 \geq T > T_c$ ($T_c \approx 0.34$ for $N = 10^5$), while for smaller s_c values they do not. The need to use large s_c to approach $P_\infty(T)$ had been realized previously (for instance Refs. [28,9]), but not commented on in any detail. We can see this behavior more clearly in Fig. 2(b), where we plot $P_\infty(T) - \langle M_p(T, s_c)\rangle$ for different values of s_c and find that $P_\infty(T)$ is an upper bound of $\langle M_p(T, s_c)\rangle$, except for large s_c (see Ref. [29], and recall that by definition $\langle M_p(T, s_c)\rangle \geq s_c$, thus care must be taken not to induce pathological situations by choosing s_c larger than expected large SIR events).

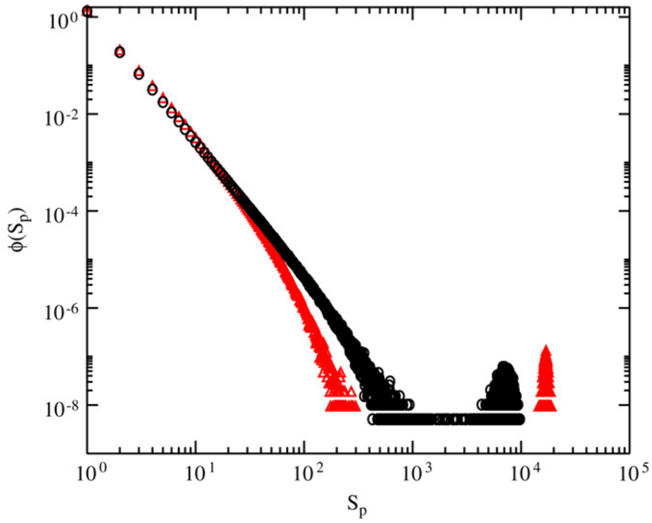

Fig. 3. Distribution of cluster sizes $\phi(S_p)$ for $T = 0.4$ (◯) and 0.5 (△, red color online). The distribution has two regions in which there is significant statistical weight. The first of the two corresponds to the presence of finite clusters, and the second to large spanning clusters. As T increases the second region moves to the right, concentrated around $S_p^\infty(T)$, and the first region becomes smaller due to the cutoff $S_p^\times(T)$ moving to the left. This also signals the decay of statistical weight of the first region and increase of the second. For $T = 1$, all the weight is concentrated on the second region. In all the simulations we used $N = 10^5$, $\lambda = 2$, $\kappa = 10$, $k_{\min} = 1$ and the averages where done over 10^4 realizations on the GC of networks of size $\simeq 0.6N$..

In order to understand these results systematically, we plot the distribution $\phi(S_p)$ for two values of T (Fig. 3). From percolation theory it is known that, for T close and above T_c, $\phi(S_p) \sim AS_p^{-\tau}\exp(-S_p/S_p^\times) + F(S_p - S_p^\infty)$, where τ has the mean field value 5/2. In the last expression, $S_p^\times$ is a characteristic maximum finite cluster size which scales as $|T - T_c|^{-\sigma}$ ($\sigma = 1/2$), A is a measure of the relative statistical weight between the two terms, F is a narrow function of its argument, and $S_p^\infty = S_p^\infty(T) \equiv \langle N_{GC}\rangle P_\infty(T)$. The value of A can be estimated from the fact that, for a system size $\langle N_{GC}\rangle$, the first term of $\phi(S_p)$ accounts for the finite clusters present, and the integral of $S_p\phi(S_p)$ must be equal to the mass of the finite clusters. Therefore

$$[\langle N_{GC}\rangle - S_p^\infty(T)] \sim A\int_1^{\langle N_{GC}\rangle} S_p^{-\tau+1}\exp(-S_p/S_p^\times)\mathrm{d}S_p$$
$$\Rightarrow A \sim \frac{(\tau-2)(\langle N_{GC}\rangle - S_p^\infty(T))}{1-(S_p^\times)^{-\tau+2}}. \tag{4}$$

Since the rest of the mass of the network is contained in a single spanning cluster, then the relative statistical weight of the first to second term of $\phi(S_p)$ is $A : 1$, justifying the choice of the integral of F to be 1. The overall normalization can be obtained from the fact that $\int_1^{\langle N_{GC}\rangle} S_p\Phi(S_p)\mathrm{d}S_p = \langle N_{GC}\rangle$. The effects shown here hold also for other networks including real networks as shown below.

In general, since $\phi(S_p) = \phi(S_p, T)$, any choice of s_c affects the value of $\langle M_p(s_c, T)\rangle$ differently for different T. The choice $s_c = S_p^\times(T = T_c)$ is generally convenient for any $T \geq T_c$ (although not perfect, as explained below) if the goal is to have $\langle M_p(s_c, T)\rangle \to P_\infty(T)$, which reflects an averaging only over large SIR events. In case the disease in question has T considerably larger than T_c, $\phi(S_p)$ is virtually bimodal, with a region of extremely low probability between $S_p^\times(T)$ and $S_p^\infty(T)$ inside of which changing s_c has virtually no consequences. If $s_c < S_p^\times(T)$, $\langle M_p(T, s_c)\rangle$ becomes the average of this bimodal, but for $S_p^\times(T) < s_c < S_p^\infty(T)$, $\langle M_p(s_c, T)\rangle$ is dominated by the second part of the distribution producing a value that reflects typical large SIR events only. Since the average of a bimodal lacks descriptive power, this analysis suggests that for large T, $S_p^\times(T) < s_c < S_p^\infty(T)$ is a good choice. On the other hand, in the case of $T \gtrsim T_c$, $\phi(S_p)$ is a truncated power-law and changes in s_c induce changes in $\langle M_p(T, s_c)\rangle$ continuously, thus making the choice of s_c less obvious. If the concern regarding a particular disease is to activate epidemiological interventions quickly, a small value of s_c should be chosen related to practical considerations such as readiness of the public health sector; if, however, the guiding principle is to analyze the statistical features of large events, s_c close to, or slightly larger than, $S_p^\times(T)$ ($s_c \gtrsim S_p^\times(T)$) guarantees averaging only over those. It is important to keep in mind that the statistical weight of events of size $S_p^\infty(T)$ becomes negligible as $T \to T_c$, making $s_c \gtrsim S_p^\times(T)$ a choice that forces $\langle M_p(T, s_c)\rangle$ to become dominated by vanishingly improbable events. Finally, choosing $s_c \gtrsim S_p^\infty(T)$ is, at best, an unsafe choice for transmissibility T or below because it forces $\langle M_p(T, s_c)\rangle \geq S_p^\infty(T)$, which is meaningless (see, for instance inset of Fig. 2(b), where $\langle M_p(T, s_c)\rangle > P_\infty(T)$ for $T_c \gtrsim T$). In essence, our analysis suggests that s_c close to and above $S_p^\times(T)$ is generally a good choice, unless special considerations are present due to a particularly dangerous disease which, in addition, satisfies $T \gtrsim T_c$.

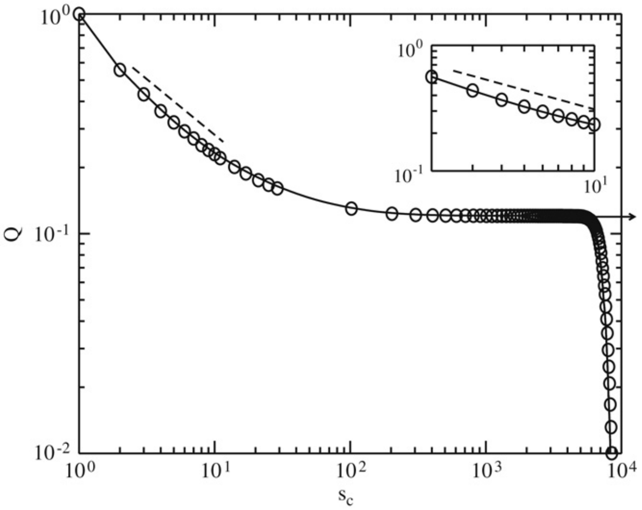

Fig. 4. Plot of Q for: SIR as a measure of the number of times an $S_{\mathrm{SIR}} \geq s_c$ divided by the number of realizations (full line). Link percolation over all clusters as in Eq. (6) (◯). We observe that both the curves are in good agreement. For small s_c, Q has a power-law decaying behavior with exponent $\tau - 2 = 1/2$. The arrow represents approximately $S_p^{\infty}/\langle N_{GC}\rangle \approx 0.12$ as predicted by the theoretical scaling.

The choice of s_c has an extra consequence, which is to change the likelihood that a given pathogen propagation be declared as an epidemic. This probability is relevant from the standpoint of readiness, because lower s_c implies that it is more likely to consider almost any disease propagation as reaching the epidemic state. Thus, we define Q which represents the probability that an SIR with transmissibility T has size $S_{\mathrm{SIR}} \geq s_c$. This quantity can be computed directly as the number of times $S_{\mathrm{SIR}} \geq s_c$ divided by the total number of realizations (see Fig. 4). Analytically, Q can be related to $\Phi(S_{\mathrm{SIR}})$ through

$$Q = \frac{\sum\limits_{S_{\mathrm{SIR}} \geq s_c} \Phi(S_{\mathrm{SIR}})}{\sum\limits_{S_{\mathrm{SIR}} \geq 1} \Phi(S_{\mathrm{SIR}})} = \sum_{S_{\mathrm{SIR}} \geq s_c} \Phi(S_{\mathrm{SIR}}) \tag{5}$$

where the last equality is a consequence of normalization. In order to calculate Q from the percolation results, we keep in mind the reweighting applied to Eq. (3). Then, Q is given by

$$Q = \frac{\sum\limits_{S_p \geq s_c} S_p\, \phi(S_p)}{\sum\limits_{S_p \geq 1} S_p\, \phi(S_p)} \tag{6}$$

where $\sum_{S_p \geq 1} S_p \phi(S_p) = \langle N_{GC}\rangle$. In Fig. 4, we plot Q for SIR for $T = 0.4$, ($T \gtrsim T_c$), using direct computation and compare it with the results obtained using Eq. (6). We can see that the agreement is excellent.

To calculate Q, we use $\phi(S_p)$ and Eq. (5), and assume the continuum limit over S_p, giving

$$\begin{aligned} Q &\sim \int_{s_c}^{\langle N_{GC}\rangle} \frac{S_p \phi(S_p)}{\langle N_{GC}\rangle} \mathrm{d}S_p \\ &\sim \int_{s_c}^{\langle N_{GC}\rangle} \frac{[A S_p^{-\tau+1} \exp(-S_p/S_p^{\times}) + S_p F(S_p - S_p^{\infty})]}{\langle N_{GC}\rangle} \mathrm{d}S_p \\ &\sim \begin{cases} A\dfrac{s_c^{-\tau+2} - (S_p^{\times})^{-\tau+2}}{\langle N_{GC}\rangle(\tau-2)} + \dfrac{S_p^{\infty}}{\langle N_{GC}\rangle} & [s_c \leq S_p^{\times}] \\ \dfrac{S_p^{\infty}}{\langle N_{GC}\rangle} & [S_p^{\times} \ll s_c \leq S_p^{\infty}] \\ 0 & [S_p^{\infty} < s_c], \end{cases} \end{aligned} \tag{7}$$

where we approximated the first term of the integral by truncating the integration at $S_p^{\times}(T)$, and simplifying F to a delta function (of integral 1, which relates to the value of A). Several Q regimes can be identified: (i) for $s_c \ll S_p^{\times}$, the contribution of $(S_p^{\times})^{-\tau+2}$ is negligible and therefore $Q \sim s_c^{-\tau+2}$; (ii) for $s_c \sim S_p^{\times}$, Q becomes dominated by a competition between the two terms of the integral and no clear scaling rules apply; (iii) for $S_p^{\times} \ll s_c < S_p^{\infty}$, $Q \sim S_p^{\infty}$, and; (iv) for $s_c > S_p^{\infty}$, $Q \to 0$. From Fig. 4 we can identify those four regimes. In the figure the arrow represents approximately $S_p^{\infty}/\langle N_{GC}\rangle \approx 0.12$ from the simulation. The agreement between the theoretical scaling (see Eq. (7)) and the simulation is excellent.

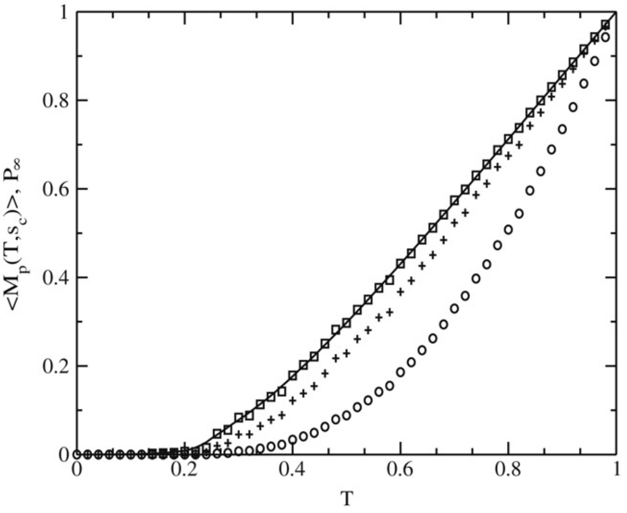

Fig. 5. Plot of $\langle M_p(T, s_c)\rangle$ as a function of T, for the Tracerouter network that has $N = 222\,934$, Links $= 279\,510$, $P(k) \sim k^{-\lambda}$ with $\lambda = 2.1$, with $s_c = 1$ ($\circ$), $s_c = 2$ ($+$) and $s_c = 100$ ($\square$). The full line represents $P_\infty(T)$.

3. Application to Tracerouter and DIMES networks

The results we have presented for our model of human infectious disease propagation is applicable to other problems in the real world. This can be well illustrated for computer networks in which information is being broadcasted.

One of the networks that describes the functional connectivity of the Internet is the Tracerouter network, where the nodes are the routers and the links are the connections between them that transport IP packets. The network, as measured in Ref. [30], has $N = 222\,934$ nodes and $L = 279\,510$ links. This network can be represented by a Scale-Free network with $\lambda = 2.1$ [30]. In order to obtain information of the Internet connectivity, a software probe is used called a Tracerouter tool, that sends IP packets on the Internet eliciting a reply from the targeted host. By citing the information of the packets' path to the various destinations, a network of router adjacencies is built [31]. Here, the SIR process can be understood as a router that has a random failure (Infected), that can produce failures on neighbor nodes that are functional (Susceptible), and these new nodes become infected. Thus, after certain time of router failure the protocol disconnects the router from the network (Removed). The DIMES network [32] uses the same algorithm of searching as the Tracerouter network, the nodes are Autonomous Systems (AS) and the links are the connections between AS. The network has $N = 20\,556$ nodes and $L = 62\,920$ links. The description of the SIR process over DIMES is the same as the one explained before for the Tracerouter network.

In Figs. 5 and 6 we plot $P_\infty(T)$ and $\langle M_p(T, s_c)\rangle$ for different values of s_c as a function of T. For $s_c = 500$, for Tracerouter and $s_c = 100$ for DIMES network we can map this problem to $P_\infty(T)$ of link percolation. We can see that the problem maps into $\langle M_p(T, s_c)\rangle$ for any size of s_c. We compute Q for both networks, those results are plotted in Fig. 7(a) and (b) for Tracerouter and DIMES networks, respectively. For DIMES, $T_c \to 0$, and thus first region cannot be seen [19]. On the other hand, if T_c is finite as in Tracerouter, Q has the four regions described for model networks (see Eq. (7)).

4. Summary

We have shown that the choice of s_c, the minimum SIR propagation size necessary to declare an epidemic, has important consequences on epidemiological predictions. Using percolation theory to calculate the average fractional size $\langle M_{\mathrm{SIR}}(T, s_c)\rangle = \langle M_p(T, s_c)\rangle$ of an epidemic, we find that the strength of the spanning link percolation cluster $P_\infty(T)$ is an upper bound to $\langle M_{\mathrm{SIR}}(T, s_c)\rangle$, provided s_c does not exceed $S_p^\infty(T)$, the typical size of finite clusters of link percolation, where pathological results can appear. When s_c is between $S_p^\times(T)$ and $S_p^\infty(T)$, $P_\infty(T)$ is a good approximation to $\langle M_{\mathrm{SIR}}(T, s_c)\rangle$. For small values of s_c, P_∞ is no longer a good approximation, and the average fractional size has to be computed directly. Our analysis suggests that for a given disease (of known T) and social network, $s_c \gtrsim S_p^\times(T)$ is generally a good choice, unless $T \gtrsim T_c$ and the disease requires special considerations by authorities. When, the goal is to have $\langle M_p(s_c, T)\rangle \to P^\infty(T)$, which reflects an averaging only over large SIR events, a convenient choice of s_c is $s_c = S_p^\times(T = T_c)$. We also study Q, the probability that an SIR propagation reaches the epidemic mass s_c, which has several interesting regimes including one that scales as $s_c^{-\tau+2}$. We apply our results to real networks (DIMES and Tracerouter) to measure the consequences of the choice s_c on predictions of average outcome sizes of computer failure epidemics.

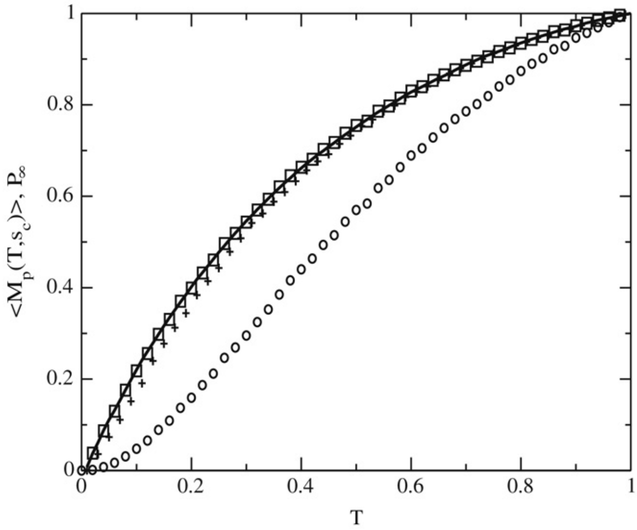

Fig. 6. Plot of $\langle M_p(T, s_c)\rangle$ as a function of T, for the DIMES network that has Scale-Free distribution with $\lambda \approx 2.15$, $N = 20\,556$, links $= 62\,920$, for $s_c = 1$ (○), $s_c = 10$ (+) and $s_c = 500$ (□). The full line represents $P_\infty(T)$.

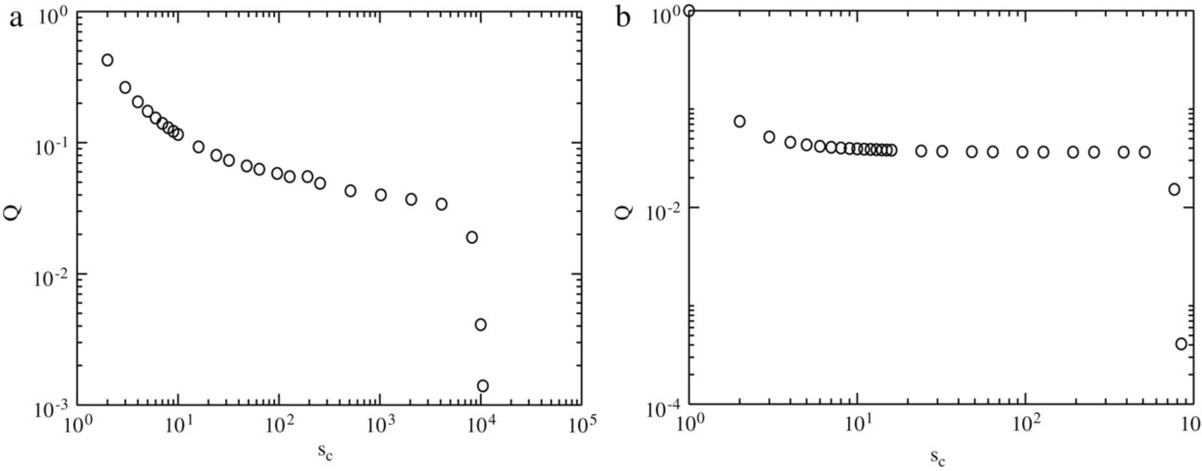

Fig. 7. Q as a function of s_c for: (a) Tracerouter network, with $T = 0.25$ (◯). (b) DIMES network, with $T = 0.02$ (◯), the exponent of the decreasing power-law is around 0.62, indicating that for this network $\tau \sim 2.62$.

Acknowledgments

E.L. acknowledges financial support from DOE (US) and EPSRC (UK). C.L., M.V.M., P.A.M. and L.A.B acknowledge financial support from PICTO-3370 (ANPCyT) and U.N.M.d.P. We also acknowledge M.E.J. Newman and E. Kenah for fruitful discussions.

References

[1] R. Pastor-Satorras, A. Vespignani, Phys. Rev. Lett. 86 (2001) 3200.
[2] M.E.J. Newman, Phys. Rev. E 66 (2002) 016128.
[3] P. Grassberger, Math. Biosci. (1983) 157–172.
[4] R.M. Anderson, R.M. May, Infectious Disease in Humans, Oxford University Press, Oxford, 1992.
[5] V. Colizza, A. Barrat, M. Barthélemy, A. Vespignani, Proc. Natl. Acad. Sci. USA 103 (2006) 2015.
[6] S. Eubank, H. Guclu, A. Kumar, M.V. Marathe, A. Srinivasan, Z. Toroczkai, N. Wang, Nature 429 (2004) 180–184.
[7] S. Bansal, B.T. Grenfell, L.A. Meyers, J. R. Soc. Interface 4 (2007) 879–891.
[8] L.M. Sander, C.P. Warren, I.M. Sokolov, C. Simon, J. Koopman, Math. Biosci. 180 (2002) 293–305.
[9] E. Kenah, J.M. Robins, Phys. Rev. E 76 (2007) 036113.
[10] R. Cohen, S. Havlin, D. ben-Avraham, Phys. Rev. Lett. 91 (2003) 247901.
[11] E. López, R. Parshani, R. Cohen, S. Carmi, S. Havlin, Phys. Rev. Lett. 99 (2007) 188701.
[12] J.C. Miller, Phys. Rev. E 76 (2007) 010101.
[13] A.L. Barabási, Rev. Modern Phys. 286 (1999) 509.

[14] R. Albert, A.-L. Barabási, Rev. Modern Phys. 74 (2002) 47.
[15] The function Li_λ is the polylogarithm function of argument λ, which emerges in this context as a consequence of the normalization condition of the probability distribution $P(k)$.
[16] M.E.J. Newman, S.H. Strogatz, D.J. Watts, Phys. Rev. E 64 (2001) 026118.
[17] We are using the notation that is usual in percolation.
[18] This approach is valid when, statistically, the number of links of a given node has no correlations to the number of links of its neighboring nodes.
[19] D. Stauffer, Introduction to Percolation Theory, Taylor & Francis, 1985.
[20] In this algorithm [22], each node is first assigned a random number of "stubs" taken from $P(k)$. Next, we connect two unused stubs from two randomly selected nodes. The only condition that we impose is that there cannot be multiple edges between two nodes.
[21] B. Bollobas, European J. Combin. 1 (1980) 311–316.
[22] M. Molloy, B. Reed, Random Structures and Algorithms 6 (1995) 161; Combin. Probab. Comput. 7 (1998) 295.
[23] M. Newman, private communication.
[24] R. Cohen, S. Havlin, D. ben-Avraham, Structural properties of scale free networks, in: S. Bornholdt, H.G. Schuster (Eds.), Handbook of Graphs and Networks, Wiley-VCH, 2002, (Chapter 4).
[25] In this article, since overall SIR dynamics are not considered, t refers to the "local time", i.e., the time of every individual node. Thus, $t = 0$ refers to the time of infection of a given node. More formally, one could refer to t_i for the time running for node i, but by context it is clear that this is not necessary.
[26] P.L. Leath, Phys. Rev B 14 (1976) 5046.
[27] Note that Eqs. (2) and (3) involve also the largest cluster. For model networks of finite size (see Eq. (1)) and for real networks, the size of the largest cluster does not diverge.
[28] E. Kenah, private communication.
[29] When s_c is very large, and given that $\langle M_p(T, s_c)\rangle$ only counts $S_p \geq s_c$, it is possible to have $P_\infty < \langle M_p(T, s_c)\rangle$, because averaging of the mass is only taking place in those rare realizations when the mass condition is satisfied. In this case, one can see a violation of $P_\infty(T)$ being a bound for $\langle M_p(T, s_c)\rangle$, but this behavior is pathological.
[30] M. Kitsak, S. Havlin, G. Paul, M. Riccaboni, F. Pammolli, H.E. Stanley, Phys. Rev. Lett. 75 (2007) 056115.
[31] R. Pastor-Satorras, A. Vespignani, Evolution and Structure of the Internet, Cambridge University Press, 2004.
[32] Yuval Shavitt, Eran Shir, DIMES—Letting the Internet Measure Itself, http://www.arxiv.org/abs/cs.NI/0506099.