J. Appl. Prob. **27**, 479–490 (1990)
Printed in Israel

POISSON APPROXIMATION FOR SOME EPIDEMIC MODELS

FRANK BALL,* *University of Nottingham*
A. D. BARBOUR,** *Universität Zurich*

Abstract

The Daniels' Poisson limit theorem for the final size of a severe general stochastic epidemic is extended to the Martin–Löf epidemic, and an order of magnitude for the error in the approximation is also given. The argument consists largely of showing that the number of survivors of a severe epidemic is essentially the same as the number of isolated vertices in a random directed graph. Poisson approximation for the latter quantity is proved using the Stein–Chen method and a suitable coupling.

STEIN–CHEN METHOD; EPIDEMIC SIZE; RANDOM GRAPH

1. Isolated vertices in a random directed graph

Let a random directed graph G on n vertices be constructed as follows. The sets E_i of links (i, j) emanating from the vertices $1 \leqq i \leqq n$ are determined independently of one another. The value of $N_i = |E_i|$ is selected at random from a given distribution F_i, and, given that $N_i = k$, a k-subset L_i of the vertices $\{j : 1 \leqq j \leqq n, j \neq i\}$ is chosen from the uniform distribution on all such k-subsets, and E_i is set equal to $\{(i, j): j \in L_i\}$. The set $\bigcup_{i=1}^{n} E_i$ thus obtained is a realization of G: the presence of a link (i, j) in G is thought of as the presence of a route in the direction from i to j, but not vice versa.

For each i, define

$$X_i = I\left[i \notin \bigcup_{j \neq i} L_j\right]; \quad W = \sum_{i=1}^{n} X_i.$$

Thus the vertices with $X_i = 1$ are isolated, in the sense that they cannot be reached from any other vertex, and W counts the number of isolated vertices. In the Bernoulli undirected random graph $G(n, p)$, the number of isolated vertices was shown to be approximately Poisson distributed by Erdős and Rényi (1960), provided that $np \gg 1$. In this section, we show that a similar result also holds for the directed graphs defined above.

Received 13 June 1989; revision received 28 July 1989.

* Postal address: Department of Mathematics, University of Nottingham, University Park, Nottingham NG7 2RD, UK.

** Postal address: Institut für Angewandte Mathematik, Universität Zürich, Rämistrasse 74, CH-8001 Zürich, Switzerland.

We start by considering the case where $\boldsymbol{N} = (N_i)_{1 \leq i \leq n}$ is not random, but consists of a given set of constants. Define

$$p_i(\boldsymbol{N}) = N_i/(n-1), \quad \pi_i(\boldsymbol{N}) = \mathbb{E}X_i = \prod_{j \neq i} (1 - p_j(\boldsymbol{N})), \quad \lambda(\boldsymbol{N}) = \mathbb{E}W = \sum_{i=1}^{n} \pi_i(\boldsymbol{N}),$$

and let d_{TV} denote the total variation distance between probability distributions: $d_{\mathrm{TV}}(P, Q) = \sup |P(A) - Q(A)|$. Then we can prove the following theorem.

Theorem 1. *In the case of fixed* $\boldsymbol{N}$,

$$d_{\mathrm{TV}}(\mathscr{L}(W), \mathrm{Po}(\lambda(\boldsymbol{N}))) \tag{1.1}$$
$$\leqq \max_{1 \leq i \leq n} \pi_i(\boldsymbol{N}) + \frac{1}{n} \sum_{i,j} \left[\prod_{l \neq i,j} \left(1 - \frac{N_l}{n-1}\right) - \prod_{l \neq i,j} \left(1 - \frac{N_l}{n-2}\right) \right],$$

where $\Sigma_{i,j}$ *denotes the sum over all pairs* (i, j) *with* $i \neq j$.

Proof. For fixed $\boldsymbol{N}$, the indicator random variables X_i are negatively related, in the sense of Barbour et al. (1990). Hence, by their Corollary G.2 of Chapter II, $1 - \mathrm{Var}\, W/\mathbb{E}W$ is an upper bound for $d_{\mathrm{TV}}(\mathscr{L}(W), \mathrm{Po}(\lambda(\boldsymbol{N})))$. Now direct calculation shows that

$$\mathbb{E}W - \mathrm{Var}\, W = (\mathbb{E}W)^2 - \mathbb{E}(W(W-1))$$
$$= \sum_{i=1}^{n} \pi_i^2(\boldsymbol{N}) + \sum_{i,j} \pi_i(\boldsymbol{N})\pi_j(\boldsymbol{N}) \left\{ 1 - \prod_{l \neq i,j} \left(1 - \frac{N_l}{n-2}\right) \Big/ \left(1 - \frac{N_l}{n-1}\right) \right\}.$$

Dividing by $\mathbb{E}W = \Sigma_{i=1}^{n} \pi_i(\boldsymbol{N})$, the first term of the estimate is immediate, and the second follows after observing that $\mathbb{E}W \geqq n\pi(\boldsymbol{N})$, where $\pi(\boldsymbol{N}) = \Pi_{i=1}^{n}(1 - p_i(\boldsymbol{N}))$.

Corollary 1.1. *Suppose now that the* N_i *are not fixed, so that the distributions* F_i *are no longer concentrated at single points. Then*

$$d_{\mathrm{TV}}(\mathscr{L}(W), \mathrm{Po}(\Lambda)) \leqq \pi \left\{ c_1 + c_2^2 \sum_{i=1}^{n} p_i/(1 - p_i) \right\},$$

where

$$p_i = \mathbb{E}p_i(\boldsymbol{N}) = \mathbb{E}N_i/(n-1); \quad \pi = \mathbb{E}\pi(\boldsymbol{N}) = \prod_{i=1}^{n} (1 - p_i); \tag{1.2}$$

$$c_1 = \exp\left\{ 1 + \frac{1}{2} \sum_{i=1}^{n} p_i^2/(1 - p_i) + \sum_{i=1}^{n} \frac{\mathrm{Var}\, N_i}{2(1 - p_i)(n-1)^2} \right\};$$

$$c_2 = \frac{1}{n-2} \sum_{i=1}^{n} \frac{1}{(1 - p_i)},$$

and $\mathrm{Po}(\Lambda)$ *denotes the mixed Poisson distribution with mean* Λ *distributed as* $\lambda(\boldsymbol{N})$.

Proof. It suffices to take the expectation of the right-hand side of (1.1). For the first term, we have the estimates

$$\begin{aligned}
\mathbb{E}\max_{1\leqq i\leqq n}\pi_i(\boldsymbol{N}) &\leqq \mathbb{E}\left\{\max_{1\leqq i\leqq n}\exp\left(-\sum_{j\neq i}N_j/(n-1)\right)\right\}\\
&\leqq e\mathbb{E}\exp\left(-\sum_{i=1}^{n}N_i/(n-1)\right)\\
&= e\prod_{i=1}^{n}\mathbb{E}\exp(-N_i/(n-1))\\
&\leqq e\prod_{i=1}^{n}\mathbb{E}\left(1-\frac{N_i}{n-1}+\frac{N_i^2}{2(n-1)^2}\right)\\
&= e\prod_{i=1}^{n}\left(1-p_i+\tfrac{1}{2}p_i^2+\frac{\operatorname{Var}N_i}{2(n-1)^2}\right)\\
&\leqq c_1\pi.
\end{aligned}$$

For the second, we obtain the estimate

$$\begin{aligned}
&\frac{1}{n}\sum_{i,j}\left[\prod_{l\neq i,j}(1-p_l)-\prod_{l\neq i,j}\left(1-\frac{n-1}{n-2}p_l\right)\right]\\
&\quad=\frac{1}{n}\sum_{i,j}\frac{\pi}{(1-p_i)(1-p_j)}\left[1-\prod_{l\neq i,j}\left(1-\frac{p_l}{(n-2)(1-p_l)}\right)\right]\\
&\quad\leqq\frac{1}{n}\sum_{i,j}\frac{\pi}{(1-p_i)(1-p_j)}\sum_{l=1}^{n}\frac{p_l}{(n-2)(1-p_l)}\\
&\quad\leqq c_2^2\pi\sum_{i=1}^{n}p_i/(1-p_i),
\end{aligned}$$

which completes the proof.

Approximation by a mixed Poisson distribution is less convenient than by a single Poisson distribution. However, if the variance of Λ is not too large, replacing Λ by its mean does not introduce much further error. This is the substance of the next theorem.

Theorem 2. *Let*

$$\lambda = \mathbb{E}\Lambda = \mathbb{E}\lambda(\boldsymbol{N}). \tag{1.3}$$

Then

$$d_{\mathrm{TV}}(\mathscr{L}(W), \mathrm{Po}(\lambda)) \leqq \varepsilon = \pi\left\{c_1 + c_2^2\sum_{i=1}^{n}p_i/(1-p_i) + c_2^2c_3\frac{1}{n}\sum_{i=1}^{n}\frac{\operatorname{Var}N_i}{(1-p_i)}\right\}, \tag{1.4}$$

where

$$c_3 = \exp\left\{\sum_{i=1}^{n} \frac{\operatorname{Var} N_i}{(n-1)^2(1-p_i)^2}\right\}.$$

Proof. We combine Corollary 1.1 with the triangle inequality and the estimate

$$d_{\mathrm{TV}}(\mathrm{Po}(\Lambda), \mathrm{Po}(\lambda)) \leqq \lambda^{-1} \operatorname{Var} \Lambda,$$

which follows from Barbour et al. (1990), Chapter I, Theorem G. Thus all that remains is to estimate Var Λ. Now it is immediate that

$$\operatorname{Var} \Lambda = \operatorname{Var} \lambda(N) = \mathbb{E}\left\{\sum_{i=1}^{n} \pi_i^2(N) + \sum_{i,j} \pi_i(N)\pi_j(N)\right\}$$

$$- \sum_{i=1}^{n} \pi^2(1-p_i)^{-2} - \sum_{i,j} \pi^2(1-p_i)^{-1}(1-p_j)^{-1},$$

and the expectation is evaluated by noting that

$$\mathbb{E}\left\{\left(1 - \frac{N_i}{n-1}\right)^2\right\} = (1-p_i)^2 + \frac{\operatorname{Var} N_i}{(n-1)^2} = (1-p_i)^2\left\{1 + \frac{\operatorname{Var} N_i}{(n-1)^2(1-p_i)^2}\right\}.$$

This yields the inequality

$$\operatorname{Var} \Lambda \leqq \pi^2\left[\sum_{i=1}^{n} \frac{1}{(1-p_i)}\right]^2\left\{\prod_{i=1}^{n}\left(1 + \frac{\operatorname{Var} N_i}{(n-1)^2(1-p_i)^2}\right) - 1\right\}$$

$$\leqq (n-2)^2 c_2^2 \pi^2 \sum_{i=1}^{n} \frac{\operatorname{Var} N_i}{(n-1)^2(1-p_i)^2} c_3,$$

and, since $\lambda \geqq n\pi$, the theorem follows.

Remark. If the distributions F_i were all the binomial distribution $B(n-1, p)$, with $1 \ll np = O(\log n)$, the case analogous to the undirected Bernoulli random graph $G(n, p)$, we would have $p_i = p$, $\pi_i \sim e^{-np}$ and $\operatorname{Var} N_i = (n-1)p(1-p)$, giving $c_1 \sim e$ and $c_2 \sim c_3 \sim 1$: hence, in this case,

$$d_{\mathrm{TV}}(\mathscr{L}(W), \mathrm{Po}(\lambda)) = O(npe^{-np}).$$

This is an estimate of the same order of magnitude as is obtained for the graph $G(n, p)$ (Barbour (1982), Theorem 1), where a rather simpler argument suffices, suggesting that not too much precision has been lost, despite the generality of the setting considered.

2. The survivors of a severe epidemic

The Martin-Löf (1986) epidemic is constructed from a random directed graph G by choosing one or more vertices initially at random to represent the set I_0 of initial infectives, the remaining set of vertices S_0 representing the initially susceptible members of the population. The sets S_j and I_j representing the susceptibles and infectives in the jth generation of the infection, $j \geqq 1$, are then generated recursively, by setting

$$I_j = \left(\bigcup_{i \in I_{j-1}} L_i \right) \cap S_{j-1}; \quad S_j = S_{j-1} \setminus I_j:$$

we define also $s_j = |S_j|$, $i_j = |I_j|$. The epidemic terminates when, for some j, $i_j = 0$. The Reed–Frost epidemic (see Bailey (1975); and En'ko (1889) for a forerunner of this model) is obtained if the F_i are all $B(n-1, p)$, and a discrete skeleton of the general stochastic epidemic if the F_i all have a common almost geometric distribution.

The final size distribution of this epidemic has been studied by Martin-Löf (1986), in the range where normal approximation is suitable. However, Daniels (1967) has shown for the general stochastic epidemic that, in severe epidemics, a Poisson approximation for the number s_∞ of survivors can also be appropriate. His argument does not yield any simple explanation as to why Poisson approximation should be expected. One intuitive justification is given in Cane (1966), and a formal, yet natural, proof is to be found in Sellke (1983): see also Ball (1986) for a heuristic argument, which allows the infectious periods of the different infectives to have different distributions. This paper is concerned with extending Daniels' result to the Martin–Löf model, and with quantifying the accuracy of the approximation. The essential idea is that, in a large epidemic, effectively only those individuals escape infection which correspond to vertices isolated in G (and not members of I_0), and the distribution of their number can thus be approximated by using Theorem 2. The main effort lies in making this intuitive idea precise, under appropriate assumptions on the contact distributions F_i. The key fact to be established is that s_∞ is small enough with high enough probability.

To simplify the argument somewhat, we consider a sequence of epidemics E_n for which the contact distributions are stochastically bounded below. This enables us, for each n, to couple E_n, sample path by sample path, to an epidemic E_n' which treats the n vertices exchangeably, in that all its contact distributions are the common lower bound distribution, and in such a way that, for each j, $S_j' \supset S_j$. We prove for E_n' that s_∞' is small with high probability in a series of lemmas; this entails a similar result for s_∞, and from this the Poisson approximation is established in Theorem 3.

The assumption on the contact distibutions is that, for each n, the $F_i^{(n)}$, $1 \leqq i \leqq n$, are stochastically larger than the distibution $F^{(n)}$ of a random variable

$$Z^{(n)} = \mu \log n + m(n)X \geqq 0, \tag{2.1}$$

where $1 \geqq \mu > 1/2$ and X do not depend on n, $\mathbb{E}X = 0$, $\operatorname{Var} X \leqq 1$ and $1 \leqq m(n) \leqq b_1 \log n$ for some $b_1 > 0$. For the Reed–Frost epidemic, taking $p_n \asymp \log n/(n-1)$, we have $m(n) \asymp (\log n)^{1/2}$; for the general stochastic epidemic, with mean contact number of order $\log n$, $m(n) \asymp \log n$ also. The condition $\mu > 1/2$ is needed to ensure that only isolated vertices are left uninfected, and the condition $\mu \leqq 1$ means that there is a chance of having a few survivors. From now on, where possible, the subscript n is suppressed.

The first step is to note some inequalities for the distribution of $T_{k\delta} = \sum_{j=1}^{k} Z_{j\delta}$, for any fixed $0 \leqq \delta < 1$, where the $Z_{j\delta}$ are independent and identically distributed, with

$$\mathcal{L}(Z_{j\delta}) = B(Z, 1-\delta). \tag{2.2}$$

Since X is bounded below, its moment generating function $\phi(t) = \mathbb{E}(e^{tX})$ exists for $t \leqq 0$, and hence, for $t \leqq 0$,

$$\mathbb{E}\{e^{tZ_{1\delta}}\} = \exp\{t_\delta \mu \log n\}\phi(m(n)t_\delta),$$

where

$$t_\delta = \log\{\delta + (1-\delta)e^t\} \leqq 0. \tag{2.3}$$

This enables us to prove the following lemma.

Lemma 2.1. *For any* $0 \leqq \delta < 1$, *there exists* $\sigma > 0$ *such that*

$$\mathbb{P}[T_{k\delta} \leqq x] \leqq \exp\left\{\frac{2\sigma x}{(1-\delta)m(n)} - \tfrac{1}{2}k\mu\sigma\frac{\log n}{m(n)}\right\}$$

uniformly in n, k *and* $x \geqq 0$.

Proof. By a routine argument, for $t \leqq 0$,

$$\mathbb{P}[T_{k\delta} \leqq x]\exp\{tx - kt_\delta \mu \log n\} \leqq \mathbb{E}\exp\{tT_{k\delta} - kt_\delta \mu \log n\} = \exp\{k \log \phi(m(n)t_\delta)\},$$

yielding

$$\mathbb{P}[T_{k\delta} \leqq x] \leqq \exp\{-tx + k[t_\delta \mu \log n + \log \phi(m(n)t_\delta)]\}.$$

Now, from (2.3), if $t_\delta = -s/m(n)$,

$$\begin{aligned} -t = -\log\{1 + [\exp(-s/m(n)) - 1]/(1-\delta)\} &\leqq 2(1 - e^{-s/m(n)})/(1-\delta) \\ &\leqq 2s/[m(n)(1-\delta)], \end{aligned}$$

provided that $s \leqq (1-\delta)/2$, since $m(n) \geqq 1$. Furthermore, since $\log \phi(-s) \sim \frac{1}{2}s^2 \operatorname{Var} X$ near $s = 0$, there exists $0 < \sigma < (1-\delta)/2$ such that $\sigma\mu/2b_1 \geqq \log \phi(-\sigma)$, and hence, for all n,

$$\frac{\sigma\mu}{2}\frac{\log n}{m(n)} \geqq \log \phi(-\sigma),$$

since $\log n/m(n) \geqq 1/b_1$. Taking $t_\delta = -\sigma/m(n)$ for this value of σ, the lemma follows.

Now fix $0 < \delta < 1$ and let σ be as in Lemma 2.1.

Corollary 2.2. *The following inequalities hold*:
(i) *if* $k = [b_2 n/\log n]$, *where* $b_2 = 8\delta/\mu(1-\delta)$, *then*, *for all* n *such that* $b_2 n \geqq 4 \log n$,

$$\mathbb{P}[T_{k\delta} \leqq n\delta] \leqq \exp\{-\tfrac{1}{8}\sigma\mu b_2 n/m(n)\};$$

(ii) *if* $k = n\delta - [b_2 n/\log n]$ *and* $x = b_3 n \log n$, *where* $b_3 \leqq \mu\delta/40$, *then*; *for all* n *such that* $\mu(1-\delta)\log n \geqq 20$,

$$\mathbb{P}[T_{k_0} \leqq x] \leqq \exp\left\{-\tfrac{1}{4}\mu\sigma\delta n\frac{\log n}{m(n)}\right\}.$$

The next step is to show that, if $s_0' = n(1-\delta)$ and $i_0' = n\delta - [b_2 n/\log n]$, then s_∞' is small, in the sense that $\mathbb{E}\{s_\infty'(s_\infty' - 1)\} \leqq k_1 n^{2-2\mu}$. In fact, we are able to show the same inequality with s_2' for s_∞'.

Lemma 2.3. *In the epidemic* E' *with* $s_0' = n(1-\delta)$ *and* $i_0' = n\delta - [b_2 n/\log n]$, $\mathbb{E}\{s_2'(s_2' - 1)\} = O(n^{2-2\mu})$.

Proof. We suppress the primes throughout. Let A_1 denote the event $\{\Sigma_{i\in I_0} N_i \geqq b_3 n \log n\}$, where b_3 is as in Corollary 2.2(ii): then $\mathbb{P}[A_1] \geqq 1 - e^{-\alpha_1 n}$ for some $\alpha_1 > 0$, by Corollary 2.2(ii). Hence we have

$$\mathbb{E}\{s_2(s_2-1)\} \leqq \mathbb{E}\{s_2(s_2-1)\,|\,A_1\} + O(n^2 e^{-\alpha_1 n}).$$

Now direct calculation shows that

$$\mathbb{E}\{s_2(s_2-1)\,|\,A_1\} = \mathbb{E}\{\mathbb{E}\{s_2(s_2-1)\,|\,s_1, A_1\}\,|\,A_1\} = \mathbb{E}\{s_1(s_1-1)(1-p_2)^{n(1-\delta)-s_1}\,|\,A_1\},$$

where

$$1 - p_2 = 1 - \sum_{j\geqq 0} \mathbb{P}[Z=j]\left(\frac{j(2n-3)-j^2}{(n-1)(n-2)}\right) \geqq 1 - \frac{2\mu \log n}{(n-2)}$$

is the chance of a given pair of susceptibles escaping infection by a given infective, and so it is enough to estimate

$$\mathbb{E}\{s_1(s_1-1)(1-p_2)^{-s_1}\,|\,A_1\}.$$

First, observe that, since $s_1 \leqq n(1-\delta)$,

$$s_1(s_1-1)(1-p_2)^{n(1-\delta)-s_1} \leqq 2s_1(s_1-1)(1-p_2)^{n(1-\delta)} + n^2 I[s_1 \geqq -2/\log(1-p_2)]. \tag{2.4}$$

Now, conditional on $\boldsymbol{N} = \{N_l; l \in I_0\}$, $s_1 = \Sigma_{i\in S_0} I[i \notin \cup_{l\in I_0} L_l]$ is a sum of indicator random variables which are negatively related in the sense of Barbour et al. (1990). Thus, by their Theorem R, Chapter II, for all $x \geqq 3\lambda$,

$$\mathbb{P}[s_1 \geqq x\,|\,\boldsymbol{N}] \leqq 2\mathbb{P}[\mathrm{Po}(\lambda) \geqq x],$$

where

$$\lambda = \lambda(\boldsymbol{N}) = \mathbb{E}(s_1\,|\,\boldsymbol{N}) = n(1-\delta)\prod_{l\in I_0}\left(1 - \frac{N_l}{n-1}\right) \leqq n(1-\delta)\exp\left\{-\sum_{l\in I_0} N_l/(n-1)\right\}.$$

This immediately implies the crude estimate $\mathbb{E}\{s_1(s_1-1)\,|\,\boldsymbol{N}\} \leqq 11\lambda^2(\boldsymbol{N})$, giving

$$\begin{aligned} 2\mathbb{E}\{s_1(s_1-1)\,|\,A_1\} &\leqq 22\mathbb{E}\{\lambda^2(\boldsymbol{N})\}/\mathbb{P}[A_1] \\ &\leqq 22n^2(1-\delta)^2 \exp\left\{-\frac{2i_0}{n-1}\mu \log n + i_0 \log \phi\left(\frac{-2m(n)}{n-1}\right)\right\} \qquad (2.5) \\ &\quad \times (1 + O(e^{-\alpha_1 n})) \asymp n^{2-2\mu\delta}. \end{aligned}$$

Also, on A_1, $\lambda(\boldsymbol{N}) \leqq n^{1-b_3}(1-\delta)$, whereas $-2/\log(1-p_2) \geqq (n-2)/[2\mu \log n]$, provided that $2\mu \log n/(n-2) \leqq 1/2$. Thus, for all n sufficiently large,

$$n^2\mathbb{P}[s_1 \geqq -2/\log(1-p_2)\,|\,A_1] \leqq 2n^2\mathbb{P}\left[\mathrm{Po}(n^{1-b_3}(1-\delta)) \geqq \frac{n-2}{2\mu \log n}\right]$$

(2.6)

$$= O(n^2 e^{-\alpha_2 n})$$

for any $\alpha_2 < b_3/2\mu$, where the last estimate uses the general inequality $\mathbb{P}[P \geqq y] \leqq \mathbb{E}(e^{\alpha P})e^{-\alpha y}$, with $\alpha = c \log n$ and $c < b_3$. Combining (2.4)–(2.6), it follows that

$$\mathbb{E}\{s_1(s_1-1)(1-p_2)^{n(1-\delta)-s_1}\,|\,A_1\} = O(n^{2-2\mu\delta}(1-p_2)^{n(1-\delta)}) + O(n^2 e^{-\alpha_2 n})$$

$$= O(n^{2-2\mu\delta} n^{-2\mu+2\mu\delta}),$$

and hence that $\mathbb{E}\{s_2(s_2-1)\} = O(n^{-2\mu+2\mu\delta}n^{2-2\mu\delta})$ also, as required.

The third step is to show that, with high probability, the epidemic E' with $i_0' = 1$ and $s_0' = n-1$ is at least as severe as that with $i_0' = n\delta - [b_2 n/\log n]$ and $s_0' = n(1-\delta)$. The idea is to compare E' in its early stages with a branching process with offspring distribution given by (2.2), an idea originating with Whittle (1955) and used also by Sellke (1983). In order to do this, we measure time in the epidemic in terms of the number of removals; that is, in terms of the number of individuals who have transmitted their infection. Thus many of the new time steps may occur during a single 'generation' of infection. In the new time scale, we stop the process either at the time $\tau_1(\delta)$, when for the first time there are fewer than $n(1-\delta)$ susceptibles left, or at the time τ_0 when there are no infectives remaining, whichever is the smaller. Until this time, the epidemic generates new infectives at least as fast as a process in which, at the jth time step, $Z_{j\delta}$ new infectives are produced and one previous infective removed, where the $(Z_{j\delta})_{j\geqq 1}$ are as defined before Lemma 2.1.

To see this, consider the evolution of the epidemic at the 'new' time k, when individuals $i(1), \cdots, i(k-1)$ have been removed, individual $i(k)$ is infecting and individuals $i(k+1), \cdots, i(U_{k-1})$ have previously been infected by one of the individuals $i(1), \cdots, i(k-1)$, but have not yet transmitted their infection. Let $S_{k-1} = \{1, 2, \cdots, n\} \setminus \{i(1), \cdots, i(U_{k-1})\}$ denote the set of susceptibles following the first $k-1$ infections. We suppose that $k \leqq \min(\tau_1(\delta), \tau_0)$, which implies that $k \leqq U_{k-1} \leqq n\delta$.

Suppose that $N_{i(k)} = l$: then the l members $j(1), \cdots, j(l)$ of $L_{i(k)}$ are chosen uniformly at random from $\{1, 2, \cdots, n\} \setminus \{i(k)\}$. Suppose that $j(1), \cdots, j(r-1)$ have already been chosen, and consider the choice of $j(r)$. If $U_{k-1} + |S_{k-1} \cap \{j(1), \cdots, j(r-1)\}| \leqq n\delta$, $j(r)$ has probability at least $1-\delta$ of belonging to S_{k-1}. Construct a Bernoulli random variable B_{kr} with distribution $\mathrm{Be}(1-\delta)$ satisfying $B_{kr} \leqq I[j(r) \in S_{k-1}]$ by, for instance, setting $B_{kr} = 1$ with probability $(1-\delta)/\mathbb{P}[j(r) \in S_{k-1}]$ if $j(r) \in S_{k-1}$, and $B_{kr} = 0$ otherwise, the additional randomization being made independently of everything else. Note that B_{kr} is thus independent of $\{B_{js}; j < k, s > 0\} \cup \{B_{ks}; s < r\}$. Then, on $U_k \leqq n\delta$, $U_k - U_{k-1} \geqq \Sigma_{r=1}^{l} B_{kr}$, and since, from (2.1), the distribution $F_{i(k)}$ is stochastically larger than F, it is possible, by further randomization, to realize a random variable $Z_{k\delta}$ with distribution given by (2.2) in such a way that, on $U_k \leqq n\delta$,

$$U_k - U_{k-1} \geqq \sum_{r=1}^{N_{i(k)}} B_{kr} \geqq Z_{k\delta},$$

with $Z_{k\delta}$ independent of $\{Z_{j\delta}; j < k\}$. Hence we arrive at the construction of a sequence of random variables $T_{k\delta} = \Sigma_{j=1}^{k} Z_{j\delta}$, with the same distributions as for Lemma 2.1, satisfying $U_k \geqq T_{k\delta} + 1$ whenever $U_k \leqq n\delta$.

Thus, if $U_k = k$, implying that $\tau_0 = k$, it follows that

$$\tau_0 \geqq \sigma_0 = \min\{j : T_{j\delta} = j - 1\},$$

and, if $U_k > n\delta$ but $U_j \leqq n\delta$ for all $j < k$, it follows that

$$\tau_1(\delta) = k \leqq \sigma_1(\delta) = \min\{j : T_{j\delta} \geqq n\delta\},$$

since then $T_{j\delta} + 1 \leqq U_j \leqq n\delta$ for all $j < k$. Hence, on $\tau_1(\delta) \leqq \tau_0$, $\tau_1(\delta)$ is stochastically smaller than $\sigma_1(\delta)$, and on $\tau_0 < \tau_1(\delta)$, τ_0 is stochastically larger than $\sigma_0(\delta)$. Thus

$$\mathbb{P}[\tau_1(\delta) > \tau_0] \leqq \mathbb{P}[\sigma_0(\delta) < \infty] \tag{2.7}$$

and, for any $K \geqq 1$,

$$\mathbb{P}[\tau_1(\delta) \leqq \tau_0, \tau_1(\delta) \geqq K] \leqq \mathbb{P}[\sigma_1(\delta) \geqq K] \leqq \mathbb{P}[T_{K\delta} \leqq n\delta]. \tag{2.8}$$

This is enough to establish the following lemma.

Lemma 2.4. *Consider the epidemic E' with $s_0' = n - 1$ and $i_0' = 1$, and let $A_2' = \{\tau_1(\delta) \leqq \tau_0 \wedge b_2 n/\log n\}$. Then we have*

(i) $\mathbb{P}[A_2'] \geqq 1 - \mathbb{P}[\sigma_0(\delta) < \infty] - e^{-\alpha_3 n/m(n)}$;

(ii) $\mathbb{E}\{s_\infty'(s_\infty' - 1) \,|\, A_2'\} = O(n^{2-2\mu})$.

Proof. Part (i) follows from (2.7), (2.8) and Corollary 2.2(i). However, if $\tau_1(\delta) \leqq b_2 n/\log n$, the number of infectives at this time is at least $n\delta - [b_2 n/\log n]$, and so part (ii) follows from Lemma 2.3.

Note that the probability $\mathbb{P}[\sigma_0(\delta) < \infty]$ can be evaluated as usual for the extinction probability of the branching process with offspring distribution that of $Z_{1\delta}$. In particular, the quick estimate

$$\mathbb{P}[\sigma_0(\delta) < \infty] \leqq \mathbb{P}[Z_{1\delta} = 0]/\mathbb{P}[Z_{1\delta} > 1], \tag{2.9}$$

combined with Lemma 2.1, gives the weak inequality

$$\mathbb{P}[\sigma_0(\delta) < \infty] \leqq e^{-\alpha_4 \log n/m(n)},$$

which can nonetheless be useful if $m(n) \ll \log n$.

With the help of Lemma 2.4, it is now possible to prove the following theorem.

Theorem 3. *In the sequence of epidemics E_n with $(s_0^{(n)}, i_0^{(n)}) = (n - 1, 1)$, and with contact distributions satisfying* (2.1),

$$d_{\mathrm{TV}}(\mathscr{L}(s_\infty), \mathrm{Po}(\lambda^{(n)})) = O(\eta_n),$$

where

$$\eta_n = \mathbb{P}[\sigma_0^{(n)}(\delta) < \infty] + n^{1-2\mu} \sum_{i=1}^{n} p_i^{(n)} + n^{-1}\lambda^{(n)} + \varepsilon^{(n)}$$

and $p_i^{(n)}$, $\lambda^{(n)}$ *and* $\varepsilon^{(n)}$ *are defined as in* (1.2), (1.3) *and* (1.4).

Remark. The first term in the expression for η_n bounds the probability that the epidemic is not severe, the second the probability that non-isolated individuals may also be left uninfected in a severe epidemic, and the third the probability that an isolated individual may be infected through being chosen to be the original infective. The last term bounds the error in the Poisson approximation of the number of isolated individuals, using Theorem 2.

Proof. Once again, we compare the epidemic E with that E' for which each contact distribution is $F^{(n)}$ as defined in (2.1), coupling sample path by sample path, so that $S_j' \supset S_j$ for each j. Using Lemma 2.4 for E', we see that, for an event A_2' of probability $1 - O(\eta_n)$, $\mathbb{E}\{s_\infty'(s_\infty' - 1) \,|\, A_2'\} = O(n^{2-2\mu})$. By the exchangeability of the vertices in E', it follows that

$$\mathbb{P}[\,\exists\, i, j \in S_\infty' : i \in L_j \,|\, s_\infty', A_2'] \leqq s_\infty'(s_\infty' - 1)\frac{1}{n}\sum_{i=1}^{n} p_i^{(n)},$$

and hence, since $S_\infty \subset S_\infty'$,

$$\mathbb{P}[S_\infty \neq J \setminus I_0 \,|\, A_2'] \leqq \mathbb{E}\left\{ s_\infty'(s_\infty' - 1)\frac{1}{n}\sum_{i=1}^{n} p_i^{(n)} \,|\, A_2' \right\} = O(\eta_n),$$

where J denotes the set of isolated vertices in the graph of E. Thus we find that $\mathbb{P}[S_\infty \neq J] = O(\eta_n)$, and the theorem now follows from Theorem 2.

Corollary 2.5. *In the Reed–Frost epidemic with* $(s_0, i_0) = (n-1, 1)$ *and with* $p_n = \mu \log n/(n-1)$, $1/2 < \mu \leqq 1$,

$$d_{\mathrm{TV}}(\mathscr{L}(s_\infty), \mathrm{Po}(n^{1-\mu})) = O(n^{1-2\mu}\log n).$$

Proof. Apply Theorem 3 with the common contact distribution $B(n-1, p)$.

Corollary 2.6. *In the general stochastic epidemic with* $(s_0, i_0) = (n-1, 1)$ *and with contact distribution given by*

$$\mathbb{P}[N_i = r] = \int_0^\infty e^{-t(1+\mu\log n)}\binom{n-1}{r}\{e^{t\mu\log n/(n-1)} - 1\}^r dt$$

$$= \left(\frac{1}{1+\mu\log n}\right)\left(\frac{\mu\log n}{1+\mu\log n}\right)^r\left\{1 + O\left(\frac{r^2+\log n}{n}\right)\right\},$$

$1/2 < \mu \leqq 1$, *we have the estimate*

$$d_{\mathrm{TV}}(\mathscr{L}(s_\infty), \mathrm{Po}(n^{1-\mu})) = O(1/\log n).$$

Proof. Apply Theorem 3 with the common almost geometric contact distribution.

Remark. The lack of precision in Corollary 2.6 is purely due to the chance of the epidemic failing to take hold. If $i_0(n)$ initial infectives are present, the error estimate becomes

$$O(n^{1-2\mu}\log n, n^{-\mu}\log^2 n, n^{-\mu}i_0(n), (\log n)^{-i_0(n)}),$$

which is the same as in Corollary 2.5 if $i_0(n) \asymp \log n$ and $\mu < 1$, for instance. The natural approximation in such circumstances is of course to use a mixture of the Poisson distribution and the 'small epidemic' branching process approximation for the final size. The argument already given shows that, with $i_0 = 1$ and conditional on there being a 'large' outbreak, the Poisson approximation holds to an accuracy of order η_n^*, where η_n^* is the same as η_n, but for the absence of the term $\mathbb{P}[\sigma_0^{(n)}(\delta) < \infty]$. If the mean of the approximating Poisson distribution is multiplied by the factor $(1 - i_0(n)/n)$, to allow for the fact that some members of I_0 may be isolated, the term $n^{-1}\lambda^{(n)}i_0(n)$ can also be eliminated from the error estimate.

The difference between the Reed–Frost epidemic and the general stochastic epidemic lies in the modelling of the infectious period. In the Reed–Frost model, the infectious period is assumed to be fixed, say of length one time unit, and, with a uniform contact rate of $\mu \log n/(n-1)$ between any given pair of individuals, and with independence between pairs, the contact distribution is $B(n-1, 1-e^{-\mu \log n/(n-1)}) \approx B(n-1, \mu \log n/(n-1))$. In the general stochastic epidemic, the assumptions are the same, except that the variable, exponentially distributed, infectious period introduces dependence between the events that i contacts j and i contacts k. A more realistic model for the length of the infectious period would be to take independent random lengths from a distribution concentrated on an interval $[a, b]$, where typically $0 < a < b < \infty$. If the distribution has a smooth density g, we find, with a uniform contact rate $\mu \log n/(n-1)$ and for $r \asymp \log n$, that

$$\begin{aligned}\mathbb{P}[N_i = r] &= \int_a^b g(t)e^{-t\mu \log n}\binom{n-1}{r}\{e^{t\mu \log n/(n-1)} - 1\}^r dt \\ &\sim \frac{1}{r!}\int_a^b g(t)e^{-t\mu \log n}\{t\mu \log n\}^r dt \\ &\sim \frac{1}{\mu \log n}\, g\left(\frac{r}{\mu \log n}\right),\end{aligned}$$

so that the distribution of $N_i/\mu \log n$ is well approximated by the density g. In particular, if $\int_a^b g(t)dt = 1$, such a contact distribution leads naturally to an approximation of the form (2.1) with $m(n) \asymp \log n$, apart from the limiting case of the Reed–Frost model, where g is concentrated at 1 and $m(n) \asymp \sqrt{\log n}$. Thus the methods of this section can easily be applied to such epidemics also. Note also that the approximation (2.9) then gives a bound of order $n^{-\mu a}/\log n$ for the probability of a small epidemic, indicating that the poor approximation in Corollary 2.6, occasioned by the relatively large probability of having only a small outbreak, is due to the possibility of very short infectious periods ($a = 0, g(0) > 0$).

Acknowledgement

The authors would like to express their warm thanks to the organizers of the meetings 'Oxford Probability Workshop', held in Oxford in July 1988, 'Stochastic processes in epidemic theory', held in Luminy in October 1988, and 'Mathematical models for infectious diseases', held in Oberwolfach in February 1989, during which much of this work was accomplished; and to the Swiss Nationalfonds for financial support under grant Nr. 21-25579.88.

References

BAILEY, N. T. J. (1975) *The Mathematical Theory of Infectious Diseases and its Applications*, 2nd edn. Griffin, London.

BALL, F. G. (1986) A unified approach to the distribution of total size and total area under the trajectory of infectives in epidemic models. *Adv. Appl. Prob.* **18**, 289–310.

BARBOUR, A. D. (1982) Poisson convergence and random graphs. *Math. Proc. Camb. Phil. Soc.* **92**, 349–359.

BARBOUR, A. D., HOLST, L. AND JANSON, S. (1990) *Poisson Approximation*. (to appear).

CANE, V. R. (1966) On the size of an epidemic and the number of people hearing a rumour. *J. Roy. Statist. Soc. B* **20**, 487–490.

DANIELS, H. E. (1967) The distribution of the total size of an epidemic. *Proc. 5th Berkeley Symp. Math. Statist. Prob.* **4**, 281–293.

EN'KO, P. D. (1889) The epidemic course of some infectious diseases. *Vrac'* **10**, 1008–1010, 1039–1042, 1061–1063.

ERDŐS, P. AND RÉNYI, A. (1960) On the evolution of random graphs. *Publ. Math. Inst. Hung. Acad. Sci.* **5**, 17–61.

MARTIN-LÖF, A. (1986) Symmetric sampling procedures, general epidemic processes and their threshold limit theorems. *J. Appl. Prob.* **23**, 265–282.

SELLKE, T. (1983) On the asymptotic distribution of the size of a stochastic epidemic. *J. Appl. Prob.* **20**, 390–394.

WHITTLE, P. (1955) The outcome of a stochastic epidemic — a note on Bailey's paper. *Biometrika* **42**, 116–122.