Epidemics and random graphs

ANDREW BARBOUR AND DENIS MOLLISON

The idea of this note is to point out that the simple random graph $G(n,p)$ (*Bollobás (1985)*) can be used as an internal description for the Reed–Frost chain–binomial epidemic model *(Bailey (1975))*. This relationship is mutually beneficial: it allows us to deduce new results on each model from old results on the other, and also points to possible extensions of work on each. It also makes clear that the Reed-Frost model, in its contact structure if not its development in time, is, *pace Jacquez (1987)*, one of the simplest and most elegant of epidemic models (see also *Lefèvre and Picard, 1989*).

We begin by defining the two processes. $G(n,p)$ is defined as the random graph on n vertices with independent undirected links of probability p. The Reed–Frost process is a Markov chain with states $\{(i,r) : i,r \geq 0, i+r \leq n\}$ and with transition probabilities given by

$$(i,r) \to (j, i+r) \text{with probability } p_j, \quad 0 \leq j \leq n-i-r,$$

where

$$p_j = \mathbf{P}[Bi(n-i-r, 1-(1-p)^i) = j].$$

However, there are several other ways of formulating the Reed–Frost model, which are equivalent in the sense that they yield the same probability distributions, and this version is what we would call a surface description. By a surface description we mean a minimal specification of the joint probability distribution of the population variables. Described as above, there is no obvious reason to suppose that the model represents an epidemic process at all. To make the connection, let the variables $I(t)$ and $R(t)$ denote the numbers of infected and removed individuals at any integer time t, and suppose that there are in addition $S(t) = n - I(t) - R(t)$ susceptibles. Then, in order to model an epidemic process, we might suppose that, up to the next time step $t+1$, each of the $I(t)$ infected individuals has a chance p of making an infectious contact with any given susceptible, and that these chance contacts occur independently of one another. Thus each of the $S(t)$ susceptibles has the chance $(1-p)^{I(t)}$ of avoiding infection by the next step, independently of all the others, and so the distribution of the number newly infected by time $t+1$ is just $Bi(S(t), 1-(1-p)^{I(t)})$. If, in addition, the $I(t)$ infected individuals from step t are assumed to be removed by step $t+1$, the resulting process has the same joint probability distributions as the Reed–Frost process defined by the surface description. However, as a result of the new, internal, description, we have not only found a way of interpreting the model in a practical context, but we have also discovered a lot of information about the hidden probabilistic structure of the process, which makes the analysis of its properties much easier.

One of the key concepts in the new description is that of individuals, each of whom make infectious contact with a randomly selected list of the other individuals. As described above, the lists associated with each infective are constructed at the time that he is infected, but it is equally possible to think of the lists as being chosen at the start of the process, provided that, at the time an individual becomes infected, those members of his list who are already infected or removed are not re–infected at

the next step, and that the lists belonging to individuals who are never infected are ignored. This in turn leads to the possibility of constructing many more epidemic processes, by defining the lists L_a associated with the different individuals a according to distributions other than that used for the Reed–Frost epidemic, which pre–supposes that each individual other than a belongs to L_a with probability p, and that these events are independent. The epidemic is then defined by choosing an individual (or individuals) at random as the initial infective, defining the members of his list as the infectives at time step 1, and then, inductively, defining the infectives at time $t+1$ to be the union of the lists belonging to the infectives at time t, excluding any individuals who have already appeared as infectives at a previous time. This definition gives a further internal description of the Reed–Frost model, in essence as a directed graph.

In the case of the Reed–Frost model, the description can be further simplified. First, observe that, in any epidemic, only one of the transitions a infects b and b infects a can occur. Thus, it does not change the description of the Reed–Frost model if the assignments of a to the list L_b and of b to L_a are allowed to be dependent, provided that each event retains the probability p, and that the assignments remain independent of all other assignments. In particular, we may make the events identical, and thus only make one independent assignment for each pair $\{a, b\}$, setting $a \in L_b$ and $b \in L_a$ with probability p. But, letting this event correspond to defining an edge between a and b, we have arrived at the definition of $G(n,p)$.

This last construction shows how one can use the random graph $G(n,p)$ to construct a realization of the Reed–Frost epidemic process: one or more initial vertices are chosen at random as the $I(0)$ initial infectives, their neighbours become the $I(1)$ infectives at time 1, and, inductively, the $I(t+1)$ infectives at time $t+1$ are taken to be those neighbours of the $I(t)$ infectives at time t which have not previously been infected. Conversely, the number of components of a realization of $G(n,p)$, together with their sizes, can be constructed using Reed–Frost epidemics. Take a vertex at random: then the size of the component containing it has the distribution of $C_i = \sum_{t\geq o} I(t)_1$ in the Reed–Frost epidemic with $I(0)_1 = 1$. If $C_1 < n$, the size of the component containing a randomly chosen vertex among the remaining $n - C_1$ is, given C_1, distributed as $C_2 = \sum_{t\geq 0} I(t)_2$ in the Reed–Frost epidemic with $I(0)_2 = 1$ in a population of size $n - C_1$. Continuing in this way until, for some K, $\sum_{k=1}^{K} C_k = n$, the quantities K and $\{C_1, \ldots, C_K\}$, realized by means of Reed–Frost epidemics, yield the number and sizes of the components of $G(n,p)$.

The simplest application of these constructions is to find the asymptotic size $n - S$ of the giant component of $G(n,p)$, when n is large, $p = c/n$ and $c > 1$. By computing the probability that a randomly chosen vertex belongs to the giant component in two different ways, one obtains

$$n^{-1}\mathbf{E}(n-S) = \mathbf{P}[\sum_{t\geq 0} I(t) > n(1-1/c) \mid I(0) = 1], \tag{1}$$

and hence, using the branching process approximation to the right hand side together with the fact that $Bi(n, c/n) \approx Po(c)$, one obtains the equation

$$n^{-1}\mathbf{E}(n-S) = 1 - \sigma = e^{-c\sigma}\{1 + O(n^{-1})\}, \tag{2}$$

where $\sigma = \mathbf{E}S/n$. This also gives a very direct demonstration of the usual 'deterministic' approximation to the final size of the epidemic, being the smaller positive root of the equation $1 - \sigma = e^{-c\sigma}$. A similar argument shows, incidentally, that for all $k \geq 1$,

$$\mathbf{E}(S)_k = (n)_k \mathbf{P}[\sum_{t \geq 0} I(t) < n(1 - 1/c) \mid I(0) = k], \tag{3}$$

where$(a)_k$ denotes the product $a(a-1)\ldots(a-k+1)$, but the branching process approximation is no longer sharp enough to yield any extra useful information from the equation.

Finer information about the distribution of S can be derived from *von Bahr and Martin-Löf (1980)*, in which it is shown that

$$\mathbf{P}[\{n - \sum_{t \geq 0} I(t)) - n\sigma\}/\tau\sqrt{n} \leq x \mid \sum_{t \geq 0} I(t) > n(1 - 1/c)] = \Phi(x)\{1 + o(1)\} \tag{4}$$

as $n \to \infty$, where Φ denotes the standard normal distribution function and τ is suitably chosen. Using the succession of Reed–Frost epidemics to generate the component sizes $(C_j)_{j \geq 1}$ of $G(n,p)$, and observing that only at most a geometrically distributed number of components of size of order 1 are constructed before the giant component, corresponding to epidemics that fail to take hold, it follows that the size of the large component is generated by a Reed–Frost epidemic in a population of size $n - M$, for a random variable M of order 1. It is then easy enough to deduce also that

$$\mathbf{P}[\{(n - s) - n(1 - \sigma)\}/\tau\sqrt{n} \leq x] = \Phi(x)\{1 + o(1)\}. \tag{5}$$

Further information about the giant component could in principle be derived by more detailed analysis of the Reed–Frost model. For the analogous Markovian epidemic process, it is possible to find asymptotic expressions for the maximum number of infectives and for the duration of the epidemic. The corresponding results for the Reed–Frost process would yield information about the structure of the giant component. In particular, it could reasonably be conjectured that the distance from a randomly chosen point in the giant component to the point furthest from it can be expressed asymptotically as $k \log n + O(1)$, where k is a constant and the randomness is confined to the $O(1)$ term. However, such problems are likely to prove more difficult than their Markovian counterparts. On the other hand, certain results for the Reed–Frost process may be more easily obtained using the $G(n,p)$ imbedding. For instance, the *Daniels (1965)* Poisson limit for the number of survivors when $p = (\log n - \log \lambda)/n$ can be simply derived from the corresponding result for the number of isolated vertices in $G(n,p)$ *(Ball and Barbour (1989))*.

References

Bailey, N.T.J., "The mathematical theory of infectious diseases and its applications.," Griffin, London, 1975.

Ball, F. and Barbour, A.D., *Poisson limit theorems for epidemic models*, Math. Biosc. **95** (1989), 27-35.

Bollobás, B., "Random graphs," Academic Press, London, 1985.

Daniels, H.E., *The distribution of the total size of an epidemic*, Proc. Vth Berkeley Symp. Math. Stats Prob. **IV** (1965), 281-293.

Jacquez, John A., *A note on chain-binomial models of epidemic spread: what is wrong with the Reed–Frost formulation?*, Math. Biosci. **87** (1987), 73-82.

Lefèvre, Claude and Picard, Philippe, *On the formulation of discrete-time epidemic models*, To appear (1989).

von Bahr, B. and Martin-Löf, A., *Threshold limit theorems for some epidemic processes*, Adv. Appl. Prob. **12** (1980), 319-349.

Inst. Angew. Math., Universität Zürich, CH-8001 Zürich and Heriot-Watt University, Dept. of Actuarial Math. and Stat., Edingburg EH14 4AS, Scotland