Spread of information and infection on finite random networks

Valerie Isham and Joanna Kaczmarska
Department of Statistical Science, University College London, Gower Street, London WC1E 6BT, United Kingdom

Maziar Nekovee[*]
Centre for Computational Science, University College London, Gordon Street, London WC1H 0AJ, United Kingdom

(Received 16 November 2010; published 28 April 2011)

The modeling of epidemic-like processes on random networks has received considerable attention in recent years. While these processes are inherently stochastic, most previous work has been focused on deterministic models that ignore important fluctuations that may persist even in the infinite network size limit. In a previous paper, for a class of epidemic and rumor processes, we derived approximate models for the full probability distribution of the final size of the epidemic, as opposed to only mean values. In this paper we examine via direct simulations the adequacy of the approximate model to describe stochastic epidemics and rumors on several random network topologies: homogeneous networks, Erdös-Rényi (ER) random graphs, Barabasi-Albert scale-free networks, and random geometric graphs. We find that the approximate model is reasonably accurate in predicting the probability of spread. However, the position of the threshold and the conditional mean of the final size for processes near the threshold are not well described by the approximate model even in the case of homogeneous networks. We attribute this failure to the presence of other structural properties beyond degree-degree correlations, and in particular clustering, which are present in any finite network but are not incorporated in the approximate model. In order to test this "hypothesis" we perform additional simulations on a set of ER random graphs where degree-degree correlations and clustering are separately and independently introduced using recently proposed algorithms from the literature. Our results show that even strong degree-degree correlations have only weak effects on the position of the threshold and the conditional mean of the final size. On the other hand, the introduction of clustering greatly affects both the position of the threshold and the conditional mean. Similar analysis for the Barabasi-Albert scale-free network confirms the significance of clustering on the dynamics of rumor spread. For this network, though, with its highly skewed degree distribution, the addition of positive correlation had a much stronger effect on the final size distribution than was found for the simple random graph.

DOI: 10.1103/PhysRevE.83.046128 PACS number(s): 89.75.Hc, 89.65.−s, 05.70.Jk, 05.10.−a

I. INTRODUCTION

The spread of infection is highly topical, with recent examples of epidemic or pandemic outbreaks including the "swine flu" outbreak of 2009, SARS in 2003, and HIV, which was first recognized in 1981. There are many earlier examples, including "black death" (plague) in the fourteenth century, and the 1918–19 influenza pandemic, in both of which huge numbers of deaths occurred. Epidemics in animals can have disastrous effects too, the UK 2001 foot and mouth epidemic being one such example. Mathematical models of the spread of infection have a major role to play in understanding the most important determinants of spread and investigating the effects of different control strategies without the need for empirical implementation ([1,2]). Many otherwise endemic infections are kept under control by the use of vaccination strategies, where the targets for the vaccinated proportion necessary to confer herd immunity are determined from epidemic models. Destructive viruses affect not only humans and other animals but also computers. On the other hand, viral marketing can be used to increase brand awareness and for fundraising purposes via social networks. Gossip algorithms are used to spread large amounts of information over the Internet ([3–5]). In such instances the aim is to maximize spread, rather than to limit it.

This paper is concerned with the effect of network structure on the spread of infection or information. The basic stochastic epidemic model for the spread of infection is the general stochastic epidemic [or Susceptible, Infected, Removed (SIR)] model; see, for example, Bailey [6] or Andersson and Britton [7]. In this model, individuals in a population of fixed size n are classified as being in one of three states: *susceptible* to infection, *infective* (that is, infected and infectious), and *removed*. We denote the numbers in these states at time t by $X(t), Y(t)$, and $Z(t)$, where $X(t) + Y(t) + Z(t) = n$. Under a homogeneously mixing assumption, new infections occur at a rate proportional to the product $X(t)Y(t)$, and these individuals are instantly infective. We may assume that each infective makes potentially infectious contacts at a rate β, choosing the contacted individual independently and uniformly over the population, so that the total rate at which infections occur is $\beta X(t)Y(t)/n$ (there is no loss of generality in allowing "self-contacts"). Each infective independently remains infectious for an identically distributed random time, assumed here to have an exponential distribution with mean $1/\delta$, so that infectives cease to be infectious and transfer to the removed state at a total rate $\delta Y(t)$. The removed individuals may be recovered and immune from reinfection, or perhaps quarantined and no longer transmitting infection. There are many generalizations of this basic epidemic model to allow for

[*] Also at Mobility Research Centre, BT, Polaris 134, Adastral Park, Martlesham, Suffolk IP5 3RE, UK.

population demography and age structure, stages of infection such as a latent period between infection and infectiousness, and population heterogeneity (see, for example, Ref. [7]), but these will not be discussed here.

The underlying model for the spread of information that we consider is a Maki-Thompson rumor model [8, Sec. 5.1], which has a very similar structure. Again there is a fixed population of size n, subdivided now into *ignorants, spreaders,* and *stiflers*, corresponding respectively to susceptibles, infectives, and removed individuals. As with the SIR model, a homogeneous mixing assumption means that ignorants become spreaders at a total rate $\beta X(t)Y(t)/n$, while spreaders spontaneously stop spreading the rumor and become stiflers independently at a per capita rate δ (this transition is sometimes called "forgetting," although the term is misleading in that such spreaders do not return to the ignorant state). In addition, though, the model includes an extra interaction in that, with probability p, a spreader contacting either another spreader or a stifler (at rate $\beta Y(t)[Y(t) + Z(t)]/n$) is stifled, thereby no longer spreading the rumor and themselves becoming a stifler. It follows immediately that the epidemic model is the special case of the rumor model in which $p = 0$. In both cases, the interest is in properties such as the probability that the infection or information will spread substantially, the final number infected or informed and, in particular, in the existence of any thresholds for such spread. Thresholds are especially relevant for control purposes, and theoretical results are available for large populations in the limit as $n \to \infty$.

These epidemic and rumor models assume that the population mixes homogeneously, an assumption that is unlikely to be realistic, even approximately, if n is at all large. One way of allowing for population structure, widely used in the context of epidemic models, is via the development of metapopulation models [9,10]. In such models the population is divided into groups (sometimes referred to as households), and individuals mix homogeneously within their own household, and again homogeneously but at a lower rate with individuals chosen at random from the whole population. In a recent paper [11], this model is generalized so that the households are located on the vertices of a network, with contacts only between connected neighbors. Interest then focuses on the limiting behavior of the epidemic when the household size stays fixed but the number of households goes to infinity. Alternatively, the population can be partitioned into classes with specified probabilities that an individual in a particular class will contact an individual in each other class, that individual then being chosen at random. This type of model is used particularly to allow for population heterogeneity [12], but the groups can also be used as a surrogate for spatial structure. For the latter, rather than having a fixed partition, it is more realistic to allow the neighborhoods of different population members to overlap [13].

In this paper, however, we will represent heterogeneous mixing by means of a network, with individuals able only to transmit information or infection to those to whom they are connected by edges of the network. This was investigated for small world networks by Ref. [14]. Here we will build on the earlier work of Isham *et al.* [15], using a range of random networks (to be described briefly in the next section) including homogeneous networks, simple random graphs, Barabasi-Albert (scale-free) networks, and geometric random graphs, to investigate the effect of different network structures, and to determine which network properties are most important in determining the spread of infection or information. In Sec. II, we describe the rumor and network models to be investigated. In Sec. III, we first investigate the adequacy of an approximate model as an approximation to the full stochastic rumor model. Then, in Sec. IV, specific results on the effects of network structure on the transmission of rumors are presented. Finally, a general discussion of the issues involved and broad conclusions are given in Sec. V.

II. RUMOR MODELS ON NETWORKS

A *homogeneous network* is one where all the nodes have the same degree, k, say (the number of nodes—termed *neighbors*—to which a node is directly connected by edges). Models for a homogeneously mixing population, such as the SIR and rumor models described above, can be thought of as taking place on a completely connected graph, that is, a homogeneous network of degree $n - 1$. In such cases, each node has the same total contact rate, denoted β above, with contacts to each neighboring node (chosen at random) taking place at rate β/n (since self-contacts were allowed there, without loss of generality). In this paper, where in general the networks are not homogeneous and the nodes have varying degrees, it is preferable to assume that each node contacts *each* neighboring node independently at rate λ, and self-contacts are not allowed. Thus, a node of degree k makes potentially infective contacts at a total rate λk, and higher-degree nodes will not only be more likely to become infected as they have more neighbors but also have the potential for greater transmission through a greater total rate of contact with their neighbors. This is often a realistic assumption, and, in particular, a degree-dependent rate of transmission will occur in simultaneous broadcasts of information.

In an earlier paper [15], we began an investigation of the stochastic spread of epidemics and rumors on networks by focusing on a stochastic approximation to the Maki-Thompson model that had recently been discussed, in a deterministic setting, by Nekovee *et al.* [16] and looked especially at the effects of network size and structure. This approximation takes into account the structure of the underlying network at the level of the degree-degree correlation function, but ignores the stochastic conditional dependence of the states of neighboring nodes given their degrees. Using embedded Markov chain techniques, a set of equations were derived for the final size of the epidemic or rumor on a *homogeneous* network that could be solved numerically. The resulting distribution was compared with the solution of the corresponding mean-field deterministic model as well as with the full Maki-Thompson model. Further investigation shows that it is possible to extend the embedded Markov chain approach to other networks. However, the enumeration of the set of equations and their boundary conditions rapidly becomes infeasible if the support of the degree distribution consists of more than a few values [17]. In addition, the time taken to solve these equations numerically increases very rapidly, in contrast to the speed with which large numbers of simulations of the full stochastic model can be generated with arbitrary degree distributions.

The advantage of having a numerically exact form for the final size distribution is that it is straightforward to determine the threshold at which the distribution changes shape from monotonically decreasing (unimodal, when the infection or information dies out rapidly from a small number of initial cases) to a bimodal form that indicates that, with nonzero probability, a substantial spread will occur. For the deterministic rumor model on a homogeneous network, the condition for rumor spread is that $R_0 = (1 + p)/(\psi + p) > 1$, where $\psi = \delta/(\lambda k)$, that is, that $\psi < 1$ [16]. Note that, when $\delta = 0$, such spread always occurs; stifling alone is never sufficient to control the outbreak. Correspondingly, in the asymptotic case, for the stochastic approximation model on a homogeneous network, it is necessary that $\delta \geqslant \lambda k$ to control the spread of the rumor, regardless of the value of p. Isham *et al.* [15] showed that the threshold value of ψ is lower than that for the deterministic model; i.e., less "forgetting" is needed to control the outbreak if the other parameters are fixed, as the stochastic fluctuations increase the chance of an extinction. In addition, the difference between the two thresholds decreases with the network size n, apparently following a $n^{-1/3}$ behavior. It was also shown that fluctuations in the final size of the epidemic are retained as the network size increases so that, even in a limiting infinite size case, the deterministic model greatly overestimates the mean of the final size of the epidemic. Isham *et al.* [15] then compared the thresholds (obtained by Monte Carlo simulation) for the full stochastic model on a homogeneous network, including density correlations at neighboring nodes, with those for the approximating stochastic model and showed that the latter can reproduce the exact simulation results with great accuracy. Finally, further Monte Carlo simulations of the full stochastic model were used in a preliminary exploration of the effects of network size and structure on the final size distribution.

In this paper we present further results to investigate the parts of the parameter space in which the stochastic approximation model provides a useful alternative to the exact rumor model on a homogeneous network. We also give a much more thorough discussion of the effects of network structure on the final size distribution. We begin by giving brief definitions of the rumor models and networks to be considered.

A. The random network models

One of the simplest random graphs is the so-called simple (Erdös-Rényi) random graph [18], in which each pair of the n nodes is independently connected by an edge with probability π, so that the degree of each node has a binomial distribution with index $n - 1$ and probability π; the degree of a node is the number of nodes to which it is connected by an edge (i.e., its number of *neighbors*). For large n, the degree distribution is approximately Poisson, with mean $n\pi$. With this construction, the degrees of distinct nodes are dependent but are asymptotically independent. For a general degree distribution, the *Molloy-Reed algorithm* ([19,20]) can be used to construct an uncorrelated graph. The idea is that n independent degrees are first generated, specifying the number of "arms" emanating from the nodes, and then these arms are joined at random to form edges (resampling may be necessary to ensure that the total number of arms is even), and multiple edges are not allowed. For such graphs, the degrees are uncorrelated; for neighboring nodes the correlation is zero once the common edge is removed.

Various degree distributions are of interest, in addition to the Poisson distribution. If an overdispersed distribution (in which the variance exceeds the mean) is required, the negative binomial distribution (which can be thought of as a Poisson distribution in which the mean is itself a gamma random variable) is an obvious possibility. The degree distributions of empirical networks are often shown to have a power-law tail and may have moments that are formally infinite. Such distributions are "scale-free" and are characterized by the presence of "hubs," where a few nodes have very high degrees while most nodes have relatively low degrees. One such example is the Yule-Simon distribution [21] in which the probability of a degree d has the form $\rho B(d,\rho + 1) \propto d^{-(\rho+1)}$ as $d \to \infty,(d = 1,2,\ldots,\rho > 0)$ where B is the beta function. This distribution was originally put forward as the equilibrium distribution of a preferential attachment process in the context of a biological application, and Simon [21] gives an example of a model for writing a book where words are added one at a time. At each step, with a fixed probability a new word is added, while with the complementary probability, the word is sampled from the existing words in the book in proportion to their current frequencies. The Molloy-Reed algorithm can then be used to generate an uncorrelated graph with this degree distribution.

Another construction for network growth that gives a scale-free degree distribution is due to Barabasi and Albert [22]. There are several variants of the algorithm. The version used in this paper starts with a small number of unconnected nodes and then adds additional nodes one at a time. These nodes each come with a fixed number of edges, m say, that are attached randomly to the existing nodes with probabilities that are in proportion to the current degrees of those nodes. Multiple edges are not permitted, and we start with m unconnected nodes at outset to ensure the network is connected. The resulting graph is asymptotically uncorrelated, and the density of the degree distribution behaves approximately as d^{-3} as $d \to \infty$, although this can be modified by generalizing the algorithm.

The final random graph that we consider in this paper is the random geometric graph [23]. The starting point for this construction is a spatial homogeneous Poisson process. The neighbors of each node are defined to be those within a fixed critical distance, and neighboring nodes are connected by edges to give the resulting graph, in which the original spatial distances are not retained. The properties of the random geometric graph are closely linked to those of the spatial Poisson process. As well as having a Poisson degree distribution, it is a matter of simple geometry to show that the correlation is $1 - 3\sqrt{3}/(4\pi) \simeq 0.59$. Graphs with a high degree of positive correlation between neighboring nodes appear highly clustered, and the words are often used interchangeably. However, the cluster coefficient measures the extent to which the graph is triangulated and is the proportion of connected triples of nodes that form a triangle. It follows from the conditional independence property of the Poisson process that, for a random geometric graph, the cluster coefficient and the degree correlation are identical, both being equal to the average (over r) area of overlap of two unit discs whose centers are a distance r apart ($0 < r < 1$).

B. The stochastic rumor models

The stochastic rumor model considered in this paper is an extension of the Maki-Thomson model described in Sec. I for a homogeneous network. The properties of the rumor dynamics on the network will be taken conditionally on the realization of the network, which is generated from one of the random mechanisms described above and then held fixed. At time t, each node is in one of three states: ignorant, spreader, or stifler. Each spreader contacts each of its neighbors independently in a Poisson process of rate λ, so that λ represents the mean number of contacts per neighbor per unit time. At a contact, two possible transitions may occur: If the node contacted is an ignorant, then that node itself becomes a spreader; if the contacted node is another spreader or a stifler then, with probability p, the node initiating the contact becomes a stifler; otherwise no transition occurs. In addition, spreaders may become stiflers spontaneously at a rate δ. If $p = 0$, the model corresponds to an SIR epidemic model on the network. At time $t = 0$, all nodes are initially ignorant, and a single node is chosen at random to be a spreader.

The approximate model [15] is a Markov model in which the state is given by the instantaneous numbers of ignorant, spreader, and stifler nodes of degree k, for each possible k. Thus, suppose that, at time t, there are n_k nodes of degree k and that of these X_k, Y_k, and Z_k are, respectively, ignorants, spreaders and stiflers (where the dependence on time is omitted and $Z_k = n_k - X_k - Y_k$). We denote the network degree-degree correlation function by

$$p_{jk} = \text{Prob(neighbor node has degree } k \mid \text{index node has degree } j).$$

For example, in an uncorrelated graph, $p_{jk} \propto kp_k$, where $\{p_k\}$ is the marginal degree distribution. Nekovee *et al.* [16] and Isham *et al.* [15] discuss an analysis of this model in which the influence of the network structure is wholly encapsulated by the p_{jk} matrix. Specifically, the total rate of transitions from (X_k,Y_k) to $(X_k - 1, Y_k + 1)$ is approximated by

$$\lambda k X_k \sum_j p_{kj} Y_j / n_j. \tag{1}$$

Effectively, the probability that, at time t, a node is a spreader given that it has degree j and is the neighbor of an ignorant node of degree k, is approximated by the unconditional probability that the degree j node is a spreader. Thus the dependence of the rumor status of the degree j node on that of its neighbor is ignored. A similar approximation is made with regard to stifling so that the total rate of transitions from (X_k,Y_k) to $(X_k,Y_k - 1)$ is

$$Y_k \left[\delta + \lambda pk \sum_j p_{kj} (n_j - X_j)/n_j \right]. \tag{2}$$

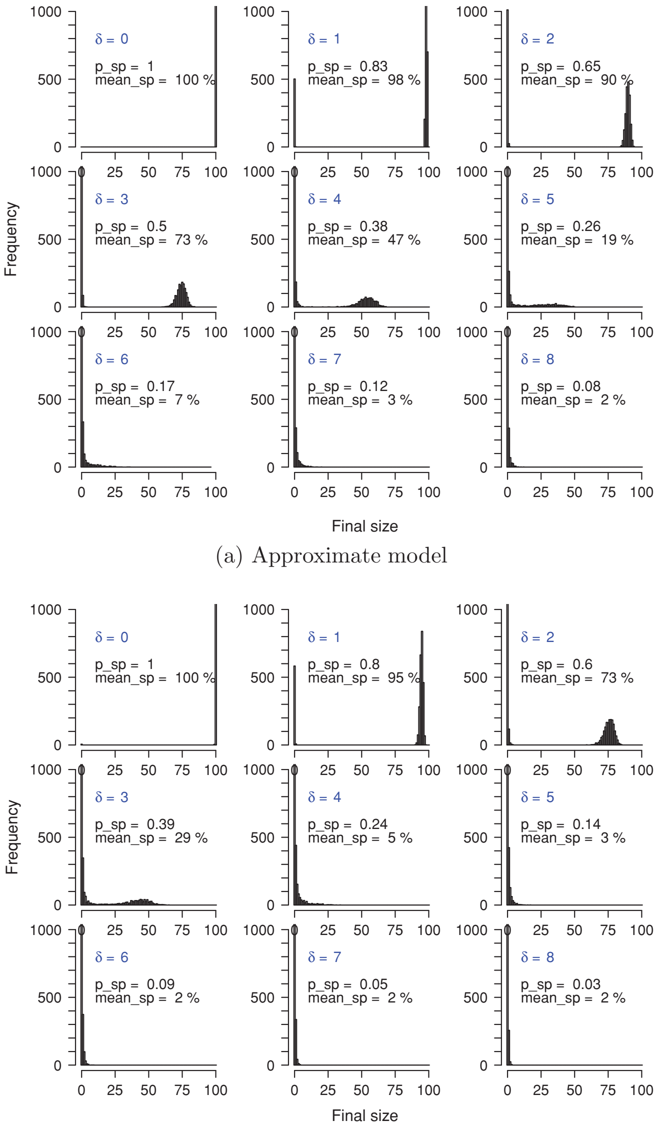
In the next section, we will discuss the adequacy of this approximate model to the full rumor model, before going on in the following section to investigate the effects of population size and network structure on the properties of the full model. Thresholds for rumor spread and the final size distribution will be of particular interest.

III. COMPARISON OF THE FULL AND APPROXIMATE RUMOR MODELS

Earlier papers [15,24] showed a good correspondence between the final size distribution for the approximate model and that for the full stochastic model. However, only rumors spreading on homogeneous networks were considered, and only limited parameter sets were used. These corresponded to widespread transmission of the rumor, and the approximation is expected to work well in this case. In the approximate model, for example, an ignorant node can be infected even if all its neighbors are stiflers, but this scenario is unlikely to arise if the contact rate λ is sufficiently high. If the dynamics are around the threshold level, the approximation is likely to have a much more dramatic effect. Similarly, the network structure is highly relevant. Ignorant nodes with many neighbors are unlikely to get in the position of being isolated from infection by stifler neighbors, whereas random effects may easily result in the isolation of those with only one or two neighbors. For all network types, the approximate model is a better approximation to the full model as the mean degree increases.

In this section we discuss some results from a simulation study used to explore the differences between the approximate and full models, and therefore concentrate on effects around the threshold level. In each case, we consider a network of 1000 nodes, as being large enough to illustrate the main effects whilst still keeping the computing time reasonable. The four network structures considered are the homogeneous graph, the simple random graph, the Barabási-Albert graph, and the random geometric graph. The homogeneous, simple, and random geometric graphs were created using the algorithms of the *R*-package *Igraph* [25], while the Barabási-Albert graphs were created using the algorithm described in Sec. II A. The first three structures are approximately uncorrelated, whereas the random geometric network is correlated. For the results reported here, the mean of the degree distribution for each network is fixed at 6, the contact rate is $\lambda = 1$, and the stifling probability is $p = 0.1$. The "forgetting" parameter δ ranges from 0 to 8, so that the parameter $\psi = \delta/(\lambda k)$, where k now denotes the *mean* degree of the nodes, varies from 0 to 4/3 (where, for the deterministic approximation to the model, $\psi < 1$ is required for rumor spread). The final size distributions are averaged first over 1000 simulations of the rumor dynamics on a fixed network, with the initial spreader chosen independently and at random for each simulation, and then over the distributions obtained by using the three independent simulations of the random network (i.e., 3000 simulations in all). Averaging over the network provides some protection against the selection of an atypical network configuration.

In Fig. 1, results are shown for a homogeneous random graph with $k = 6$. Note that although frequencies of up to 3000 are possible, the scale is capped at 1000 to show the patterns of the smaller frequencies more clearly. As can be seen, for both the full and approximate stochastic models,



(a) Approximate model

(b) Full model

FIG. 1. (Color online) Comparison of final size distributions as percentages of the total population for the homogeneous network using the approximate and full models on a network with 1000 nodes, $\lambda = 1$, $k = 6$, $p = 0.1$; p_sp is the probability of spreading to above 1% of the population; mean_sp is the conditional mean final size, given that the rumor has spread to above 1% of the population.

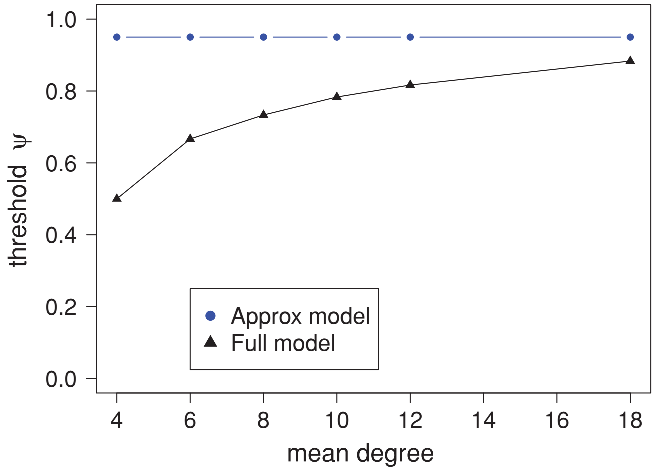


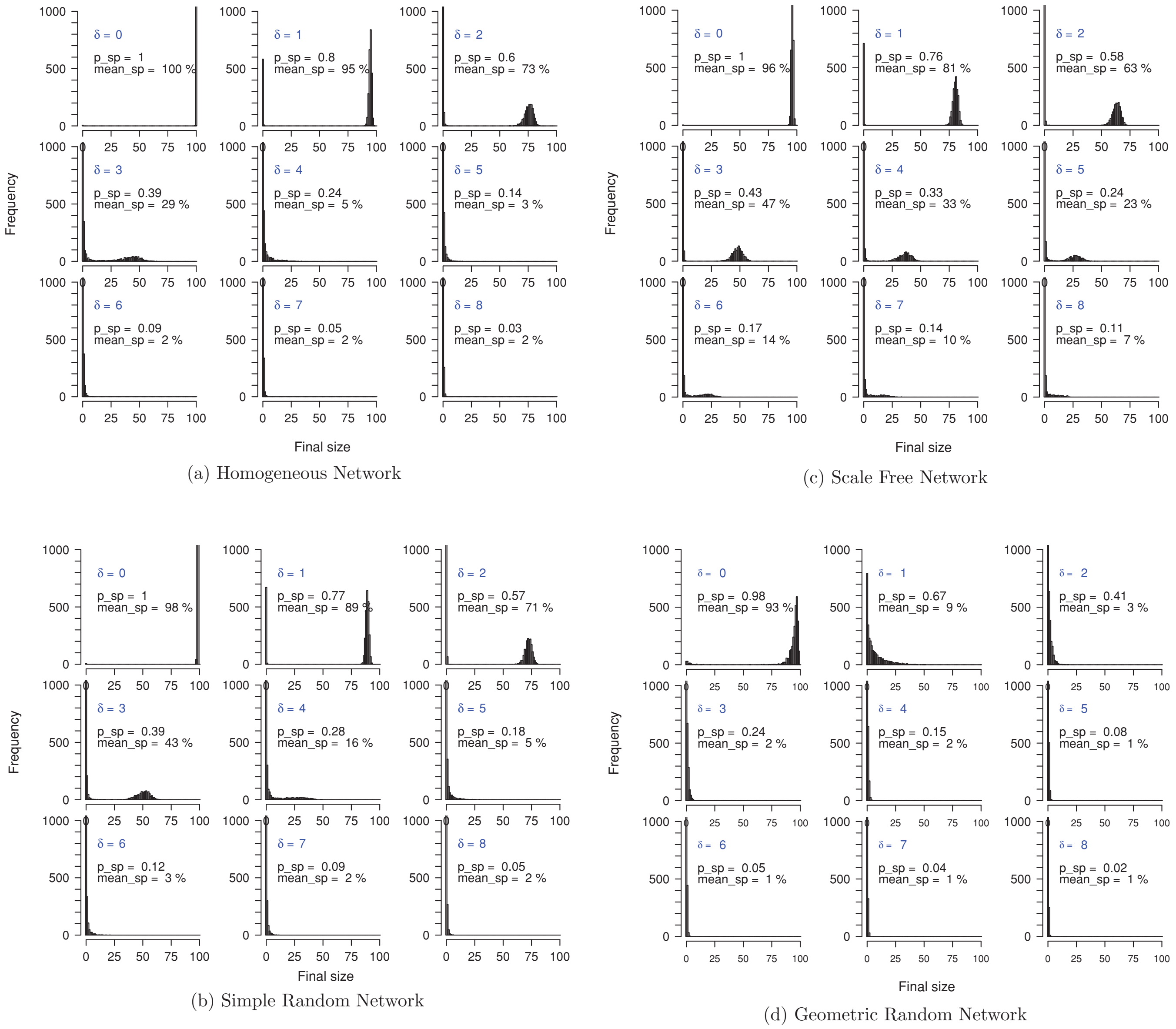
FIG. 2. (Color online) Thresholds for the homogeneous network with λk fixed at 6, $p = 0.1$.

the threshold values of ψ between unimodal and bimodal forms for the final size distribution are lower than for the corresponding deterministic model (for which $\psi = 1$). This is because stochastic effects make it more likely that the rumor will die away soon after its introduction, and therefore less "forgetting" is required to control its spread. As described above, the approximate model allows spreading to isolated ignorants whereas the full model does not, and its threshold value of ψ (seen to lie in the interval $(\frac{5}{6},1)$) is therefore much closer to the deterministic critical value $\psi = 1$ and substantially overestimates the value for the full model [which can be seen to lie in $(\frac{3}{6},\frac{4}{6})$]. These ranges for the threshold values of ψ can be narrowed by using a much more extensive set of simulations, but the graphs are intended to show the qualitative picture, which is clear. In Fig. 2 for comparison, we show the threshold value of ψ for a homogeneous network as a function of the degree (k), where λ is adjusted so that the total contact rate per node is kept fixed ($\lambda k = 6$). Here we have determined more exact threshold levels by running a further set of simulations for each model, with δ varying within the appropriate narrower range identified in the initial set. The threshold value of ψ for the approximate model is a constant in Fig. 2 as it depends only on the product λk, while the threshold for the full model approaches this value as k increases. A mean degree of 6 has been chosen for most of the comparisons as an interesting compromise. Note that the error in the approximate model is primarily an effect of network connectivity and remains an issue for larger network sizes.

Results for the other three network structures show that the threshold value of ψ for the approximate model consistently overestimates that for the full model. For the simple random graph the corresponding ranges of values for ψ are $(\frac{4}{6},\ \frac{5}{6})$ for the full model and $(1,\ \frac{7}{6})$ for the approximate model; for the Barabási-Albert network, the approximate model has a threshold of around $\psi = \frac{11}{6}$, while that for the full model is about $\frac{8}{6}$; for the random geometric graph the corresponding threshold is close to $\frac{1}{6}$ for the full model, with a range of $(1,\ \frac{7}{6})$ for the approximate model. The final size distributions for these networks for the full model are shown in Fig. 3.

It is interesting to consider separately the probability of spread to above 1% of the population, and the conditional mean final size given this level of spread (denoted as p_sp and mean_sp, respectively, in the histograms). In Fig. 4 we have plotted these for all four network types, showing that the former is fairly accurately represented by the approximate model, whereas the latter is less well approximated, especially for the random geometric network.

As can be seen from the figures, the approximation for the scale-free Barabási-Albert network is relatively good. This is because the presence of the hubs in the network drives the



(a) Homogeneous Network

(c) Scale Free Network

(b) Simple Random Network

(d) Geometric Random Network

FIG. 3. (Color online) Final size distributions as a percentage of total population for different network structures with 1000 nodes, $\lambda = 1$, $k = 6$, $p = 0.1$; p_sp is the probability of spreading to above 1% of the population; mean_sp is the conditional mean final size, given that the rumor has spread to above 1% of the population.

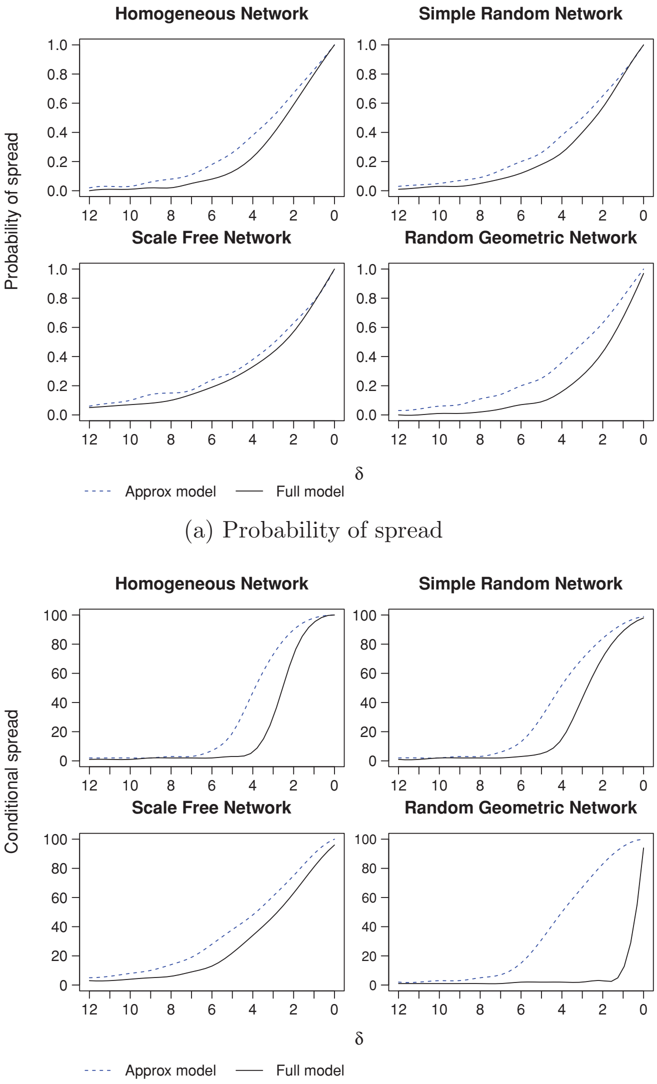
dynamics, with little scope for ignorant hubs to get isolated. On the other hand, for the random geometric graph, with its structure of weakly connected groups, it is relatively easy for ignorant nodes to become permanently separated from spreaders, and the approximate model does not allow for this.

The threshold point for each of the models can roughly be identified from the plot of the conditional spread against δ, as the point at which the curve flattens out. This could be made clearer by increasing the network size, which would sharpen the point at which the slope changes from positive to zero (with the exception of the scale-free network, where the nature of the structure means that the threshold increases as the network size grows; see Sec. IV).

IV. THE EFFECTS OF NETWORK STRUCTURE ON RUMOR DYNAMICS

A. Different network types

It is often thought that simple random graphs are similar to uncorrelated homogeneous networks, and, for large networks with a fairly high mean degree, the variability of the binomial degree distribution is relatively unimportant. However, for a small mean degree, this variability has noticeable effects. For a mean degree below about 6, a simple random graph will often not be connected. Even for connected graphs, the presence of nodes with very low degrees makes the rumor less likely to spread than for the corresponding homogeneous network. On the other hand, the presence of other nodes with relatively high



(a) Probability of spread

(b) Conditional spread as a percentage of the total population

FIG. 4. (Color online) A comparison of the full and the approximate models for different network structures; all have 1000 nodes; $\lambda k = 6$, $p = 0.1$.

degrees means that, if the rumor does spread initially, then it is more likely to spread extensively.

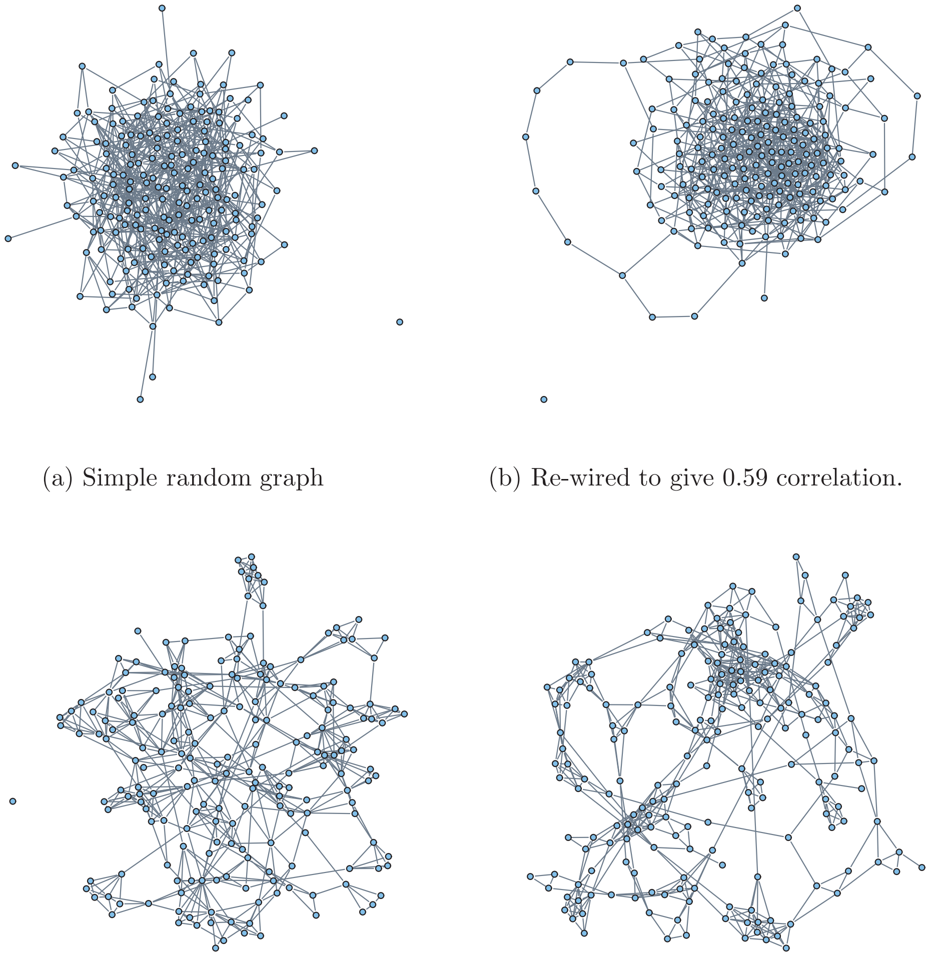
The rumor dynamics of the Barabási-Albert network are very different, particularly for large networks where the truncation effect of the finite number of nodes on the power-law tail of the degree distribution still permits the presence of substantial hubs in the network. The network is characterized by a number of very well-connected nodes, with the majority of nodes having low degrees. The effect is that the threshold in ψ is much higher than for the homogeneous network and simple random graph, with much more forgetting needed to overcome the effect of the hubs. For example, with $n = 1000$, a mean degree of 6, and other parameter values as described in Sec. III, simulations show that the threshold is approximately 1.3 as compared with values under 1 for the homogeneous network and simple random graph. The difference would be much greater for larger networks. Using simulations for networks with $n = 1000$ and various mean degrees (and correspondingly adjusted values of λ so that the mean contact rate, λk is fixed), we have found that for low values of ψ, broadly those below the threshold for the homogeneous network, given that it spreads at all, the rumor reaches a smaller proportion of the population for the the Barabási-Albert network than the other two structures. Intuitively, in this case the network has relatively many low-degree nodes, some of which will not hear the rumor before the better-connected nodes become stiflers.

The deterministic analysis of Nekovee *et al.* [16] shows that, for uncorrelated networks, to a first approximation the threshold for ψ is $1 + c_K^2$, where c_K is the coefficient of variation of the degree distribution. Thus, it is to be expected that the thresholds for the Barabási-Albert network in these comparisons will be much higher than those for the simple random graph, which in turn will be higher than those for the homogeneous network. The random geometric graph has a Poisson degree distribution as has, to a good approximation, the simple random graph once n is sufficiently large. Nevertheless, the former is highly correlated, the network structures are very different, and, as can be seen from Fig. 3, the rumor dynamics on the two graphs are very different. For the parameter set considered in Fig. 3, the simple random graph has a threshold in ψ somewhere between $\frac{4}{6}$ and $\frac{5}{6}$, while that for the random geometric graph is less than $\frac{1}{6}$. (As discussed above, however, the approximate models, which do not properly take account of the dependence of the node states on their connectivity, give very similar results for the two networks.)

This result raises the question of what properties of the network are driving this difference in dynamics. It is often assumed to be a result of the high correlation (0.59) of the node degrees of the random geometric graph, but the graph also has a high cluster coefficient (proportion of connected triples that are triangles) and a much greater mean geodesic distance (the average path length between connected pairs of nodes). For example, in the network simulations underlying Fig. 3, the mean geodesic distance for the homogeneous network was 4.2, for the simple random graph 4.0, for the Barabási-Albert network 3.5, and for the random geometric graph 16.0. This provides a numerical quantification of the underlying much less interconnected structure of the latter. As a first step in investigating this question, in the next section we will look at the impact of separately adding positive correlation and clustering to the simple random graph.

B. Impact of adding positive correlation and clustering to the simple random graph

We start with our 1000 node simple random graph, with its negligible correlation and clustering, and use network "rewiring" algorithms, first to increase the correlation while keeping the clustering coefficient unchanged, and second to increase the clustering coefficient while keeping the correlation unchanged. In both cases, the node degree distribution remains constant. This allows us to isolate the impact of two of the key differences between the simple random graph and the random geometric graph.



(a) Simple random graph

(b) Re-wired to give 0.59 correlation.

(c) Re-wired to give 0.59 clustering coefficient

(d) Random geometric graph

FIG. 5. (Color online) Illustration of the effect of correlation and clustering on network structure.

The algorithm to introduce positive correlation is due to Xulvi-Brunet and Sokolov [26]. Essentially, the idea is as follows. The starting point is an uncorrelated network with a particular degree distribution; here we start with a simple random graph, with its Poisson degree distribution. At each step of the algorithm, two edges (with four corresponding nodes) of the network are chosen at random. With probability α, the four nodes are rewired by deleting the two edges and joining the two nodes with the highest degrees, and the two nodes with the lowest degrees. Otherwise the four nodes are rewired at random. If one or both new edges already exists, the step is discarded. The larger α the larger the (limiting) positive correlation that can be obtained by this algorithm. To obtain negative correlations, the nodes with the highest and lowest degrees must be joined, but we will not follow this option here. For our simple random graph with 1000 nodes and mean degree 6, this algorithm has no significant effect on the clustering coefficient, although this is not necessarily the case for all network structures.

The algorithm to increase the clustering coefficient, due to Bansal *et al.* [27], works as follows. Again, the starting point is a random network with a given degree distribution.

TABLE I. Impact on threshold of adding correlation and clustering.

Network	Correlation coefficient	Clustering coefficient	Mean geodesic distance	Range for ψ threshold
Simple random (ER)	0.02	0.00	4.1	$(\frac{4}{6},\frac{5}{6})$
ER with added correlation	0.59	0.01	4.3	$(\frac{5}{6},1)$
ER with added correlation	0.94	0.04	6.1	$(\frac{5}{6},1)$
ER with added clustering	0.04	0.58	7.7	$(\frac{1}{6},\frac{2}{6})$
Geometric	0.58	0.59	16.0	$(0,\frac{1}{6})$

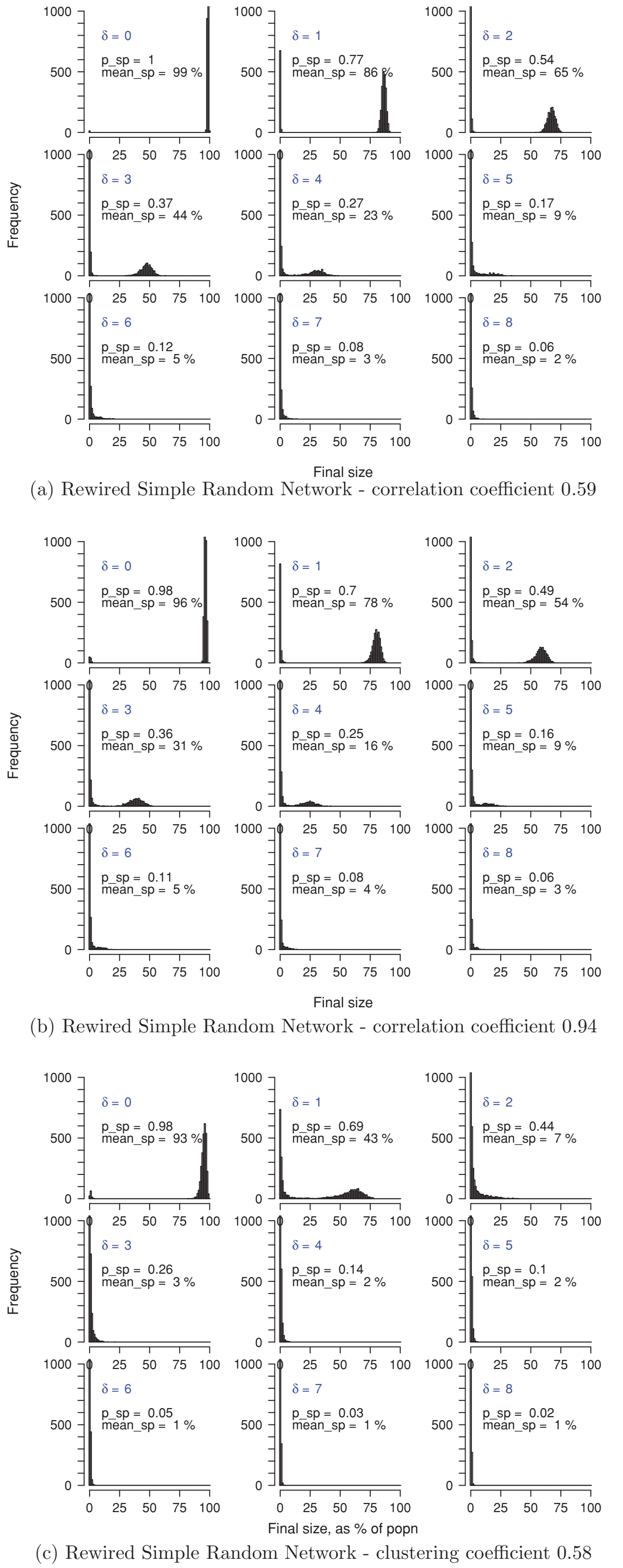
At each step we pick a random node x with degree greater than 1, and randomly select two of its neighbors, y_1 and y_2 (also with degree greater than 1). We then randomly select a neighbor z_1 of y_1, and z_2 of y_2, and rewire the nodes by deleting the edges (y_1,z_1) and (y_2,z_2), and adding new edges (y_1,y_2) and (z_1,z_2) (assuming these do not exist already), which creates the triangle (x,y_1,y_2). If the step increases the clustering coefficient (which depends on the other edges of the selected nodes), it is retained; otherwise it is discarded. The algorithm is continued until the desired level of clustering is achieved, or a maximal level is reached, if this is sooner.

Figure 5 illustrates the effect on the network structure of adding correlation and clustering to a simple random graph with mean degree 6 and 200 nodes. The Xulvi-Brunet and Sokolov algorithm is used to add correlation (up to the level, 0.59, of the random geometric graph). Restarting from the simple random graph, we then use the Bansal algorithm to add clustering (again up to 0.59, the level of the random geometric graph). The graphs are plotted using the *R*-package *Igraph* [25]. The random geometric graph with the same number of nodes and mean degree can be seen to include both the triangles of Fig. 5(c), and the more linear components of Fig. 5(b).

The impact of adding correlation and clustering on the final size distribution is shown in Table I, where the starting point is a simple random graph with $n = 1000$ nodes, $\lambda = 1$, $p = 0.1$, and the mean degree is 6. As before, three independent simulations of each network were used, so that in total we had 3000 simulations. The table gives the values for the node degree correlation (0.02), cluster coefficient (0), mean geodesic distance (4.1), and threshold region for ψ $(\frac{4}{6}, \frac{5}{6})$. The Xulvi-Brunet and Sokolov algorithm was then used with $\alpha = 0.8$ to give a correlation of 0.59, at which point the sample cluster coefficient remains almost zero, and the mean geodesic distance has increased slightly to 4.3. Rewiring with $\alpha = 1$ to give the maximal correlation further increases the mean geodesic distance to 6.1, with the clustering coefficient still close to zero. Figure 6 shows the final size distributions for the two networks with added correlations.

Although adding correlation can be seen to reduce the spread for low values of ψ, it increases it for higher values and ultimately slightly increases the threshold at which the final size distribution becomes unimodal. Intuitively, the reasoning is that some of the low-degree nodes that are now connected in a much more linear structure are likely not to hear the rumor even when the level of forgetting is low. Conversely, at the higher levels of forgetting, the dominant impact is that, if the rumor reaches a highly connected node, it is more likely to spread, due to the connections to other high-degree nodes.

Increasing the clustering coefficient to 0.58 (the maximum achievable for our network) using the Bansal algorithm can be seen to have a much more significant impact, with an increase in geodesic distance to 7.7, and a reduction in the threshold to $(\frac{1}{6},\frac{2}{6})$ (see also Fig. 6). These results are to be compared with those for a random geometric graph, where the sample correlation was 0.58, the clustering coefficient was 0.59, and the mean geodesic distance was 16. For this graph, there is substantial rumor spread only for ψ close to zero. Thus we see that the much lower spread observed for the random geometric graph compared to the simple random graph, which has a



(a) Rewired Simple Random Network - correlation coefficient 0.59

(b) Rewired Simple Random Network - correlation coefficient 0.94

(c) Rewired Simple Random Network - clustering coefficient 0.58

FIG. 6. (Color online) Final size distributions as a percentage of total population for the ER network with added correlation or clustering; 1000 nodes, $\lambda = 1$, $k = 6$, $p = 0.1$; p_sp is the probability of spreading to above 1% of the population; mean_sp is the conditional mean final size, given that the rumor has spread to above 1% of the population.

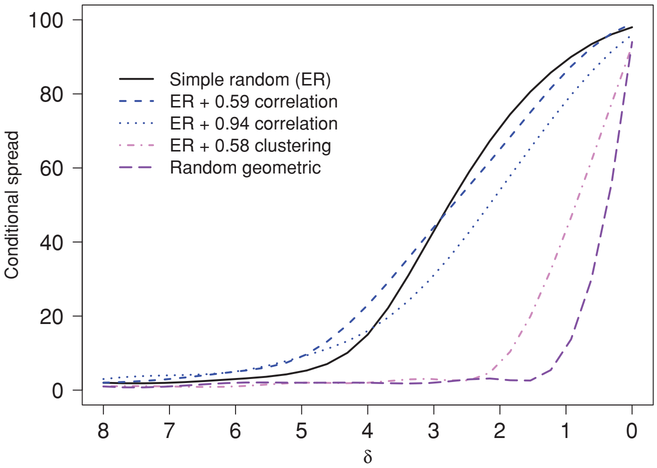


FIG. 7. (Color online) The impact of correlation and clustering on the conditional spread (expressed as a percentage of the total population) for networks with a Poisson degree distribution; networks have 1000 nodes; $\lambda k = 6$, $p = 0.1$.

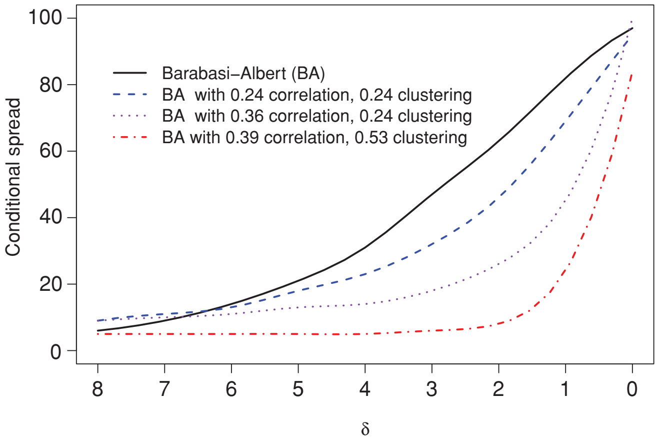


FIG. 8. (Color online) The impact of correlation and clustering on the conditional spread (expressed as a percentage of the total population) for scale-free networks; networks have 1000 nodes; $\lambda k = 6$, $p = 0.1$.

common degree distribution, is mainly due to clustering, with correlations having only a weak effect. The relative impacts can be seen clearly in Fig. 7 where we have plotted the conditional mean given the rumor spreads to above 1% of the population for the simple random network, the variants with added correlation or clustering, and the random geometric network. Even the maximal degree-degree correlation of 0.94 is seen to have a fairly limited impact, whereas the addition of clustering significantly reduces the spread of the rumor.

Analysis of the effect of introducing positive correlation and clustering to the simple random network has offered insight into the differences in rumor dynamics of this network compared with the random geometric graph. The simple random graph was also particularly suitable for this kind of analysis, since these features could readily be added independently of each other. However, social networks are mainly scale free, and rumor models are of particular interest on social networks. In the next section therefore we extend our analysis to the Barabási-Albert scale-free network.

C. Impact of adding positive correlation and clustering to the Barabási-Albert scale-free network

Again we start with a 1000-node network and apply the algorithms described in the section above. For a network of this type, size, and mean degree, the algorithm for adding positive correlation was found also to increase the clustering coefficient. Similarly, the algorithm for increasing the clustering coefficient also increased the correlation. Thus it is not possible entirely to separate the two. Nevertheless, by targeting different levels of each coefficient, we gained useful insights for this network. Table II gives a summary of the networks in our comparison, showing the algorithm that created each, the correlation and clustering coefficients, and the mean geodesic distance. Figure 8 plots the conditional mean given the rumor spreads to above 1% of the population for these networks. We start with a Barabási-Albert scale-free network, with 1000 nodes, and the parameters $\lambda = 1, p = 0.1$ as before. As before, three independent simulations of each network were used. Before rewiring, our network has a slightly negative correlation coefficient, negligible clustering and a mean geodesic distance of 3.5. First we used the Xulvi-Brunet and Sokolov algorithm to add maximal correlation, which was limited by the size of the network and the highly skewed degree distribution, giving an average of 0.36 across our three network simulations. This rewired network was found to have an average clustering coefficient of 0.24, so we then targeted a clustering coefficient of 0.24 using the Bansal algorithm, at which point the correlation coefficient was also at 0.24. A comparison of the behavior of the rumor on these two rewired networks gives an indication of the impact of correlation (since they have the same clustering coefficient). No thresholds are given, since we found that there was a second mode even when the conditional spread was very low. However, in contrast to the results for the simple random graph, here

TABLE II. Summary of BA network variants investigated.

Network	Correlation coefficient	Clustering coefficient	Mean geodesic distance
Barabási-Albert (BA)	−0.07	0.02	3.5
BA rewired to add clustering	0.24	0.24	4.3
BA rewired to add correlation	0.36	0.24	7.6
BA rewired to add further clustering	0.39	0.53	8.2

increasing correlation reduces both the probability of spread and the conditional spread of the rumor significantly. This is in line with intuition, given the degree distribution. In order to increase the correlation coefficient, the network is rewired such that the low-degree nodes tend to be connected to each other, rather than to the hubs. This makes them more likely to become isolated. The bimodal nature of the final size distribution, even at high levels of forgetting is due to the fact that if the rumor starts in one of the well-connected nodes, it will tend to spread within its (also well-connected) neighbors before dying out.

In the final network, the maximal level of clustering using the Bansal algorithm is added. This gave a clustering level of 0.53, with a correlation coefficient of 0.36. Thus we can compare two networks with broadly similar levels of correlation to see the impact of adding clustering. As for the simple random graph, this is shown to have a highly significant effect in reducing the spread of the rumor, particularly at the lower levels of forgetting.

V. DISCUSSION AND CONCLUSIONS

In this paper we have investigated the spreading of rumors on networks. We examined the approximate stochastic model, which has been used by a number of earlier authors, particularly in the physics community. This model does not require the actual generation of a network, but uses equations involving just the degree distribution and degree correlation matrix. Although other authors had found that the approximate model gave very close results to the full model for the mean final size, the examples on which such conclusions were based were limited to homogeneous networks with parameter sets for which spreading was widespread, i.e., those with ψ far below the threshold. In order to get a more complete picture, we undertook a simulation study for a variety of networks and parameters to investigate the situations in which the approximate model can be considered to provide a good solution to the final size distribution, and where it is less appropriate.

We found that the approximate model is reasonably accurate at predicting the probability of spread for all the network types. However, it performs poorly in identifying the position of the threshold and the conditional mean final size for networks that are not well connected. This includes networks with low mean degree of all types. More interestingly, it includes the random geometric network, with its weakly connected groups of nodes and high geodesic distances. The performance of the approximate model improves with increasing mean degree, and results are close for the small-world type networks with mean degree of 12 or more. However, although improvement is also seen for the random geometric network, the quality of the approximation remains very poor even at this level.

As well as comparing it against the approximate model, we also used the full stochastic model, where the rumor spreads along the edges of a network, to investigate how the spreading process is affected by the nature of the underlying network. We found the following key results:

(i) For the same level of λk, the rumor will spread to a greater percentage of the population for networks with higher k (i.e., higher mean number of edges), assuming the same network structure, and that all other parameters remain the same.

(ii) Scale-free networks require a greater level of forgetting in order to control a rumor than exponential-type networks (such as the homogeneous or Erdös-Rényi) given the same spreading rate and mean degree. However, for low levels of the forgetting parameter δ, the rumor is spread to a lower proportion of the population with the scale-free network.

(iii) The simple random and geometric random networks share a common (Poisson) degree distribution, and yet exhibit very different behavior in terms of rumor spread. The former is uncorrelated and has no clustering, whereas the latter has correlation and clustering coefficients of 0.59. In order to isolate the effects of these two features, we used edge rewiring algorithms to introduce each separately to the simple random network. Our results showed that it was the introduction of clustering, rather than correlation, that significantly reduced the threshold. Interestingly, the spreading behavior on the network with added clustering was very similar to that of the random geometric network, despite the fact that its mean geodesic distance was still much lower. Intuitively, this is because it is the "local" properties such as clustering and mean number of neighbors that affect the spread of the rumor, rather than the more "global" properties, such as geodesic distances.

(iv) Introducing clustering to the Barabási-Albert scale-free network was similarly found to reduce the spread of a rumor. However, for this network, adding positive correlation also reduces the spread, with a bigger impact than for the simple random graph. This is due to the highly skewed degree distribution, which has a large number of low-degree nodes. If the network is rewired in order to increase correlation, these low-degree nodes become much less likely to hear the rumor before it is stifled.

In this paper we have focused on the final size distribution. However, in some cases, for example in a viral marketing campaign, we may be more interested in achieving a certain level of spread within a given time. Future work is required to gain an understanding of how the various network types perform in terms of the initial speed of spreading and the overall time pattern.

Here we have isolated the effects of correlation and clustering. Since we have used specific rewiring algorithms, in principle it is possible that the effects that we have observed are due in part to other features introduced into the networks. Further analysis to identify the components that uniquely define a network would clarify this and may suggest additional investigations of interest. Another line of future work is to look at the connection between the structure of these networks and their spectra, and in particular their largest eigenvalues, as these are linked to the thresholds for spread.

Other possible analyses, not included here, are the impact of the choice of initial spreader and looking at different possible modes of spreading (e.g., broadcast rather than the independent spreading we have assumed here).

[1] L. A. Meyers, B. Pourbohloul, M. Newman, D. M. Skowronski, and R. Brunham, J. Theor. Biol. **232**, 71 (2005).

[2] R. R. Kao, D. M. Green, J. Johnson, and I. Z. Kiss, J. R. Soc. Interface **4**, 907 (2007).

[3] D. Kempe, J. Kleinberg, and A. Demers, in *Proceedings of the 33rd Annual ACM Symposium on Theory of Computing* (ACM, New York, NY, USA, 2001), pp. 163–172.

[4] D. Kempe, J. Kleinberg, and E. Tardos, in *Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (ACM Press, Washington, DC, 2003), p. 137.

[5] M. Nekovee and Y. Moreno, J. Comput. Math. **85**, 1165 (2008).

[6] N. T. J. Bailey, *The Mathematical Theory of Infectious Diseases and its Applications* (Griffin, London, 1975).

[7] H. Andersson and T. Britton, *Stochastic Epidemic Models and Their Statistical Analysis*, Lecture Notes in Statistics no. 151 (Springer, New York, 2000).

[8] D. J. Daley and J. Gani, *Epidemic Modelling* (Cambridge University Press, Cambridge, 1999).

[9] F. Ball, D. Mollison, and G. Scalia-Tomba, Ann. Appl. Prob. **7**, 46 (1997).

[10] V. Colizza and A. Vespignani, Phys. Rev. Lett. **99**, 148701 (2007).

[11] F. Ball, D. Sirl, and P. Trapman, Adv. Appl. Probab. **41**, 765 (2009).

[12] J. A. Jacquez, C. P. Simon, and J. Koopman, in *Mathematical and Statistical Approaches to AIDS Epidemiology*, Lecture Notes in Biomathematics no. 83, edited by C. Castillo-Chavez (Springer, Berlin, 1989), p. 301.

[13] F. Ball and P. Neal, Math. Biosci. **180**, 73 (2002).

[14] D. H. Zanette, Phys. Rev. E **65**, 041908 (2002).

[15] V. Isham, S. Harden, and M. Nekovee, Physica A **389**, 561 (2009).

[16] M. Nekovee, Y. Moreno, G. Bianconi, and M. Marsili, Physica A **374**, 457 (2007).

[17] J. M. Kaczmarska, Master's thesis, Department of Statistical Science, University College London (2009).

[18] P. Erdös and A. Rényi, Publ. Math. Inst. Hungar. Acad. Sci. **5**, 17 (1960).

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

[20] M. Molloy and B. Reed, Comb. Prob. Comput. **7**, 295 (1998).

[21] H. Simon, Biometrika **42**, 425 (1955).

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

[23] M. Penrose, *Random Geometric Graphs* (Oxford University Press, Oxford, 2003).

[24] Y. Moreno, M. Nekovee, and A. F. Pacheco, Phys. Rev. E **69**, 066130 (2004).

[25] G. Csardi and T. Nepusz, InterJ. Compl. Syst. 1695 (2006), [http://igraph.sf.net].

[26] R. Xulvi-Brunet and I. M. Sokolov, Phys. Rev. E **70**, 066102 (2004).

[27] S. Bansal, S. Khandelwal, and L. Meyers, BMC Bioinformatics **10**, 405 (2009), [http://www.biomedcentral.com/1471-2105/10/405].