Methodol Comput Appl Probab
DOI 10.1007/s11009-013-9325-z

Applications of the Variance of Final Outbreak Size for Disease Spreading in Networks

Lilia L. Ramírez-Ramírez · Mary E. Thompson

Received: 19 June 2012 / Revised: 7 February 2013 / Accepted: 15 February 2013

Abstract The assumption that all susceptible individuals are equally likely to acquire the disease during an outbreak (by direct contact with an infective individual) can be relaxed by bringing into the disease spread model a contact structure between individuals in the population. The structure is a random network or random graph that describes the kind of contacts that can result in transmission. In this paper we use an approach similar to the approaches of Andersson (Ann Appl Probab 8(4):1331–1349, 1998) and Newman (Phys Rev E 66:16128, 2002) to study not only the expected values of final sizes of small outbreaks, but also their variability. Using these first two moments, a probability interval for the outbreak size is suggested based on Chebyshev's inequality. We examine its utility in results from simulated small outbreaks evolving in simulated random networks. We also revisit and modify two related results from Newman (Phys Rev E 66:16128, 2002) to take into account the important fact that the infectious period of an infected individual is the same from the perspective of all the individual's contacts. The theory developed in this area can be extended to describe other "infectious" processes such as the spread of rumors, ideas, information, and habits.

Keywords Epidemic · Random graph · Social network · Outbreak size

AMS 2000 Subject Classifications 60J85 · 65C05 · 92D30 · 05C80

L. L. Ramírez-Ramírez (✉)
Actuaría y Seguros, ITAM, Rio Hondo 1, Col. Tizapán San Angel, Del. Alvaro Obregón, C.P. 01080, Mexico City, Mexico
e-mail: lilialeticia.ramirez@itam.mx

M. E. Thompson
Statistics and Actuarial Science, University of Waterloo, 200 University Avenue West, Waterloo, ON N2L 3G1, Canada
e-mail: methomps@uwaterloo.ca

Published online: 08 March 2013

1 Introduction

Some models that have been proposed to relax the law of mass action for the spread of disease consider the division of the population into subpopulations by their level of mixing (interaction) (Brauer and Watmough 2009; Ball and Neal 2002; Ball et al. 1997; Britton et al. 2011) or are based on a population network of contacts.

Some important early treatments of network based epidemic models are Anderson and May (1991), Molloy and Reed (1995), Andersson (1998) and Newman et al. (2001). The networks around which the models are constructed are those in which the vertices represent individuals (or units) and the edges represent the contacts that can lead to the transmission of the infectious agent.

These network based epidemic models can also relax the assumption that the infectious period has an exponential distribution (Anderson and May 1991; Newman et al. 2001), and have provided expressions for the expectation of final outbreak size for general infectious period distributions. However, knowledge of the first two moments of outbreak size is a minimum requirement for portraying the range of most likely scenarios (with the use of probability intervals) for the evolution of outbreaks. Since social and "real-world" networks are usually very heterogeneous, one focus of this work is to study the impact of heterogeneity induced by a non-Poisson degree distribution on the variability around the mean of the final outbreak size.

The paper is organized as follows: Section 2 introduces the basic concepts associated with random graphs and real-world networks. In Section 3 we present some results on the mean and variability of final outbreak size, assuming that the contacts are described by a random network and the rates and infectious periods are independent random variables. Based on these derived expressions, we study the dependence of final size variation on network heterogeneity. The results apply also to the mean and variance of outbreak size of small outbreaks that can occur even when epidemics are possible. We use the results to compute probability bounds on outbreak sizes under various scenarios.

The expressions for the mean and variance of outbreak size depend on approximations to the outbreak process, whereas simulations of the contact networks and the spread of disease can yield more accurate estimates of the distribution of outbreak size in principle. At the same time, simulation of the contact network and outbreak processes is challenging for large populations and general degree distributions. We find that the results of simulations and our moment calculations agree well in moderate size populations where the approximations are reasonable, and that the moment calculations can be useful guides when simulations are too resource-consuming or unstable.

An interesting result of Newman et al. (2001) has to do with the differences in the means of the degree distributions of infected and uninfected individuals during outbreaks that evolve into epidemics. We extend this result to the case of regarding the infectious period as a random variable that it is unique to each infectious individual, and provide a new proof.

Finally, we generalize previous results to the case where the distributions of the infectious rates and periods can depend on the individual (through covariates), independently of the individual's degree.

The Appendix presents the algorithm to simulate random networks used in Section 3.

2 Background

In the disease spread model that we study, we first set up a structure for the contacts among individuals or other units that can lead to infection. The term "contact" is defined according to the transmission mode for the specific agent under study, and the contact is declared an "infective contact" if it results in the transmission of the infectious agent.

The contact structure is formally described as a random graph (we use the terms "random graph" and "random network" interchangeably) where the vertices represent the individuals and the edges those contacts that the infective agent may use to transmit the infection to other individuals; not all edges between infected and susceptible vertices will result in transmission.

As in Bailey (1975) we consider that infected individuals make transitions through different infectious stages or compartments (SIR or SEIR models), and that infectious contact or transmission may occur only between susceptible (S) and infective (I) individuals that are in contact.

An outbreak occurs when an individual ("patient zero") becomes infected and may transmit the disease to others, who in turn may spread the disease. An outbreak which grows indefinitely or to a very large size is called an "epidemic".

The next subsection presents some of the most important terminology of random graphs applied in the context of the spread of disease.

2.1 Networks

2.1.1 Random Graphs

A *graph* G consists of an ordered pair (V, E) of vertices (points or vertices) $V = \{v_1, \ldots, v_n\}$ and edges E (lines or links) that connect pairs of vertices, so that $E \subset V^2$. The number n is called the *order* of G and the *size* of G refers to the number q of edges.

The graph G is *undirected* if its edges are undirected, and is called *simple* if no more than one edge can connect two different vertices and there are no *self-loops*.

The *degree of a vertex* is the number of edges of which it is an endpoint. A *random graph* is the name given to a graph in which the network vertices are randomly connected by edges. The degree of a vertex is then a random variable.

In one kind of random graphs proposed by Erdős and Rényi (denoted as $G_{n,p}$), each possible edge between two vertices is present independently with probability p, and absent with probability $1 - p$. In the random graphs $G_{n,p}$ each vertex has degree distribution which is binomial with mean $(n-1)p \approx np$. If p is small the degree distribution can be taken to be Poisson. For large n, in an asymptotic sense, the majority of vertices have approximately the same degree, close to the average degree $(n-1)p$. More results concerning the properties of random graphs $G_{n,p}$ can be found in Ivčhenko (1973) and Bollobás (1985), for example.

Some other authors have studied real-world networks, such as networks of citations in the academic literature (Lotka 1926; Gilbert 1997), the World Wide Web (Albert et al. 1999; Faloutsos et al. 1999), and sexual contacts (Liljeros et al. 2001), among others (Watts and Strogatz 1998; Amaral et al. 2000), and showed that the degree sequences of such networks typically approximate a non-Poisson

distribution (See Section 2.1.2). Motivated by the non-Poisson degree distribution graphs present in some real-world networks, various researchers have studied the viability of defining random graphs with any degree sequence (Bender and Canfield 1978; Łuczak 1992; Molloy and Reed 1995) and characterizing some properties of a network by its degree distribution (Molloy and Reed 1998; Albert et al. 2000; Callaway et al. 2000; Cohen et al. 2000; Newman et al. 2001; Pastor-Satorras and Vespignani 2001).

In this paper, the contact network in the disease spread model is described as a simple, undirected and static random graph with general degree distribution. Conditional on the realization of the random graph, the spread of disease proceeds on the contact network according to the stochastic mechanism described in Section 3. (In the alternative framework of Andersson (1998) the set of infected vertices evolves deterministically given a random graph of infectious contacts.)

An important feature of a graph is what is called a *component*. A component in a graph is a subset of vertices each of which is reachable from all others by paths through the network. It is natural to think that if there are few edges in the graph, most of the vertices are disconnected from one another, and the components have small sizes. Erdős and Rényi (1960) (see also Molloy and Reed 1995) provided a threshold result for graphs $G_{n,m}$, where the n vertices are joined by m edges which are placed between pairs of vertices chosen uniformly at random: as the number of vertices n increases to infinity, if $m = cn + o(n)$ for $c < 1/2$, then a.s. the graph has no component with size greater than $O(\log n)$, and no component has more than one cycle. If $m > cn$ for $c > 1/2$ there are constants ϵ, $\delta > 0$ depending on c such that a.s. as n tends to infinity $G_{n,m}$ has a component of at least ϵn vertices and at least δn cycles, known as a *giant component*, and no other component has more than $O(\log n)$ vertices or more than one cycle.

From the definitions of a component and an epidemic, it is clear that epidemics can occur only in graphs where giant components are present.

Molloy and Reed (1995) generalized Erdős and Rényi's findings for a random graph with general degree sequence having a limiting distribution. For the mean degree above a certain threshold, in general there will be only one giant component, and other components will be small and with total size fraction $S_0 \to 0$ as $n \to \infty$.

The threshold results of Molloy and Reed are understood also to apply to random graphs with general degree distribution in the sense used by Newman (2002), where an independent and identically distributed (i.i.d.) sequence $K_1, K_2, \ldots$ of degrees is generated from the distribution, and the graph is randomly selected from among all graphs with a given set of n vertices having degrees given by the set $\{K_1, \ldots, K_n\}$; and similarly, they apply to random graphs with general degree distribution in the sense used by Durrett (2007), in which the degree sequence $K_1, K_2, \ldots$ is assigned randomly to vertices, each vertex is given a number of half-edges equal to its degree, and half-edges are randomly paired.

In Molloy and Reed (1995), given a degree sequence, the authors regard the graphs as generated in an iterative process in which two vertices are connected in each step by randomly selecting the edge that connects them from the set of all possible edges. The selection of an edge in each step is weighted with respect to the product of the respective residual degrees that still have to be allocated for each of its endpoints. The algorithm we use in this paper to simulate the networks is based on the algorithm of Molloy and Reed, and presented in the Appendix.

2.1.2 Non-Poisson Degree Distributions

For a large number of networks the degree distribution is well described with a discrete *power law distribution* (also known as a *zeta distribution*):

$$\Pr(K = k) \propto k^{-\delta}; \qquad \delta > 1, \quad k \in \{1, 2, \ldots\}.$$

Barabási and Albert (1999) called the networks with power-law degree distribution *scale-free* (SF) networks since their degree kernel distribution function $\{p_k = \Pr(K = k)\}$ remains unchanged when scaling k with any constant a. That is, $p_{ak} \propto a^{-\delta} k^{-\delta} \propto p_k$.

The identification of networks with power-law degree distributions has given impetus to the development of new studies of the dynamics that are naturally associated to network structures. Barabási and Albert (1999) pointed out that this kind of network can potentially model generic properties of many real-world networks, and they proposed that the properties of these networks could be explained, as the final product of a model in which a network grows dynamically.

There are major topological differences between the random graphs $G_{n,p}$ and SF networks. For the former, most vertices have approximately the same number of links $E(K)$ since the decay of the distribution's tail guarantees the absence of vertices with appreciably more links than $E(K)$. In contrast, the power-law distribution implies the existence of numerous vertices with only a few links, and a few vertices with a very large number of links. Thus SF networks are extremely heterogeneous.

The power law distribution has finite r-th moment only if its parameter δ is larger than $r + 1$. Since in later results we require a distribution with finite first moments, we illustrate some outbreak characteristics on networks using the polylogarithmic (δ, γ) distribution

$$\Pr(K = k) = \frac{k^{-\delta} \exp(-k/\gamma)}{\mathrm{Li}_\delta(\exp(-1/\gamma))}; \qquad \delta > 1, \gamma > 0, \quad k \in \{1, 2, \ldots\},$$

where $\mathrm{Li}_\delta(t) = \sum_{i=1}^{\infty} t^i / i^\delta$.

The polylogarithmic distribution tends to a power law distribution as $\gamma \to \infty$, and like the power law distribution, it can generate very heterogeneous networks. However all its moments are finite for any value of δ and γ.

3 Outbreaks in Networks

Since some infections require very specific contact between individuals to propagate, their outbreaks are heavily affected by the population connectivity patterns that characterize the type of contact that can result in infection transmission (Bailey 1975; Anderson and May 1991).

The disease models we consider map this contact pattern in terms of random graphs or random networks. The vertices in the graph represent individuals or units (such as hospital wards, institutions and cities) susceptible to becoming infected and transmitting the illness. The links between them represent the kind of contacts that can lead to a transmission of the infection between two individuals or units.

This section studies some characteristics of the final results for an outbreak happening in a population with a simple random graph contact structure among

individuals. Key characteristics include the probability of an outbreak to develop into a large outbreak (epidemic), and the mean and variance for the outbreak size, which is the size of the total affected population.

Andersson (1998) proved that when the number of initial patients is negligible, and the degree distribution has finite $(4+\epsilon)$ moment for some $\epsilon > 0$, the infectious process can be regarded as a branching process for large networks. This result is based on McKay (1985), and with more recent results (Janson 2009) we only require the second degree to be finite to be able to justify this approximation.

Using the branching process approximation Newman (2002) obtained the mean outbreak size for outbreaks that remain small (*sub-critical outbreak*).

As well, for the case of outbreaks that evolve into epidemics, Newman (2002) showed that the degrees of individuals who are infected during an outbreak have a larger mean than the overall mean degree for the vertices in the graph. We present an alternative proof of this fact taking into account that the realized infectious period of an infected individual is the same period the agent uses for transmission to any of the susceptible neighbouring vertices. We also study the variance for the degree distribution of infected individuals.

3.1 Fundamentals

We denote by K the degree random variable for the graph and by $\{p_k\}$ the degree distribution of a randomly chosen vertex. We denote the probability generating function (p.g.f.) of this distribution by $G_K(s)$.

If instead of directly choosing a vertex, we randomly select it by choosing an edge, we have that its degree distribution K_e has probability function proportional to kp_k. Since the p.g.f. of K_e is equal to

$$G_{K\mathrm{e}}(s) = \frac{sG'_K(s)}{E(K)}, \tag{1}$$

then the degree obtained after excluding the edge we arrived along has p.g.f.

$$G_{K_1}(s) = \frac{G'_K(s)}{E(K)}. \tag{2}$$

We call the latter degree the *excess degree* and denote it as K_1.

Consider a pair of individuals who are connected, one of whom, i, is infective and the other, j, is susceptible. Suppose that the infectious contacts for i and j occur as a Poisson process with rate r, and that the infective individual remains infective for a time l (*infectious period*). After this period the infective individual is considered removed. Then the probability that the individual i transmits the infection to j at some time during its infectious period is

$$\Pr(\text{disease is transmitted from } i \text{ to } j) = 1 - \mathrm{e}^{-rl}. \tag{3}$$

In pursuance of introducing individual variation of the infectious rate and period, we assume that the infectious contact rates $\{R_{ij}\}$ and infectious periods $\{I_i\}$ are random variables that are independent with distributions $F_{R_{ij}}(\cdot)$ and $F_{I_i}(\cdot)$, respectively. In this case, the conditional probability that the individual i transmits the infection to j during its entire infectious period, given that $R_i = r$ and $I_i = l$, corresponds to Eq. 3.

From here and up to Section 3.3 we consider that $\{R_{ij}\}$ and $\{I_i\}$ are independent random variables with distributions $F_R(\cdot)$ and $F_I(\cdot)$, respectively. Then the marginal probability that the infective i transmits the disease to the connected susceptible j (the *transmissibility*) is

$$\begin{aligned}\pi &:= \Pr(\text{disease is transmitted}) \\ &= \int_0^\infty \int_0^\infty \left(1 - \mathrm{e}^{-rl}\right) dF_R(r) dF_I(l) = E_{(R,I)}\left(1 - \mathrm{e}^{-RI}\right).\end{aligned}$$

An *occupied edge* is an edge of the network on which the disease is transmitted. Then an edge between an initially susceptible and an infected individual becomes occupied with probability π.

The infectious period is a random variable, but with the same value affecting all the neighbouring susceptible individuals. Then, for example, the probability that two given susceptible individuals connected to the same infectious individual become infected is

$$\int_0^\infty \left(1 - \int_0^\infty \mathrm{e}^{-rl} dF_R(r)\right)^2 dF_I(l) = E_I\left[\left(1 - E_R\left(\mathrm{e}^{-RI} | I\right)\right)^2\right].$$

In the particular case that the infectious period is constant ($I \equiv l$), the last probability is just π^2.

Due to the fact that an infected individual, or *occupied vertex*, is a vertex that is reachable by an occupied edge, we can say that the outbreak size corresponds to the size of the cluster of occupied vertices. In order to study the outbreak size we examine the occupied edges at the end of the outbreak. We define the *occupied degree* K_T and *occupied excess degree* K_{T1} of a vertex (respectively) as the number of occupied edges connected to the vertex, and the number of occupied edges resulting from the infection of the vertex.

It is important to notice that although the contact structures we study in this paper are undirected, K_T and K_{T1} are defined in terms of the infectious flow in the population. All following analyses also take into account this natural flow.

To obtain analytic results, we assume, as a simple approximation, that the contact network approximates a forest, so that once an infection spreads to a vertex, it cannot reach that vertex by any other route. Once infection spreads to a vertex along an edge, all other edges with that vertex as an endpoint can be thought of as directed away from the vertex, so that the component of occupied edges grows tree-like (Andersson 1998).

This approximation is appropriate for sub-critical outbreaks (Andersson 1998; Molloy and Reed 1995), since below that value the number of cycles in the occupied component is at most one. To derive the expressions for the mean and variance of the final outbreak size we restrict to these kinds of outbreaks.

The main consequence of infected cases growing tree-like is that the number of possible new infections coming directly from a secondary case depends only on its excess degree and the agent's transmissibility.

If T_j is the final status (1 = infected, 0 = not infected) of a single vertex connected to the patient zero i then the total number of occupied edges connected to patient zero i having neighbor vertices $j_1, \ldots, j_K$ is $\sum_{s=1}^{K} T_{j_s}$.

Since the variables $\{T_{j_s}|I=l\}$ are i.i.d. random variables with distribution which is Bernoulli $\left(1-E_R(\mathrm{e}^{-RI})\right)$, the p.g.f $G_{K_T}(s)$ of K_T is

$$G_{K_T}(s)=E_I\left[G_K\left(s+(1-s)E_R\left(\mathrm{e}^{-RI}|I\right)\right)\right]. \tag{4}$$

Similarly, the p.g.f. of the new cases originated by a secondary case is

$$G_{K_{T1}}(s)=E_I\left[G_{K_1}\left(s+(1-s)E_R\left(\mathrm{e}^{-RI}|I\right)\right)\right]. \tag{5}$$

If the infectious rate is a characteristic of the infectious individual, that is, like the infectious period, it is considered to be the same for all the connections to an infectious vertex, then Eqs. 4 and 5 become

$$G_{K_T}(s)=E_{(R,I)}\left[G_K\left(s+(1-s)\mathrm{e}^{-RI}\right)\right], \quad \text{and}$$
$$G_{K_{T1}}(s)=E_{(R,I)}\left[G_{K_1}\left(s+(1-s)\mathrm{e}^{-RI}\right)\right].$$

Under this scenario, the main relationships which follow in Section 3.2 remain the same.

In another particular case, when I is constant $\left(I\equiv l_0,\ r_0, l_0\in\mathbf{R}^+\right)$ then Eqs. 4 and 5 reduce to

$$G_{K_T}(s)=G_K\left(1+(s-1)\pi\right) \qquad \text{and} \qquad G_{K_{T1}}(s)=G_{K_1}\left(1+(s-1)\pi\right),$$

as expressed in Newman (2002).

3.2 Final Outbreak Size Mean and Variance

Let Z be the total number of infected individuals, including the initial cases. In the following computations we assume that there exists only one patient zero, but the results can easily be extended when we have more cases, as noted in Andersson (1998).

Newman et al. (2001) derived an expression for the first moment of Z based on the p.g.f.'s of Z, the occupied degree, the final total number of infected at the end of an occupied edge Z_1, and the excess occupied degrees. In this section we also obtain the variance of Z since both moments are straightforward to obtain and can provide useful information about the distribution of the final outbreak size.

Using the derivative of Z's p.g.f. Newman et al. (2001) extracted the expression

$$E(Z)=1+\frac{E(K_T)}{1-E(K_{T1})}=1+\frac{\pi E(K)}{1-\pi E(K_1)}. \tag{6}$$

The last relation can be also be derived considering that the outbreak evolves as a branching process, in discrete time by generations. If there exists only one patient zero at time 0, then the p.g.f of the number of infected at time 1 is $G_{K_T}(s)$; at time 2 the p.g.f. is $G_{K_T}\circ G_{K_{T1}}(s)$; at time 3 the p.g.f. is $G_{K_T}\circ G_{K_{T1}}\circ G_{K_{T1}}(s)$, and so on. Then the mean numbers of individuals infected at times 1, 2, 3 are $E(K_T)$, $E(K_T)E(K_{T1})$ and $E(K_T)[E(K_{T1})]^2$, respectively. Hence the mean of the total number of individuals that have been infected up to time n is

$$1+\sum_{m=0}^{n-1}E(K_T)[E(K_{T1})]^m,$$

that converges to Eq. 6 as $n\to\infty$ when $E(K_{T1})<1$.

From Eq. 6 we have that $E(Z)$ diverges when $E(K_{T1}) = \pi E(K_1)$ approaches 1. The quantity $\pi E(K_1)$ coincides with the replacement number $\mathcal{R}$, representing the expected excess number of occupied edges of a typical infected vertex. In fact, it is easy to see that the outbreak has a finite mean size if and only if $\mathcal{R} < 1$.

Furthermore, from Eq. 6 we have the critical transmissibility threshold for the outbreak to become an epidemic:

$$\pi_c := \frac{1}{G'_{K_1}(1)} = \frac{E(K)}{E(K^2) - E(K)}. \tag{7}$$

Then if $\pi < \pi_c$ the outbreak remains finite with probability 1 and $E(Z) < \infty$, while if $\pi > \pi_c$, the outbreak becomes an epidemic (in the sense that $E(Z) = \infty$) with positive probability.

Following Pastor-Satorras and Vespignani (2001) (also Pastor-Satorras and Vespignani 2003), we note that the transmission threshold (Eq. 7) approaches zero as $E(K^2)$ increases. Thus, in a network with a large enough degree variance, any infectious disease outbreak will have the potential to turn into an epidemic.

For some specific cases, Newman (2002) and Meyers et al. (2003) obtained the mean final outbreak size, and they compared their results using the average final outbreak size observed in a large number of computational simulations. However, the authors did not explicitly describe the behavior of the deviations from the mean.

The outbreak size variance allows us to describe the most likely final scenarios, providing more information to help decision makers prepare the necessary resources to handle future outbreaks. In Section 3.2.1 we describe how the range of scenarios can change for a fixed mean outbreak size. Thus, in order to complement Newman's results (Newman 2002), we next derive the expression for the variance of the final outbreak size.

The variance of the final outbreak size is

$$\begin{aligned} \mathrm{Var}(Z) &= \mathrm{Var}(K_T E(Z_1) + 1) + E(K_T \mathrm{Var}(Z_1)) \\ &= E(Z_1)^2 \mathrm{Var}(K_T) + \mathrm{Var}(Z_1) E(K_T) \end{aligned} \tag{8}$$

where Z_1 is the size of the component at the end of a randomly chosen occupied edge.

Similarly

$$\mathrm{Var}(Z_1) = E(Z_1)^2 \mathrm{Var}(K_{T1}) + \mathrm{Var}(Z_1) E(K_{T1}).$$

Now, since

$$E(Z_1) = \frac{1}{1 - E(K_{T1})},$$

then

$$\mathrm{Var}(Z_1) = \frac{E(Z_1)^2 \mathrm{Var}(K_{T1})}{1 - E(K_{T1})} = E(Z_1)^3 \mathrm{Var}(K_{T1}),$$

and it follows that

$$\mathrm{Var}(Z) = E(Z_1)^2 \left[\mathrm{Var}(K_T) + E(Z_1) \mathrm{Var}(K_{T1}) \pi E(K) \right], \tag{9}$$

where

$$\mathrm{Var}(K_T) = \left[\mathrm{Var}(K) + E(K)^2 - E(K)\right]\pi_2 - E(K)^2\pi^2 + E(K)\pi, \tag{10}$$

$$\mathrm{Var}(K_{T1}) = \left[\mathrm{Var}(K_1) + E(K_1)^2 - E(K_1)\right]\pi_2 - E(K_1)^2\pi^2 + E(K_1)\pi, \tag{11}$$

and

$$\pi_2 = E_I\left[\left(1 - E_R\left(\mathrm{e}^{-RI}|I\right)\right)^2\right],$$

$$E(K_1) = \frac{E(K^2)}{E(K)} - 1, \text{ and}$$

$$\mathrm{Var}(K_1) = \frac{1}{E(K)}\left(E(K^3) - \frac{E(K^2)^2}{E(K)}\right).$$

Hence $\mathrm{Var}(Z)$ can be computed using the expressions 9, 10 and 11.

From Eq. 9 it is clear that the variance for the final outbreak size depends on the infectious rate and period through π and π_2 and on the connectivity structure through the first three moments of K. Hence, fixing the distributions of R and I, the final outbreak results can vary widely depending only on the heterogeneity of the connectivity pattern (skewness of the degree distribution). Then even if $E(Z) < \infty$ with probability 1, a large number of individuals may be infected for large values of $E(K^3)$.

In order to limit the number of infected individuals, control measures usually have as a goal to decrease $\mathcal{R}$ to a value less than 1 by reducing the agent's transmissibility, the expected number of contacts or the size of the susceptible population. The above results suggest that it is important to reduce concomitantly the variance of the excess degree. To achieve this goal, for example, some vaccination procedures target individuals with high degree since this can result in the use of fewer vaccines than when randomly selecting them.

3.2.1 Output Variation Due to Network Heterogeneity ($\mathcal{R} < 1$)

To illustrate the possible outbreak outcomes we consider three networks with different heterogeneity levels. They are networks with degenerate (m), Poisson (λ) and polylogarithmic (δ, γ) degree distributions. That is,

$${}_1p_k = \begin{cases} 1 & \text{if } k = m, \quad m \in \{2, 3, ...\}, \\ 0 & \text{otherwise;} \end{cases}$$

$${}_2p_k = \begin{cases} \dfrac{\lambda^k \exp(-\lambda)}{\lambda!} & \text{if } k \in \{0, 1, \ldots\}, \quad \lambda > 0, \\ 0 & \text{otherwise;} \end{cases}$$

$${}_3p_k = \begin{cases} \dfrac{k^{-\delta}\exp(-k/\gamma)}{\mathrm{Li}_\delta(\exp(-1/\gamma))} & \text{if } k \in \{1, 2, \ldots\}, \quad \delta > 1, \gamma > 0, \\ 0 & \text{otherwise.} \end{cases}$$

The three kinds of networks are paired to compare the variability of the final outbreak sizes, in the cases when their mean final outbreak sizes are the same.

We pair the degenerate and Poisson networks and we call this pair ($N1$, $N2$). We consider also the pair formed by the Poisson and polylogarithmic networks ($N2$, $N3$). We suppose that R and I are constants ($R \equiv r$, $I \equiv l$), so that $\pi = 1 - \exp(-rl)$.

For the first pair of networks, the parameter of the degenerate network is $m = 4$ and the Poisson network's parameter is set equal to

$$\lambda(\pi) = n/(1+\pi).$$

In Fig. 1a the black line is the *natural logarithm* of their theoretical $E(Z)$ (denoted as $log(E(Z))$), that coincides for all values of π.

Similarly, for the second pair ($N2$, $N3$) we fixed the parameters of the polylogarithmic distribution and obtained the parameter of the Poisson network that will result in the same $log(E(Z))$. The parameters of the polylogarithmic network are $\delta = 1.3$, $\gamma = 50$, and for a given value of π, the parameter of the Poisson network is

$$\lambda(\pi) = \frac{\mathrm{Li}_{\delta-1}(a)/\mathrm{Li}_{\delta}(a)}{1-\pi\left[\mathrm{Li}_{\delta-2}(a)/\mathrm{Li}_{\delta-1}(a) - \mathrm{Li}_{\delta-1}(a)/\mathrm{Li}_{\delta}(a) - 1\right]},$$

where $a = \exp(-1/\gamma)$.

As in Fig. 1a, Fig. 2a depicts the common $log(E(Z))$ for this second pair of networks.

In order to verify the agreement between the expressions derived for the mean and variance of the final outbreak sizes and their empirical values in network simulations, we simulated 10,000 outbreaks for each element in the pair of networks to be compared (each network had 100,000 vertices), considering a handful of different values of π. The logarithms of the average simulated final outbreak sizes are depicted in Figs. 1a and 2a as crosses and circles.

Figures 1b and 2b display the logarithms of the standard deviation differences, for the final outbreak size for networks belonging to the pairs ($N1$, $N2$) and ($N2$, $N3$), respectively. The logarithms of the standard deviation differences based on the

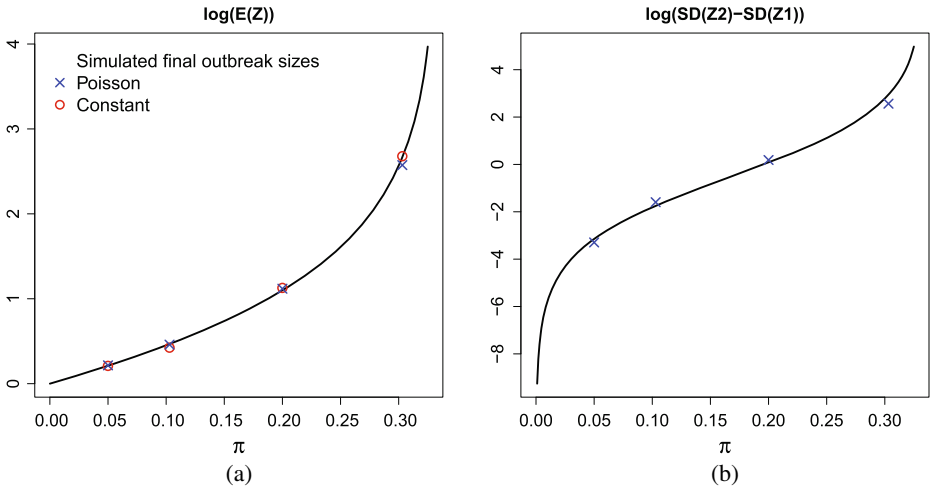

Fig. 1 Log mean final outbreak size and log difference of the variances for the paired networks ($N1$, $N2$)

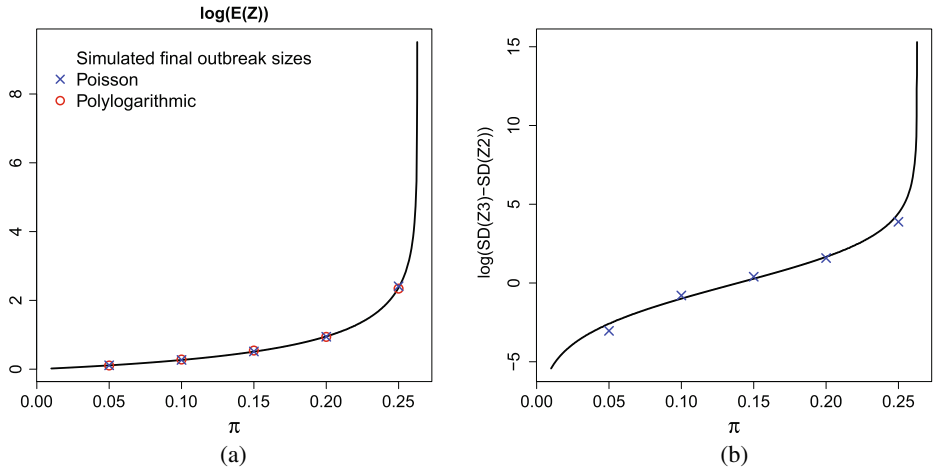

Fig. 2 Log mean final outbreak size and log difference of the variances for the paired networks ($N2$, $N3$)

simulated outbreaks are included as blue crosses. The variances for the pair ($N1$, $N2$) are very similar for small π and they increase with π. The same increasing tendency is observed in the second pair of networks, but in this case the difference grows more rapidly.

In the case of the first pair ($N1$, $N2$), the outbreaks will tend to affect similar numbers of individuals, while in the second case ($N2$, $N3$) we see that the mean final outbreak sizes are equal, but the variance of outbreak size for the network with polylogarithmic degree distribution can be much larger. This characteristic points to the fact that for larger values of π, while most of the outbreaks occurring in this network will die out almost immediately after starting, a few outbreaks will result in a large number of infected individuals. To put this another way, when $\mathcal{R} < 1$ it is usual to observe more infected individuals in a homogeneous than in a highly heterogeneous network. In this sense the heterogeneity tends to *protect* the population from larger outbreaks, but at the same time there exists the possibility that the outbreak evolves into a much larger outbreak than if evolving in a homogeneous network. This property reflects the shape of the degree distribution since, clearly, one factor closely related to the total number of infected individuals is the degree of the first individual(s) that is (are) infected.

Based on the one-sided Chebyshev inequality we can obtain crude probability intervals for the final outbreak size Z based only on its mean μ and variance σ^2 as

$$P(Z < k\sigma + \mu) \geq \frac{k^2}{1 + k^2}, \tag{12}$$

where $k \in \mathbf{R}^+$.

The bounds provided by the Chebyshev inequality are conservative but in general cannot be improved when distributions are arbitrary.

Where the mean is the same, as for the paired graphs ($N1$, $N2$) and ($N2$, $N3$), then for a fixed value of k in Eq. 19, the Chebyshev probability intervals for Z depend exclusively on the values of the variances.

3.2.2 Conditional Mean and Variance ($\mathcal{R} > 1$)

As noted above, when $\mathcal{R} < 1$, the final outbreak size has a proper distribution and the mean and variance are given by Eqs. 6 and 9.

From the properties of the p.g.f., the probability that the final size outbreak grows to epidemic levels is $1 - G_{K_T}(u)$ where u is the smallest root of

$$G_{K_{T_1}}(u) = u. \tag{13}$$

See Andersson (1998) and Durrett (2007).

If $\mathcal{R} \leq 1$ then $u = 1$ and all outbreaks remain small. On the other hand, if $\mathcal{R} > 1$, then $u < 1$ and there exists a positive probability that an outbreak becomes an epidemic. Based on this result, Meyers (2007) noted that the probability of an epidemic, when $\mathcal{R} > 1$, conditioned on patient zero having degree k, is $1 - (1 - \pi + \pi u)^k$.

On the other hand, as pointed out by Yan (2008), when $\mathcal{R} > 1$ some outbreaks will become epidemics and some will stay small and those outbreaks that remain small will be indistinguishable from a branching process with replacement number:

$$\mathcal{R}^* = G'_{K_{T_1}}(u) < 1. \tag{14}$$

For details refer to Athreya and Ney (1972) or Grimmett and Stirzaker (2001).

Then Eqs. 6 and 8 can be generalized to the conditional mean and variance:

$$E(Z|\text{small outbreak, prob small outbreak} = u) = 1 + \frac{E(K_T)}{1 - E(K_{T1}|u)}, \tag{15}$$

$$\begin{aligned} &\text{Var}(Z|\text{small outbreak, prob small outbreak} = u) \\ &= [E(Z_1|u)]^2 \,\text{Var}(K_T) + \text{Var}(Z_1|u)E(K_T), \end{aligned} \tag{16}$$

where

$$\begin{aligned} E(K_{T1}|u) &= G'_{K_{T_1}}(u), \\ \text{Var}(K_{T1}|u) &= G''_{K_{T_1}}(u) + E(K_{T1}|u)(1 - E(K_{T1}|u)) \\ E(Z_1|u) &= \frac{1}{1 - E(K_{T1}|u)}, \text{ and} \\ \text{Var}(Z_1|u) &= [E(Z_1|u)]^3 \text{Var}(K_{T1}|u). \end{aligned}$$

3.2.3 Example

We consider simulated outbreaks with constant transmissibility π, in Poisson($\lambda = 2.101$) and Polylogarithmic($\delta = 2$, $\gamma = 13$) networks with 100,000 vertices. For each kind of network and each of three different transmissibility values, we simulated 10,000 outbreaks using 100 networks, and obtained the probability intervals derived from Eq. 12 and their coverage (Table 2).

Table 1a, b presents the theoretical and empirical mean and standard deviation of the final outbreak sizes in each network and Table 2 displays the upper bounds for the final outbreak sizes derived from these statistics and their respective coverage for three different probability levels.

Table 1 Theoretical and empirical final outbreak size mean and SD

		Theoretical	Empirical
(a) Poisson(2.101)-network			
$\pi = 0.10, \mathcal{R} = 0.21$	Mean	1.26	1.27
	SD	0.65	0.67
$\pi = 0.20, \mathcal{R} = 0.42$	Mean	1.72	1.71
	SD	1.47	1.43
$\pi = 0.30, \mathcal{R} = 0.63$	Mean	2.70	2.71
	SD	3.53	3.65
(b) Polylog(2,13)-network			
$\pi = 0.10, \mathcal{R} = 0.38$	Mean	1.31	1.33
	SD	1.09	1.17
$\pi = 0.20, \mathcal{R} = 0.76$	Mean	2.59	2.56
	SD	8.52	7.88
$\pi = 0.25, \mathcal{R} = 0.95$	Mean	10.71	9.06
	SD	117.46	66.79
(c) Polylog(2,13)-500,000 vertices network			
$\pi = 0.10, \mathcal{R} = 0.38$	Mean	1.31	1.32
	SD	1.09	1.25
$\pi = 0.20, \mathcal{R} = 0.76$	Mean	2.59	2.93
	SD	8.52	10.76
$\pi = 0.25, \mathcal{R} = 0.95$	Mean	10.71	12.57
	SD	117.46	117.59

In the Poisson case, from Table 1a, we can observe a remarkable agreement between the theoretical and empirical final outbreak size statistics. However, there exists an appreciable difference between these results when considering the Polylogarithmic networks (Table 1b), especially as π increases.

In order to measure the impact that networks with finite number of vertices have on potentially large outbreaks, we simulated a single Polylogarithmic($\delta = 2, \gamma = 13$) network with 500,000 vertices. Using this network, a total of 10,000 outbreaks were simulated (always with a randomly selected patient zero) and the corresponding mean and standard deviation of the final outbreak sizes are presented in Table 1c. From this table we can appreciate that by increasing the network order to 500,000 vertices, we have good agreement between the derived theoretical two moments and their empirical counterparts, even in the case where $\mathcal{R}$ is close to 1.

In Table 1a–c all of the mean final outbreak sizes are very small (less than 11 individuals). However, the variances of the final outbreak sizes capture widely differing scenarios. In the case of a Poisson network, most of the outbreaks would have final size not far from the mean value, but in the case of the polylogarithmic network, for example when $\pi = 0.25$, it is clear from Table 1b, c that many of the outbreaks will die out almost immediately and a few will expand to infect hundreds of individuals.

In Table 2, the second column (in each case) corresponds to the upper bound for Z, namely $(k\sigma + \mu)$ as in Eq. 12, for three different probability levels $l = k^2/(1 + k^2)$. The values μ and σ correspond to Eqs. 6 and 9 presented in Table 1.

The third columns of Table 2, like the second ones, show upper bounds for Z, but here we use the empirical mean and variance of the final outbreak size based on the 10,000 simulated outbreaks, given in Table 1 for the Poisson network and Table 1c for the Polylogarithmic network.

Table 2 Probability intervals for the final outbreak size

Prob. level	Upper bound	Emp. upper bound	Coverage (%)
Poisson(2.101)-network			
$\pi = 0.10$			
80 %	2.57	2.61	94.95
90 %	3.22	3.27	98.20
95 %	4.11	4.19	99.36
$\pi = 0.20$			
80 %	4.66	4.57	95.21
90 %	6.12	6.05	98.28
95 %	8.12	7.97	98.86
$\pi = 0.30$			
80 %	9.76	10.01	96.60
90 %	13.30	13.67	98.06
95 %	18.10	18.64	99.12
Polylog(2,13)-network			
$\pi = 0.10$			
80 %	3.49	3.36	97.42
90 %	4.59	4.39	98.42
95 %	6.08	5.80	98.94
$\pi = 0.20$			
80 %	19.63	15.91	97.76
90 %	28.15	22.63	98.57
95 %	39.73	31.77	99.10
$\pi = 0.25$			
80 %	245.62	142.65	98.84
90 %	363.07	209.44	99.12
95 %	522.68	300.20	99.33
Polylog(2,13)-500,000 vertices network			
$\pi = 0.10$			
80 %	3.49	3.82	97.12
90 %	4.59	5.07	98.82
95 %	6.08	6.77	99.25
$\pi = 0.20$			
80 %	19.63	24.45	98.17
90 %	28.15	35.21	98.83
95 %	39.73	49.84	99.29
$\pi = 0.25$			
80 %	245.62	247.78	99.13
90 %	363.07	365.37	99.35
95 %	522.68	525.17	99.48

As can be observed, in the polylogarithmic case the theoretical probability intervals tend to be larger than the empirically derived intervals for all levels.

Finally, the fourth column, in each part of Table 2, is the proportion of simulated outbreaks whose size falls in the probability interval given by the third column.

For the Poisson settings, the means and variances of the final outbreak sizes estimated from the simulated outbreaks are close to the means and variances of Eqs. 6 and 9, based on the branching process approximation; this fact is reflected in the closeness of the second and third columns in the Poisson part of Table 2.

For the polylogarithmic settings with 100,000 vertices, there is a large discrepancy between second and third columns, particularly in the last case where the theoretical upper bound is much larger. This results from a difference between the estimated variance in the simulation and the theoretical variance based on the branching process approximation. We see that for a network of 100,000 vertices with the parameter settings of Table 2, to obtain relatively tight Chebyshev bounds, simulation is really necessary in the case of the polylogarithmic network, while the theoretical calculations are sufficient in the Poisson case. For a network of 500,000 vertices, the theoretical calculations are adequate for both. Clearly, the Chebyshev bounds are conservative, and entirely empirical probability bounds would be smaller in all cases, but obtainable at greater computing cost.

3.3 Mean Degree of Infected and Uninfected Individuals

Since the infection transmits through the network edges, it is natural to expect that the larger the degree of a vertex, the more likely the vertex will be affected during the outbreak. A kind of converse is also true, namely that when the outbreak has taken its course, the mean degree of affected vertices is larger than the mean degree of unaffected vertices. In this section we present a new proof of this result, in the context of our model. The result is important for inference about the contact network from samples of affected and unaffected vertices and supports control measures that target the vaccination (protection or isolation) of individuals that are highly connected.

Since the outbreaks transmit along randomly chosen edges, when the outbreak remains small ($\mathcal{R} < 1$), the degree distribution of secondary infected individuals (K_e) is $P(K_e = k) = kp_k/E(K)$. Then, the degree mean of infected individuals is $E(K_e) = E(K^2)/E(K)$, and by Jensen's inequality

$$E(K_e) \geq \frac{[E(K)]^2}{E(K)} = E(K).$$

If $\Pr(K = k)$ is symmetric or negatively skewed then $\mathrm{Var}(K_e) < \mathrm{Var}(K)$. In the case that the degree distribution is positively skewed, we have the following cases

$$\mathrm{Var}(K_e) \quad \left\{ \begin{matrix} > \\ = \\ < \end{matrix} \right\} \quad \mathrm{Var}(K) \quad \text{if} \quad \gamma \quad \left\{ \begin{matrix} > \\ = \\ < \end{matrix} \right\} \quad SD(K)/E(K),$$

where γ is the third standardized moment for $\Pr(K = k)$. Then, for example, if the degree distribution is Poisson(λ), $\gamma = 1/\sqrt{\lambda} = SD(K)/E(K)$, and hence $\mathrm{Var}(K_e) = \mathrm{Var}(K)$ for all λ.

When $\mathcal{R} > 1$, a different argument can be used to compare the degree distribution of vertices inside and outside an epidemic.

Assuming that an infection begins at a randomly chosen vertex in a randomly generated graph with degree distribution $\{p_k\}$, and assuming that there is an epidemic, the probability that any particular vertex is not in a component of occupied vertices is a constant.

Again assuming that there is an epidemic, the probability that a vertex is not affected by the epidemic via one of its edges, given that its degree is k, is

$$\begin{aligned}\Pr\left(\text{vertex not in epidemic} \mid \text{its degree is } k\right) &= \sum_{i=0}^{k} \binom{k}{i} q^{k-i} \left[(1-q)\, E_{(R,I)}\left(\mathrm{e}^{-RI}\right)\right]^i \\ &= \left[q + (1-q)\, E_{(R,I)}\left(\mathrm{e}^{-RI}\right)\right]^k \\ &= \left[q + (1-q)(1-\pi)\right]^k, \end{aligned} \tag{17}$$

where q is the probability that a vertex selected by randomly choosing one edge (from which our vertex can be reached) is not in the epidemic; it is assumed that this event can be regarded as independent from edge to edge.

The parameter q can be computed as follows. We know that if $\mathcal{R} = G'_{K_{T1}}(1) > 1$ the probability that an outbreak does not evolve into an epidemic is $G_{K_T}(u)$, where u is the smallest root of Eq. 13. As before, let K_e be the degree of a vertex selected by randomly choosing one edge. Then K_e has p.f. $kp_k/E(K)$ and

$$\begin{aligned}&\Pr\left(\begin{array}{c|c} \text{vertex selected by randomly} & \text{its degree is } k_\mathrm{e} \\ \text{choosing one edge is not in epidemic} & \end{array}\right) \\ &= G_{K_T}(u) + (1 - G_{K_T}(u))\left[q + (1-q)(1-\pi)\right]^{k_\mathrm{e}}. \end{aligned}$$

Then q can be obtained solving

$$q = G_{K_T}(u) + (1 - G_{K_T}(u)) G_{K_\mathrm{e}}\left(q + (1-q)(1-\pi)\right). \tag{18}$$

Hence the degree distribution of a vertex given that it *was not* and *was* infected in the epidemic are respectively

$$\Pr\left(\text{degree is } k \mid \text{ not in epidemic}\right) = \frac{\left[q + (1-q)(1-\pi)\right]^k p_k}{G_K\left(q + (1-q)(1-\pi)\right)}$$

and

$$\Pr\left(\text{degree is } k \mid \text{ in epidemic}\right) = \frac{\left(1 - \left[q + (1-q)(1-\pi)\right]^k\right) p_k}{1 - G_K\left(q + (1-q)(1-\pi)\right)}.$$

Based on the conditional degree distributions we can easily calculate the respective conditional degree means. They are

$$E(K|\text{not in epidemic}) = \frac{\left[q + (1-q)(1-\pi)\right] G'_K\left(q + (1-q)(1-\pi)\right)}{G_K\left(q + (1-q)(1-\pi)\right)}$$

and

$$E(K|\text{in epidemic}) = \frac{E(K) - \left[q + (1-q)(1-\pi)\right] G'_K\left(q + (1-q)(1-\pi)\right)}{1 - G_K\left(q + (1-q)(1-\pi)\right)}.$$

Next we will prove that the mean degree of affected vertices is larger than the mean degree of the vertices that are unaffected.

To prove that $E(K|\text{in epidemic}) \geq E(K|\text{not in epidemic})$ it is enough to show

$$E(K)G_K(z) - zG'_K(z) \geq 0 \tag{19}$$

where $z = q + (1-q)(1-\pi)$.

Due to the fact that $E(K) > 1$ there exists u such that $G'_k(u) = u$. If $z < u$ then $z < G_K(z)$, and since $G'_K(z) < G'_K(1)$, then Eq. 19 immediately follows.

If $z > u$ we use the expansion of the function

$$H(z) = E(K)G_K(z) - zG'_K(z),$$

expressed as

$$\begin{aligned} H(z) &= \sum_k \sum_j k p_k p_j z^j - \sum_j j p_j z^j \\ &= \sum_j z^j \left\{ \sum_k k p_k p_j - j p_j \right\} = \sum_j z^j \left\{ p_j [E(K) - j] \right\}. \end{aligned} \tag{20}$$

The coefficients of z^j: $c_j = p_j[E(K) - j]$ are non-negative for values of $j \leq E(K)$, and negative for $j > E(K)$, and they sum up to 0.

Now, since $z \in (0, 1)$, the function z^j is decreasing in j, and then the sum of non negative terms in Eq. 20 is larger that the absolute value of sum of the rest of the elements. Therefore $H(z) \geq 0$.

Then inequality in Eq. 19 is then strict for values of $z \in (0, 1)$ and equal to 0 in the trivial case $z = 1$.

3.4 Individual Transmission Rates

There are several infections in which is important to consider that transmission rates are not i.i.d. r.v.'s, that is, that the probability of transmission from a given individual i to another j could be drawn from different distributions for different individuals.

A particular case of this variation is related to age-dependent epidemic models. These kinds of models define, by age levels or as a function of age, parameters such as mortality and birth rates, attack rates and infectious period parameters (as in Capasso 2008, Ch. 6). Age-dependent models are important for some diseases, such as the so-called childhood infections, that tend to target subpopulations. In the case of smallpox, the age levels define not only the susceptible population but also the subpopulations corresponding to different susceptibility levels: children younger and older than 12 months.

In the context of network epidemics, the age-dependent models can also be extended to describe the mixing patterns between and within age groups. An example is the use of bipartite networks (Meyers et al. 2003; Hyman et al. 2007) to describe group membership, which could be age-dependent. However, in this section we extend our models consider only different finite population distribution functions for the infectious period and transmission rates, within a random contact network. That is, we consider the contact network to be essentially independent of the individual characteristics that define the distributions for the infectious period and transmission

rate. We take the transmission rate to depend on the transmitter but not on the recipient.

Generalizing the model studied in the last sections, here we consider that the infectious contact rate and infectious period $\{R_i\}$ and $\{I_i\}$ are two series of independent random variables with distributions $\{F_{R_i}(\cdot)\}$ and $\{F_{I_i}(\cdot)\}$, so that the probability of transmission is

$$\pi_{ij} := \Pr(i \text{ transmit to } j) = \int_0^\infty \int_0^\infty \left(1 - \mathrm{e}^{-rl}\right) dF_{R_i}(r) dF_{I_i}(l).$$

In contrast to the model presented in the last two subsections, suppose that the distribution of R_i does vary with respect to the infective individual. Thus, assume that the transmission rate from an infective i to each of the k_i others to whom it is connected follows a distribution $F_{R_i}(\cdot)$, which can be different for each individual i. The individual distribution variations can be modeled depending on covariates such as sex, age, ethnicity or health variables.

Similarly we can assume that the distribution of the infectious period I_i is different for each individual i. Thus the set of individual transmission probabilities is

$$\pi_i := \Pr(i \text{ transmit to } j) = \int_0^\infty \int_0^\infty \left(1 - \mathrm{e}^{-rl}\right) dF_{R_i}(r) dF_{I_i}(l).$$

Now, let N be the number of vertices in the graph, where N is large. Since the occupied degree of a patient zero that is randomly selected is equal to m with probability

$$\begin{aligned}
&\Pr\left(m \text{ occupied edges proceeding from a randomly chosen vertex}\right) \\
&= \sum_{i=1}^{N} E_{I_i}\left[E_{R_i}\left[\sum_{k=m}^{\infty} \Pr\left(m \text{ occupied edges} \;\middle|\; \begin{array}{l}\text{selected vertex is } i,\\ \text{vertex has degree } k,\ R_i,\ I_i\end{array} \right) \right.\right. \\
&\qquad \left.\left. \times \Pr\left(\text{degree } k \mid \text{select vertex } i\right) \times \Pr(\text{select vertex } i) \right]\right] \\
&= \sum_{i=1}^{N} E_{I_i}\left(\sum_{k=m}^{\infty} \binom{k}{m} \left(1 - E_{R_i}\left(\mathrm{e}^{-R_i I_i} | I_i\right)\right)^m \left(E_{R_i}\left(\mathrm{e}^{-R_i I_i} | I_i\right)\right)^{k-m} p_k \frac{1}{N} \right),
\end{aligned}$$

then the p.g.f. for the occupied degree distribution of patient zero is

$$G_{K_T}(s) = \frac{1}{N} \sum_{i=1}^{N} E_{I_i}\left[G_K\left(s + (1-s) E_{R_i}\left(\mathrm{e}^{-R_i I_i} | I_i\right)\right)\right]. \tag{21}$$

For a secondary case, the distribution of the number of vertices this vertex infects is

$$\begin{aligned}
&\Pr\left(\text{excess occupied degree of vertex is } m \mid \text{vertex was infected}\right)\\
&= \sum_{i=1}^{N} E_{(R_i,I_i)}\left[\sum_{k=m+1}^{\infty} \Pr\left(m \text{ exc. occ. degree} \mid \text{infected, degree } k,\, R_i,\, I_i\right)\right.\\
&\qquad\left.\times \Pr\left(\text{degree } k \mid \text{infected } i\right) \Pr\left(i \mid \text{infected}\right)\right]\\
&= E_{I_i}\left(\sum_{i=1}^{N}\sum_{k=m+1}^{\infty}\binom{k-1}{m}\left(1-E_{R_i}\left(\mathrm{e}^{R_iI_i}|I_i\right)\right)^m\left(E_{R_i}\left(\mathrm{e}^{-R_iI_i}|I_i\right)\right)^{k-1-m}\frac{kp_k}{E(K)}\frac{1}{N}\right).
\end{aligned}$$

Hence the p.g.f. for the occupied excess degree of a vertex that is a secondary case is

$$\begin{aligned}
G_{K_{T1}}(s) &= \frac{1}{NE(K)}\sum_{i=1}^{N} E_{I_i}\left[\sum_{k=0}^{\infty} kp_k\left(s\left(1-E_{R_i}\left(\mathrm{e}^{-R_iI_i}|I_i\right)\right)+E_{R_i}\left(\mathrm{e}^{-R_iI_i}|I_i\right)\right)^{k-1}\right]\\
&= \frac{1}{NE(K)}\sum_{i=1}^{N} E_{I_i}\left[G_k'\left(s+(1-s)E_{R_i}\left(\mathrm{e}^{-R_iI_i}|I_i\right)\right)\right]\\
&= \frac{1}{N}\sum_{i=1}^{N} E_{I_i}\left[G_{K_1}\left(s+(1-s)E_{R_i}\left(\mathrm{e}^{-R_iI_i}|I_i\right)\right)\right]. \qquad (22)
\end{aligned}$$

Newman (2002) also derived the p.g.f. for the occupied degree and occupied excess degree, but effectively under the assumption that an infected vertex shows independent infectious periods to each of its edges. Our expressions 21 and 22 are different. These are the weighted average of the individual p.g.f.'s for the occupied degrees and occupied excess degrees, and it is easy to see that when $\{R_i\}$ and $\{I_i\}$ are both identically distributed they reduce to Eqs. 4 and 5.

Using Eqs. 21 and 22 the results from Section 3.2 can be used immediately. For example, the mean final outbreak size Eq. 6 becomes

$$E(Z) = 1 + \frac{E(K)\frac{\sum\pi_i}{N}}{1-E(K_1)\frac{\sum\pi_i}{N}},$$

and its variance 8 has similar changes:

$$\mathrm{Var}(Z) = \frac{\mathrm{Var}(K_T)}{\left[1-E(K_1)\frac{\sum\pi_i}{N}\right]^2} + \mathrm{Var}(Z_1)\frac{E(K)\sum\pi_i}{N},$$

where

$$\mathrm{Var}(Z_1) = \frac{\mathrm{Var}(K_{T1})}{\left[1 - E(K_1)\frac{\sum \pi_i}{N}\right]^3},$$

$$\mathrm{Var}(K_T) = \left[\mathrm{Var}(K) + E(K)^2 - E(K)\right]\frac{\sum \pi_{i(2)}}{N} - E(K)\frac{\sum \pi_i}{N} \times \left(E(K)\frac{\sum \pi_i}{N} - 1\right),$$

$$\mathrm{Var}(K_{T1}) = \left[\mathrm{Var}(K_1) + E(K_1)^2 - E(K_1)\right]\frac{\sum \pi_{i(2)}}{N} - E(K_1)\frac{\sum \pi_i}{N} \times \left(E(K_1)\frac{\sum \pi_i}{N} - 1\right),$$

and

$$\pi_{i(2)} = E_{I_i}\left[\left(1 - E_{R_i}\left(\mathrm{e}^{-R_i I_i} | I_i\right)\right)^2\right].$$

4 Discussion and Future Work

Based on our results we argue that in order to picture the most likely outbreak scenarios it is of high value to compute the variance of the final outbreak size. As shown by some examples, the variability is not only related to the variability of R and I but is also strongly linked to the contact network's heterogeneity.

In terms of control measures, the probability intervals allow us to measure the impact that control measures can have on the upper bounds for the most likely outbreak sizes, whether they aim to decrease the transmissibility-susceptibility or to modify the population contact structure.

In this paper we have obtained the expression for the mean and variance of the final outbreak sizes for a random network with any distribution on the integers.

Here we have also reformulated some of the most important results in Andersson (1998) and Newman (2002), modifying some of them to take into account the important fact that the transmissibility of a vertex i to any of its neighbors is affected by the same realized infectious period of i.

The principal assumption of the model considered here is that the probability of loops in the graph is negligible and the infection grows tree-like. This characteristic, as noted by Keeling (1999), Newman et al. (2001), Watts (2002), can be achieved approximately by considering graphs of high order.

However, in practical cases we may have data from small networks (populations in long term care facilities, hospitals, etc.). The estimation of the outbreak sizes can be drawn from simulations as presented in Section 3.2.1 and the example in Section 3.2.3. Nevertheless, a valuable extension would be to correct the results, adding the effect of cycles on the excess degree distribution as the infection spreads.

Since in communities or large populations we have a combination of networks with different connectivity (schools, hospitals, work centers, etc.), it is necessary to

generalize the results for more complex structures (one special case is the bipartite population also considered by Newman (2002)).

Due to the fact that the epidemic curve (number of infected individuals over time) is an important outbreak characteristic used for inferring the infectivity of an agent during the first stages of the outbreak, it would be of great value to study the outbreak dynamic in networks over time.

An important assumption in the epidemic model presented here is that the outbreak evolves very fast compared to demographic and social changes, so that the network of contacts remains static except for the elimination of edges and vertices arising from vertices entering into the removal stage after being infected and infectious. Although many infectious agents of interest have a short evolution in the host, it is necessary to relax this supposition in order to include infections such as AIDS.

Acknowledgements The research presented in this paper was funded by CONACYT (Grant number: 119868) and NSERC (Grant number: RGPIN8146-04) and made possible by the facilities the Shared Hierarchical Academic Research Computing Network (SHARCNET: www.sharcnet.ca) and Compute/Calcul Canada.

Appendix

The random graphs simulated in this work are constructed based on the algorithm described in Molloy and Reed (1995). To increase its computational efficiency, it is divided into two iterative steps. The first one selects a group of edges, where groups are defined by the degrees of the vertices they connect. In the second step a specific edge is randomly chosen from the already selected group. A detailed description of this algorithm is presented next.

1. A sequence of degrees with length n is simulated randomly from the specified distribution.
2. The sequence is assigned to the nodes identified with labels from 1 to n.
3. A table $\mathcal{D}$ with the frequencies ($F_{\mathcal{D}}$) of degree values ($V_{\mathcal{D}}$) is generated. Based on $\mathcal{D}$, the table $\mathcal{PD}$ with the frequencies ($F_{\mathcal{PD}}$) of the possible product values ($V_{\mathcal{PD}}$) is obtained. Table $\mathcal{PD}$ does not completely aggregate $V_{\mathcal{PD}}$ by value, but by the pairs of $V_{\mathcal{D}}$ leading to the values. This makes possible to identify for each edge selected from $\mathcal{PD}$ the degrees of its two endpoints. The frequency $F_{\mathcal{PD}}$ of edges with endpoints with degree d_1 and d_2 corresponds to the number of possible edges that connect to vertices with degrees d_1 and d_2. Thus, it is equal to $F_{\mathcal{D}}(d_1) \times F_{\mathcal{D}}(d_2)$ if $d_1 \neq d_2$, and $F_{\mathcal{D}}(d_1) \times [F_{\mathcal{D}}(d_1) - 1]/2$ if $d_1 = d_2$, where $F_{\mathcal{D}}(d_i)$ is the number of vertices with degree d_i.
4. In each step an edge in $\mathcal{PD}$ is randomly selected with weighted probability proportional to $F_{\mathcal{PD}} \times V_{\mathcal{PD}}$.
5. If the selected edge connects vertices with degree d_1 and d_2 then two vertices are randomly sampled from the sets of vertices with degree d_1 and d_2, respectively.
6. The degree tables $\mathcal{D}$ and $\mathcal{PD}$ are updated and the process repeats from step 4 until no more edges remain or can be allocated.

As proven by the fit shown in Tables 1a–c, this network generating algorithm proves to be adequate to randomly simulate networks with specified degree sequences and as implemented is shown to be time and memory efficient.

The algorithm is implemented in R Development Core Team (2007) and it can be provided by request to lilialeticia.ramirez@itam.mx.

References

Albert R, Jeong H, Barabási A-L (1999) Diameter of the world-wide web. Nature 400:130–131

Albert R, Jeong H, Barabási A-L (2000) Attack and error tolerance of complex networks. Nature 406:378–382

Amaral LAN, Scala A, Barthélémy M, Stanley HE (2000) Classes of small-world networks. Proc Natl Acad Sci USA (PNAS) 97(21):11149–11152

Anderson RM, May RM (1991) Infectious diseases of humans: dynamics and control. Oxford Science Publications

Andersson H (1998) Limit theorems for a random graph epidemic model. Ann Appl Probab 8(4):1331–1349

Athreya KB, Ney PE (1972) Branching processes. Springer Verlag

Bailey NTJ (1975) The mathematical theory of infectious diseases, 2nd edn. Charles Griffin & Company Ltd

Ball F, Mollison D, Scalia-Tomba G (1997) Epidemics in populations with two levels of mixing. Ann Appl Prob 7:46–89

Ball F, Neal PJ (2002) A general model for stochastic sir epidemics with two levels of mixing. Math Biosci 180(1–2):73–102

Barabási A-L, Albert R (1999) Emergence of scaling in random networks. Science 286:509–512

Bender EA, Canfield ER (1978) The asymptotic number of labeled graphs with given degree sequences. J Comb Theory A 24:296–307

Bollobás B (1985) Random graphs. Academic Press, New York

Brauer F , Watmough J (2009) Age of infection epidemic models with heterogeneous mixing. J Biol Dyn 3:324–330

Britton T, Kypraios, O'Neill PD (2011) Inference for epidemics with three levels of mixing: methodology and applicatio to a measles outbreak. Scand J Statist 38(3):578–599

Callaway DS, Newman MEJ, Strogatz SH, Watts D (2000) Network robustness and fragility: percolation on random graphs. Phys Rev Lett 85(25):5468–5471

Capasso V (2008) Mathematical structures of epidemic systems. Lecture notes in biomathematics. Springer

Cohen R, Erez K, ben Avraham D, Havlin S (2000) Resilience of the internet to random breakdowns. Phys Rev Lett 85(21):4626–4628

Hyman JM, Hethcote HW, DelValle SY, Eubank SG (2007) Mixing patterns between age groups in social networks. Soc Netw 29:539–554

Durrett R (2007) Random graph dynamics. Cambridge Series in Statistical and Probabilistic Mathematics

Erdős P, Rényi A (1960) On the evolution of random graphs. Publ Math Inst Hung Acad Sci 5:17–61

Faloutsos M, Faloutsos P, Faloutsos C (1999) On power-law relationships of the internet topology. Comput Commun Rev 29:251–263

Gilbert N (1997) A simulation of the structure of academic science. Sociol Res Online 2(2). http://www.socresonline.org.uk/2/2/3.html

Grimmett G, Stirzaker D (2001) Probability and random processes, 3rd edn. Oxford

Ivčhenko G (1973) On the asymptotic behaviour of degrees of vertices in a random graph. Theory Probab Appl 18:188–195

Janson S (2009) The probability that a random multigraph is simple. Comb Probab Comput 18(1–2):205–225

Keeling MJ (1999) The effects of local spatial structure on epidemiological invasions. Proc R Soc Lond B 266(1421):859–867

Liljeros F, Edling CR, Amaral LAN, Stanley HE, Åberg Y (2001) The web of human sexual contacts. Nature 411:907–908

Lotka AJ (1926) Thre frequency distribution of scientific productivity. J Wash Acad Sci 16(12):317–323

Łuczak T (1992) Sparse random graph with a given degree sequence. In: Fieze AM, Łuczak T (eds) Proceedings of the symposium of random graphs, Poznań 1989. New York. John Wiley, pp 165–182

McKay BD (1985) Asymptotics for symmetric 0–1 matrices with prescribed row sums. Ars Comb 19A:15–26

Meyers LA (2007) Contact network epidemiology: bond percolation applied to infectious disease prediction and control. Bull Am Math Soc 44:63–86

Meyers LA, Newman MEJ, Martin M, Schrang S (2003) Applying network theory epidemics: control measures for outbreaks of Mycoplama pneumoniae. Technical report, Santa Fe Institute

Molloy M, Reed B (1995) A critical point for random graphs with a given degree sequence. Random Struct Algorithms 6(2–3):161–180

Molloy M, Reed B (1998) The size of the giant component of a random graph with a given degree sequence. Comb Probab Comput 7:295–321

Newman MEJ (2002) Spread of epidemic disease on networks. Phys Rev E 66:16128

Newman MEJ, Strogatz SH, Watts DJ (2001) Random graphs with arbitrary degree distributions and their applications. Phys Rev E 64:26118

Pastor-Satorras R, Vespignani A (2001) Immunization of complex networks. Phys Rev 65:36104

Pastor-Satorras R, Vespignani A (2003) Chapter 5: Epidemics and immunization in scale-free networks. In: Bornholdt S, Schuster G (eds) Handbook of graphs and networks: from the genome to the internet. Heinz. Wiley-VCH

R Development Core Team (2007) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0

Watts DJ, Strogatz SH (1998) Collective dynamics of 'small-world' networks. Nature 393:440–442

Watts DJ (2002) A simple model of global cascades on random networks. Proc Natl Acad Sci 99(9):5766–5771

Yan P (2008) Chapter 10: Distribution theory, stochastic processes and infectious disease modelling. In: Brauer F, Van den Driessche P, Wu, J (eds) Lecture notes in mathematical epidemiology. Springer-Verlag