The duration of a sequence of epidemics

A. D. Barbour and Roman Kälin

Universität Zürich

Abstract

The typical duration of an epidemic in a sequence of linearly ordered populations shows a surprising non-monotonic behaviour with respect to population size, which was noted by Swinton (1998). This paper gives the sketch of a proof of the phenomenon.

1. Introduction

We consider the time course of an epidemic running through a succession of populations arranged in linear order, a situation motivated by the epidemic of phocine distemper, which spread through the seal colonies around the North Sea coast in 1988. Swinton (1998) modelled the process by taking k populations each of size N, formulating the individual epidemics within them as realizations of a Markov S–E–I–R stochastic epidemic (Diekmann & Heesterbeek (2000), p.32), with potentially infectious contacts between one population and the next occurring at rate $\beta\varepsilon I$, where I is the current number of infectives in the infecting population and β is the rate at which an infective makes infectious contacts within his own population. Swinton's simulations show that the duration of the epidemic, when plotted as a function of $\log N$, with all other parameters remaining fixed, is close to being linear, except for a jump of magnitude roughly proportional to k, occurring when $\log N$ is near $\log(1/\varepsilon)$. The approximate linear dependence on $\log N$ of the duration of an epidemic in a single population is shown in Barbour (1975); this result covers values of $\log N$ which are much smaller than $\log(1/\varepsilon)$, since the epidemic then remains confined to the single initially infected population 1. For $\log N$ much bigger than $\log(1/\varepsilon)$, on the other hand, the duration consists of that of the epidemic in the single population k, added to the sum of the $(k-1)$ transmission times T_i from first infection of population i until first infection of population $i+1$, $1 \le i \le k-1$, the latter sum being responsible for the jump.

However, Swinton (1998, Figure 1, p.221) notes a further phenomenon; that, for k large enough, there is a range of $\log N$, just above the jump, in which the median duration of the epidemic process *decreases.* The purpose of this note is to explain this rather surprising observation, highlighting the essential points of the argument without presenting all details.

The explanation is broadly as follows. The duration of a single epidemic, one element in the total duration, increases at an almost constant rate with $\log N$, so that any decrease must come from the transmission times T_i. At low population sizes N, T_1 is (almost) always infinite, but this contributes *nothing* to the total duration, because populations $2, 3, \ldots, k$ are then never infected, and the total duration is just that of the epidemic in population 1. As N increases, transmission times become finite with ever greater probability, and thus

contribute more often to the total duration; indeed, if $\log N \gg \log(1/\varepsilon)$, the probability Q_k that population k is never infected becomes exponentially small with εN. However, as N increases, the transmission times are stochastically *decreasing*, and for large εN, the rate of decrease becomes much larger than any increase in typical duration resulting from $1 - Q_k$ increasing. As a result, the typical contribution of each of the $(k-1)$ transmission times T_i to the total duration decreases with N for all N large enough, although ever more slowly as N increases. The total rate of decreases from all the transmission times together is thus typically $(k-1)$ times the individual rate of decrease, and hence, for any large enough value of N, k can be chosen large enough to make this rate of decrease outweigh the rate of increase with $\log N$ of the duration of the single population epidemic. Nevertheless, because the rate of decrease in the transmission times becomes small with increasing N, its effect on the total duration becomes negligible for any fixed k as N increases, and the almost linear growth with respect to $\log N$ at the rate appropriate to that of the duration of a single population epidemic is re-established.

2. Results

We now make this interpretation more precise. Our argument runs as follows. There are two key parameters in the model, ε and N, whose values are to be carefully chosen. We start by treating ε as fixed, and letting N increase. For large $N\varepsilon$, we show that values $k = k(N\varepsilon)$ can be found such that the total duration of the epidemic decreases with N in a certain range. We do this first in a model which is not quite the same as the model of real interest, but is a close approximation to it. The accuracy of the approximation depends only on the value of N, and by increasing N and decreasing ε while keeping $N\varepsilon$ constant, any desired accuracy of approximation can be achieved. Hence, in the model of actual interest, the total duration of the epidemic also decreases with N in a certain range, for suitably chosen values of k and ε. In particular, we shall always suppose that N and ε are chosen in such a way that $N^{\frac{7}{12}} \le 1/\varepsilon$, which, for $N\varepsilon$ held fixed, merely requires N to be chosen large enough that $N^{\frac{5}{12}} \ge N\varepsilon$.

For simplicity, and unlike Swinton, we work with the Markov S–I–R epidemic without a latent period; this does not alter the essential argument. We begin by considering the development of the epidemic in population 1. In the initial stages, with appropriate choice of time scale to make $\beta = 1$, the number $Y^{(1)}(t)$ of infected individuals develops like a birth and death process $Z(t)$, with *per capita* birth rate 1 and death rate η, where η denotes the relative removal rate. Indeed, fixing $Y^{(1)}(0) = Z(0)$, the likelihood ratio of the distribution of the $Y^{(1)}$ process with respect to that of the Z process has variance of order $O(N^{-2}e^{3(1-\eta)t}\mathbb{E}\{W^3\})$ at time t (Barbour (1980)), where W is the almost sure limit of $Z(t)e^{-(1-\eta)t}$ (Athreya & Ney (1972), Theorem 1, p.111), so that the two processes are indistinguishable off a set of probability of order $O(N^{-\frac{1}{8}})$ in the range

$$0 \le t \le t_1 := \frac{7}{12(1-\eta)} \log N,$$

at the end of which $\mathbb{E}Z(t_1) = e^{(1-\eta)t_1}Z(0) = \mathbb{E}W N^{\frac{7}{12}}$. Thereafter, the process Z behaves

almost deterministically like $We^{(1-\eta)t}$, in the sense that

$$\mathbb{P}[\sup_{s\geq t}|W - e^{-(1-\eta)s}Z(s)| > \delta] = O(\delta^{-2}e^{-(1-\eta)t})$$

for any $\delta, t > 0$; the essential randomness in Z, reflected in the random variable W, accrues during its very early development.

The process $Y^{(1)}$ also behaves almost deterministically after t_1, until the time at which it once again becomes small. This is to be understood in the sense that,

$$\mathbb{P}[\sup_{t_1\leq t\leq t_3}|N^{-1}Y(t) - \hat{y}(W^{(1)}e^{(1-\eta)t}/NK(\eta))| > N^{-\frac{1}{12}}(\log N)^2] = O(N^{-(\log N)^{\frac{1}{2}}}),$$

where $t_3 := (\frac{1}{1-\eta} + \frac{5}{12(\eta-\theta)})\log N$; this can be proved much as in Barbour (1975, p.481), by way of Barbour (1974, Corollary 1, p.32), see also Kurtz (1970). In the above, $W^{(1)}$ has the distribution of $Z(t_1)e^{-(1-\eta)t_1}$, and thus almost the same distribution as W; $\hat{y}$ is given by $\hat{y}(v) = y((1-\eta)^{-1}\log v)$, where $(x(t), y(t))$ solve the deterministic differential equations for the Markov epidemic process (Bailey (1975), equations (6.1), p.82),

$$\frac{dx}{dt} = -xy; \qquad \frac{dy}{dt} = (x-\eta)y,$$

with initial condition $x(0) = \eta$, $y(0) = 1 - \eta + \eta\log\eta$; and $K(\eta) = \eta^{-1}(1-\eta)^2e^{g(\eta)}$, with

$$g(\eta) = \eta\int_\eta^1 \frac{(z - 1 - \log z)}{z(1-z)(1-z+\eta\log z)}\,dz.$$

Note that the solution $x(t)$ of the above differential equation satisfies

$$-t = \int_\eta^{x(t)} \frac{dv}{v(1+\eta\log v - v)},$$

from which the approximation $x(t) \sim 1 - \frac{1-\eta}{\eta}e^{(1-\eta)t+g(\eta)}$ $(t \to -\infty)$ follows. Since also $y(t) = 1 + \eta\log x(t) - x(t)$, we get $y(t) \sim K(\eta)e^{(1-\eta)t}$ as $t \to -\infty$ and hence

$$\hat{y}(W^{(1)}e^{(1-\eta)t}/NK(\eta)) = y(t + (1-\eta)^{-1}\log(W^{(1)}/NK(\eta))) \sim \frac{W^{(1)}}{N}e^{(1-\eta)t}$$

as $t \to -\infty$, this way matching and continuing properly the early birth-death process approximation.

We also have $\hat{y}(z) \sim K(\eta)z(1-cz)$ as $z \to 0$, and $\hat{y}(z) \sim Cz^{-\gamma}$ with $\gamma = (\eta-\theta)/(1-\eta)$ as $z \to \infty$, for constants c and C, with θ the solution in $(0,\eta)$ to the final size equation $\theta - \eta\log\theta = 1$.

Our approximation to the S–I–R epidemic process in population i thus consists of a birth and death process (with immigration from the epidemic in population $i-1$), running for a time $\frac{7}{12}(1-\eta)^{-1}\log N$, followed by a deterministic process $N\hat{y}(W^{(i)}e^{(1-\eta)t}/NK(\eta))$, where the random variable $W^{(i)}$ is taken to have the distribution of the limiting random variable

for the birth and death process with immigration. We shall also suppose that the process of infection into population $i+1$ is confined to this second, deterministic phase; the probability of an infection in the birth and death phase is of order $O(N\varepsilon/N^{\frac{5}{12}})$, and can be made small for fixed $N\varepsilon$ by increasing N. There is a final birth and death phase in the last population infected which also contributes to the duration; however, this has already been treated in Barbour (1975), and we do not need here to consider it further.

The infection process from the first population into the second can now be analyzed in this framework. If it were the birth and death process Z generating infections at *per capita* rate ε, then the first infection would occur at a time close to the time at which Z first reaches the level $1/\varepsilon$; that is, near the time $\tau := (1-\eta)^{-1}\log(1/W\varepsilon)$, with the randomness in the early history of Z reflected in the appearance of W in the formula for τ. For the epidemic process, we define the 'effective transmission time' from population i to population $i+1$ to be $\tau^{(i)} := (1-\eta)^{-1}\log(1/W^{(i)}\varepsilon)$, and consider this time to be the time origin for the epidemic process in population $i+1$. The actual infections from population i into population $i+1$ still take place at random times, relative to $\tau^{(i)}$ as origin, because of the randomness of the infection process, which is a Poisson process with time varying rate given by

$$N\varepsilon\hat{y}(e^{(1-\eta)u}/N\varepsilon) \tag{1}$$

at time u; this randomness is simply incorporated into the initial growth random variable $W^{(i+1)}$ for the epidemic in population $i+1$. Note that, with the new choice of time origin, u may be either positive or negative, and that, for moderate u and $N\varepsilon$ large,

$$N\varepsilon\hat{y}(e^{(1-\eta)u}/N\varepsilon) \sim K(\eta)e^{(1-\eta)u}.$$

In particular, the total number of infections occurring at times $u < 0$ has a Poisson distribution with mean

$$K(\eta)\int_{-\infty}^{0} e^{(1-\eta)u}(1+O(e^{(1-\eta)u}/N\varepsilon))\,du = \{K(\eta)/(1-\eta)\}(1+O(1/N\varepsilon)),$$

and is finite almost surely. However, an infection at time u contributes an amount $e^{-(1-\eta)u}W_u$ to the random variable $W^{(i+1)}$, where W_u has distribution close to $\mathcal{L}(W \,|\, Z(0)=1)$ and is independent of all other contributions to $W^{(i+1)}$. Hence the total contribution $W_-^{(i+1)}$ to $W^{(i+1)}$ which arises from infections at times $u<0$ is not helpfully described by means of moments; however, it is an almost surely finite infinitely divisible random variable, with Lévy measure

$$\frac{N\varepsilon}{z}\int_0^\infty \hat{y}\left(\frac{v}{z(1-\eta)N\varepsilon}\right)e^{-v}\,dv \quad \text{on} \quad 1 \le z < \infty, \tag{2}$$

and has Laplace transform close to $\{s/(s+1)\}^{sK(\eta)}$, for $N\varepsilon$ large. For $u>0$, the moments of the total contribution $W_+^{(i+1)}$ are better behaved: the variance is bounded uniformly in N,

and the mean is

$$\begin{aligned}\int_0^\infty \varepsilon N\hat{y}(e^{(1-\eta)u}/N\varepsilon)e^{-(1-\eta)u}du &= \frac{1}{1-\eta}\int_{1/N\varepsilon}^\infty z^{-2}\hat{y}(z)\,dz \\ &= \frac{1}{1-\eta}\left\{K(\eta)\int_{1/N\varepsilon}^1 z^{-1}dz + \int_{1/N\varepsilon}^1 z^{-2}\{\hat{y}(z)-K(\eta)z\}dz + \int_1^\infty z^{-2}\hat{y}(z)\,dz\right\} \\ &= \frac{1}{1-\eta}\left\{K(\eta)\log(N\varepsilon)+O(1)\right\}.\end{aligned}$$

Note also that we can assume that all the infection takes place during the birth and death phase of the epidemic in population $i+1$, since the probability of there being an infection at any time $u > \frac{7}{12}(1-\eta)^{-1}\log N$ is at most

$$\frac{N\varepsilon}{1-\eta}\int_{1/\varepsilon N^{\frac{5}{12}}}^\infty z^{-1}\hat{y}(z)\,dz = O((N\varepsilon)^{\gamma+1}N^{-7\gamma/12}),$$

with $\gamma = (\eta-\theta)/(1-\eta)$ as before, and can be made arbitrarily small for $N\varepsilon$ fixed by choosing N to be large.

Thus, since $W^{(i+1)} = W_-^{(i+1)} + W_+^{(i+1)}$, we have

$$\log W_+^{(i+1)} \le \log W^{(i+1)} \le \log(1+W_-^{(i+1)}) + \log(1+W_+^{(i+1)}),$$

so that $\log W^{(i+1)} - \log\log N\varepsilon$ is a random variable which is stochastically bounded in N. Furthermore, since the Poisson infection rate can be written as

$$N\varepsilon\hat{y}(e^{(1-\eta)u}/N\varepsilon) = e^{(1-\eta)u}\left\{N\varepsilon e^{-(1-\eta)u}\right\}\hat{y}(e^{(1-\eta)u}/N\varepsilon),$$

and since $z^{-1}\hat{y}(z)$ is a decreasing function of z, because

$$\frac{d}{dz}(z^{-1}\hat{y}(z)) = \frac{\hat{y}(z)}{(1-\eta)z^2}\left\{\left(\frac{1}{y}\frac{dy}{dt}\right)\left(\frac{1}{1-\eta}\log z\right) - (1-\eta)\right\} \le 0,$$

and

$$\frac{1}{y}\frac{dy}{dt} = x - \eta \le 1-\eta$$

for all t , it follows that the distribution of $W^{(i+1)}$ is stochastically increasing in N for fixed ε. Hence the effective transmission time $\tau^{(i)} := (1-\eta)^{-1}\log(1/W^{(i)}\varepsilon)$ is stochastically decreasing in N for fixed ε, with mean like $(1-\eta)^{-1}\{\log(1/\varepsilon) - \log\log(N\varepsilon) + O(1)\}$, for $2 \le i \le k-1$. The effective transmission time $\tau^{(1)}$ is different, since here there is no immigration of infection from a previous population, and the random variable $W^{(1)}$ is naturally taken to be that obtained from the initial birth and death process with a single initial individual, conditioned not to hit zero; this has an exponential distribution with mean $(1-\eta)^{-1}$ (Kendall (1966), equation (13), p.395). On the other hand, the epidemic in population k runs faster than a single S–I–R epidemic, because of the infection process from population $k-1$, again by an amount of approximately $-(1-\eta)^{-1}\log\log(N\varepsilon)$, so that this decrease in duration actually occurs $(k-1)$ times.

We now show that the presence of the term $-(k-1)(1-\eta)^{-1}\log\log(N\varepsilon)$ in the total duration of the epidemic is enough to explain the phenomenon observed by Swinton (1998). To do so, we need to show that events such as the failure of transmission from one population to cause a large epidemic in the next are sufficiently rare. This is relatively easily seen. We condition on the first population having a large epidemic. Then the number of infections immigrating into the second population is a Poisson distributed random variable with mean $N\varepsilon\int_{t_1}^{t_3}\hat{y}(W^{(1)}e^{(1-\eta)t}/NK(\eta))\,dt = O(N\varepsilon)$. So we have more than $c_1N\varepsilon$ infections immigrating into the second population, except for a set of probability $O(e^{-c_2N\varepsilon})$. Each of these infections fails to cause a large epidemic in the initial birth and death process with probability of at most η. So the chance that none of these new infections results in a large epidemic is at most

$$\eta^{c_1N\varepsilon} = O(e^{-c_3N\varepsilon}).$$

Hence a large epidemic in one population leads to a large epidemic in the next with probability at least $1-Ce^{-\alpha N\varepsilon}$ for some $C,\alpha>0$, and thus the chance that not all populations become infected, conditional on the first having a large epidemic, is of order $O(ke^{-\alpha N\varepsilon})$.

The effect of this on the median of the total duration is correspondingly small. The total duration is largely made up of the sum of the effective transmission times $\tau^{(i)}$, which are multiples of the $\log W^{(i)}$. Now the distribution of each $W^{(i)}$ is infinitely divisible, with Lévy measure as in (2), but on the whole of $(0,\infty)$. This is the distribution of a normalized sum $\{(1-\eta)N\varepsilon\}^{-1}\sum_{j=1}^{M}U_j$ of independent random variables U_j, $j\geq 1$, with density

$$g_U(u) := \frac{1}{K_0u}\int_0^\infty \hat{y}(v/u)e^{-v}\,dv, \qquad 0<u<\infty,$$

where $M\sim \mathrm{Po}\,(K_0N\varepsilon)$ is independent of the U_j's and $K_0=\int_0^\infty v^{-1}\hat{y}(v)\,dv$. Note that g_U is continuous and decreasing, with

$$g_U(u)\sim\begin{cases} c_1u^{\gamma-1}, & \text{if } u\to 0;\\ c_2u^{-2}, & \text{if } u\to\infty,\end{cases}$$

so that U_1 is in the domain of attraction of a positive stable law of index $\alpha=1$, and the distribution of $\sum_{j=1}^{r}U_j$ has bounded continuous density for every $r\geq 1/\gamma$. Hence, and in view of Gnedenko and Kolmogorov (1954, Chapter 8, §46, Theorem 2), it follows that $\mathcal{L}(W^{(i)})$ can be written as a mixture of two distributions, in which the main component, with weight

$$\mathbb{P}\,[M\geq K_0N\varepsilon/2] = 1-O(Ce^{-\alpha N\varepsilon})$$

for some $C,\alpha>0$, has bounded continuous density on $(0,\infty)$. The same is then also true of the distribution of $\log W^{(i)}$, the main component now with support $(-\infty,\infty)$ and with finite variance; the minor component gives some mass to $-\infty$, again reflecting the event of no transmission of the epidemic from one population to the next. Hence, and in view of Gnedenko and Kolmogorov (1954, Chapter 8, §46, Theorem 1), the probability density of $\sum_{i=1}^{k-1}\tau^{(i)}$ is uniformly bounded below by $c(k-1)^{-1/2}$ in the neighbourhood of its median,

for some $c > 0$, provided that $k\mathbb{P}[M < K_0 N\varepsilon/2] < 1/4$, and thus its median can be shifted by at most $O(k^{3/2}e^{-\alpha N\varepsilon})$ as a result of incomplete transmission, for some $\alpha > 0$.

It now remains to consider what happens when N increases and $N\varepsilon$ is large. Changing N to $N + 1/\varepsilon$ leads to an increase in the duration of the single population epidemic of approximately $C_0/(N\varepsilon)$ for $C_0 = \{(1-\eta)^{-1} + (\eta-\theta)^{-1}\}$, and the worst case increase resulting from possibly incomplete transmission at population size N becoming complete at population size $N + 1/\varepsilon$ is at most $C_1 k^{3/2} e^{-\alpha N\varepsilon}$. To set against this is a decrease of at least $C_2(k-2)\{N\varepsilon \log(N\varepsilon)\}^{-1}$ from the transmission times. Thus, provided that $N\varepsilon$ is large enough, there is a (large) choice of $k = k(N\varepsilon)$ such that the total duration of the epidemic process decreases when changing N to $N + 1/\varepsilon$.

All these calculations have been carried out using a model which is only close to the truth. However, as discussed above, the differences all become small as functions of N, if N is increased while $N\varepsilon$ remains fixed. Hence their influence is negligible provided that ε is chosen to be small enough.

3. Discussion.

Swinton (1998, Section 3.5) remarks on the non–monotonicity observed in his simulation studies, both in general terms, and with reference to an approximate description of the transit time distribution, introduced in his Section 3.4. For the latter, he notes that non–monotonicity does not appear to be a feature of his approximation at the parameter values used in the simulations. The general discussion is always difficult, because of the trade–off, as N increases, between a higher number of patches being affected by the epidemic, increasing the overall duration, and the transit times (if finite) between patches becoming shorter.

In our analysis, the most important factor in reducing the overall duration T_N in the non–monotonic régime is one that is not allowed for by Swinton. The transit time from one population to the next comes to an end with the first 'successful' infection into the second population, but the time course of the epidemic in the second population is also substantially influenced by subsequent infections from the first population. Individually, an infection occurring at time t after the first has an effect of relative order $e^{-(1-\eta)t}$ on the development of the epidemic, but the intensity of such infections increases initially at a rate proportional to $e^{(1-\eta)t}$, as in the argument between Equations (1) and (2), and it is these subsequent infections which result (to leading asymptotic order) in the reduction in duration of $(1-\eta)^{-1} \log\log N\varepsilon$, relative to that of an epidemic brought about by a single successful infection. Note that the quantity $N\varepsilon$ is itself proportional to the total number of infections from the first population to the second, reflecting the part that they all play in transmission.

Precise comparison of our asymptotics with Swinton's simulations is not possible, for two reasons. First, we consider a model without latent period, whereas his simulations included one. Secondly, we have used only leading term asymptotics as $N\varepsilon \to \infty$. However, to get a rough idea of what is to be expected, we matched his parameter values as far as possible to

ours, taking

$$\eta = 1/R_0 = 1/2.8, \quad \text{implying} \quad \theta = 0.075, \quad \text{and} \quad \varepsilon = \tfrac{1}{2}\rho = 5 \times 10^{-5}.$$

We then used the asymptotic formula

$$T_N \approx \left\{ \left(\frac{1}{1-\eta} + \frac{1}{\eta - \theta} \right) \log N + \frac{k-1}{1-\eta} \left(\log(1/\varepsilon) - \log\log N\varepsilon \right) \right\} \frac{5}{280},$$

which is our approximation when the probability is very small that not all populations become infected, this being close to $(k-1)\theta^{N\varepsilon}$; here, the factor 5 is to allow for Swinton's latent period, four times longer than the infectious period, and $1/\beta = 1/280$, corresponding to his infection rate. The agreement of this formula with Swinton's simulations (Swinton 1998, Figure 1), which it seems were actually conducted with $\rho = 10^{-4}$, and not with the quoted value of $\rho = 10^{-3}$, was found to be remarkably good throughout the range $10^5 \le N \le 10^7$ ($5 \le N\varepsilon \le 500$).

Acknowledgement

We wish to thank a referee for encouraging us to add some discussion. This work was supported in part by Schweizer Nationalfondsprojekte Nr. 20-61753.00 and 20-67909.02.

References.

Athreya, K. B. & Ney, P. E. (1972). *Branching Processes.* Springer, Berlin.

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

Barbour, A. D. (1974). On a functional central limit theorem for Markov population processes. *Adv. Appl. Prob.* **6**, 21–39.

Barbour, A. D. (1975). The duration of the closed stochastic epidemic. *Biometrika* **62**, 477–482.

Barbour, A. D. (1980). Density dependent Markov population processes. In: Biological growth and spread, Eds W. Jäger, H. Rost and P. Tautu. *Lecture Notes in Biomathematics* **38**, 36–49. Springer, Berlin.

Diekmann, O. & Heesterbeek, J. A. P. (2000). *Mathematical Epidemiology of Infectious Diseases: Model Building, Analysis and Interpretation.* Wiley, Chichester.

Gnedenko, B. V. & Kolmogorov, A. N. (1954). *Limit Distributions for Sums of Independent Random Variables.* Addison–Wesley, Cambridge, Mass.

Kendall, D. G. (1966). Branching processes since 1873. *J. London. Math. Soc.* **41**, 385–406.

Kurtz, T. G. (1970). Solutions of ordinary differential equations as limits of pure jump Markov processes. *J. Appl. Prob.* **7**, 49–58.

Swinton, J. (1998). Extinction times for epidemics in spatially structured closed populations. *Bull. Math. Biol.* **60**, 215–230.