Available at
www.ElsevierMathematics.com
POWERED BY SCIENCE DIRECT®
Computational Statistics & Data Analysis 45 (2004) 3–23
COMPUTATIONAL STATISTICS & DATA ANALYSIS
www.elsevier.com/locate/csda

# Modeling computer virus prevalence with a susceptible-infected-susceptible model with reintroduction

John C. Wierman[a,*], David J. Marchette[b]

[a] *Department of Mathematical Sciences, The Johns Hopkins University, Baltimore, MD 21218-2682, USA*
[b] *Naval Surface Warfare Center, Code B10, Dahlgren, VA 22448, USA*

Received 10 January 2002; received in revised form 28 August 2002

**Abstract**

Computer viruses are an extremely important aspect of computer security, and understanding their spread and extent is an important component of any defensive strategy. Epidemiological models have been proposed to deal with this issue, and we present one such here. We consider a modification of the Susceptible–Infected–Susceptible (SIS) epidemiological model as a model of computer virus spread. This model includes a reintroduction parameter, which models the rerelease of a computer virus, or the introduction of a new virus. This is a more realistic model of computer virus spread than the standard SIS model, and can be used to understand the behavior of the quasi-stationary regime of the SIS model.


*Keywords:* Epidemic; Computer virus; SIS Model; Quasi-stationary

## 1. Introduction

Computer viruses have cost billions of dollars since their invention in the 1980s. Actual figures are somewhat speculative, but have been reported to be $12.1 billion in 1999, $17.1 billion in 2000 and $10.7 billion for the first three quarters of 2001 (Abreu, 2001). Thus, methods to analyze, track, model, and protect against viruses are of considerable interest. This paper describes some methods from epidemiology that are of value in understanding the spread of computer viruses. We will discuss some

* Corresponding author. Tel./fax: +1-410-516-7211.
*E-mail addresses:* wierman@jhu.edu (J.C. Wierman), marchettedj@nswc.navy.mil (D.J. Marchette).

0167-9473/$ - see front matter 
doi:10.1016/S0167-9473(03)00113-0

elementary epidemic models, and show how they relate to the problem of modeling the spread of computer viruses.

The model investigated is the Susceptible–Infected–Susceptible (SIS) model. In this model, susceptibles (labeled S) are susceptible to infection from any infected individual. When a susceptible becomes infected (labeled I), it is immediately infectious. Upon "cure", an individual is labeled S and is immediately once again susceptible. This is a homogeneous model, where every individual has the same probability of cure or, if susceptible, of infection, and each infected individual has the opportunity to infect each susceptible individual.

The deterministic SIS model was introduced by Ross (1915). It leads to a logistic curve which predicts extinction of the infection whenever a basic reproductive ratio $R < 1$, and predicts a steady-state endemic infection level if $R > 1$ whenever the initial proportion of infected individuals is positive.

The stochastic SIS model was introduced by Weiss and Dishon (1971). It is a continuous time Markov birth-and-death process (Cavender, 1978) used to model epidemics, see for example Ball (1999) Jacquez and Simon (1993) Kryscio and Lefèvre (1989) and Nåsell (1996, 1999) and also transmission of rumors (Bartholomew, 1976) and chemical reactions (Oppenheim et al., 1977).

The long-term behavior of the deterministic and stochastic versions of the SIS model are entirely different. In the stochastic SIS model, the infection becomes extinct with probability one, regardless of the parameters of the model. However, the time to extinction depends on the infection and cure rate parameters, and can be extremely large. The probability distribution of the number of infected individuals, during the long time until extinction, is sometimes approximated by the distribution under the condition that extinction has not occurred, which has been called the quasi-stationary distribution. The concept of the quasi-stationary distribution of a continuous-time Markov process was introduced by Darroch and Seneta (1967) for finite state-space chains, and was first applied to epidemics by Kryscio and Lefèvre (1989), whose work was extended by Nåsell (1999) using asymptotic approximation.

Epidemic models for computer viruses have been investigated since at least 1988. Murray (1988) appears to be the first to suggest the relationship between epidemiology and computer viruses, although he did not suggest any specific models. More recently, Kephart and White (1991, 1993) and Kephart et al. (1993) have investigated SIS models for computer virus spread.

In this paper we analyze a birth-and-death process with reintroduction to model computer virus spread. This is a variation on the model of Kephart and White (1991, 1993). We assume an SIS model, where any susceptible computer can become infected, and any cured computer is immediately susceptible. The model is introduced in Section 2, followed by an analysis of the model in Section 3. The results are illustrated by simulations in Section 4, followed by a discussion of potential further research.

## 2. The SIS model with reintroduction

Let the number of computers be $n$, the infection rate for any infected–susceptible pair $r$, and the cure rate for any infected computer $c$. We will model the SIS model as an

$n+1$ state continuous time Markov process, where the states denote the number of infected machines. We can represent the process as a birth-and-death process with birth rates

$$\lambda_i = ri(n-i)$$

and death rates

$$\mu_i = ci,$$

for $i = 1,\ldots,n-1$. The parameter $c$ is interpreted as the cure rate for a single infected computer, and $r$ is interpreted as the infection rate from a particular infected computer to a particular susceptible computer. It is important to note that the total infection rate for a susceptible computer is $r$ times the size of the infected population, which may be of order $rn$.

Kephart and White studied a similar model both analytically and with simulations. They found that, under certain conditions, the infected population initially grew rapidly. In fact, in the early stages of an epidemic, the process is well approximated by a branching process, and thus exhibits an exponential growth rate. The branching process approximation is sometimes called Kendall's approximation (Kendall, 1956), and has been rigorously established for many stochastic epidemic processes (Ball, 1983a, b; Ball and Donnelly, 1995; Martin-Löf, 1986; Scalia-Tomba, 1985; von Bahr and Martin-Löf, 1980). After the rapid growth phase, the epidemic appeared to reach an equilibrium. However, since state 0 is absorbing and the state space is finite, the process will become absorbed in state 0 with probability one, i.e. the infection will become extinct. Since extinction is certain, this apparent equilibrium is temporary, and the situation represents a case of "metastability", where a temporary equilibrium or quasi-stationary distribution persists for a long time until a transition occurs to a final equilibrium, which in our model is extinction. In order to study this temporary equilibrium analytically, we modify the model by allowing the infection to restart with a rate $a$ when the infected population size is zero. This is mathematically convenient, because it eliminates the absorbing state, so the infected population size has a non-trivial limiting distribution, which serves as an approximation to the temporary equilibrium distribution. The addition of this "reintroduction parameter" may also make the model more realistic. This corresponds to the possibility that the infection is "archived" either intentionally or unintentionally and reintroduced at a later time. If we consider infection to be more broadly defined than infection by a specific virus, but rather infection by any computer virus, then reintroduction corresponds to the introduction of a new virus. This is perhaps the most appropriate application to consider for this model.

Note that the SIS model is particularly appropriate if one treats "virus infection" as any infection by any virus. In this case, short of taking a computer off the network, any cured computer is, in principle, immediately susceptible, due to the introduction of new viruses.

The revised model is a birth-and-death Markov process with

$$\lambda_0 = a,$$

$$\lambda_i = ri(n-i), \quad i = 1,\ldots,n-1,$$

$$\mu_i = ci, \quad i = 1,\ldots,n.$$

The epidemic process with reintroduction can serve as an approximation for the metastable state in the epidemic process without reintroduction. First, the two processes are identical until the epidemic process without introduction becomes extinct, after which the epidemic with reintroduction again becomes an active epidemic after a random waiting period. Thus, the epidemic with reintroduction will always have a greater or equal number of infected individuals than the epidemic without reintroduction. So, the distribution of the epidemic with reintroduction is stochastically larger than that of the epidemic without reintroduction at all times. On the other hand, the epidemic with reintroduction does occasionally become extinct for a short time, then grows rapidly up to the temporary equilibrium of the epidemic without reintroduction. Because of the (relatively short) time spent with low infected population sizes, the stationary distribution of the epidemic with reintroduction will tend to have (slightly) lower population sizes than the temporary equilibrium of the epidemic without reintroduction. Thus, the stationary distribution can be expected to be a good approximation to the distribution in the metastable state, but has the advantage that it can be studied analytically in terms of its parameters. Figs. 4 and 5 show simulation results that confirm the validity of the approximation for some realistic parameter values.

The form of the stationary distribution is given by the standard formula (see, e.g. Ross, 1996, pp. 253–254)

$$P_0 = \frac{1}{1 + \sum_{i=1}^{n} \lambda_0\lambda_1 \cdots \lambda_{i-1}/\mu_1\mu_2 \cdots \mu_i},$$

$$P_k = P_0 \frac{\lambda_0\lambda_1 \cdots \lambda_{k-1}}{\mu_1\mu_2 \cdots \mu_k}.$$

Calculating the factor in $P_k$ yields

$$\frac{\lambda_0\lambda_1 \cdots \lambda_{k-1}}{\mu_1\mu_2 \cdots \mu_k} = \frac{ar^{k-1}(n-1)!}{kc^k(n-k)!}.$$

Further calculations can be made, with the help of Mathematica, to get a closed form solution to the probabilities. It can be shown that

$$P_0 = \frac{c}{c + a_pF_q[\{1,1,n-1\},\{2\},-r/c]}, \tag{1}$$

where ${}_pF_q$ is a generalized hypergeometric function (see Wolfram, 1996, pp. 750–751). In this case $p = 3$ and $q = 1$. Similarly

$$P_k = \frac{ar^{k-1}(n-1)!}{c^{k-1}k(n-k)!(c + a_pF_q[\{1,1,n-1\},\{2\},-r/c])}. \tag{2}$$

These expressions can be used to calculate various functionals of the distribution. For example, the mode can be obtained by solving Eq. (2) for a maximum using Mathematica's optimization solver (noting that Mathematica treats the function as a continuous one, while in reality it is discrete). The mean and variance of the distribution are also available, either in closed form using Mathematica or numerically. Section 4 demonstrates a close agreement between this analysis and simulations.

We can get an instructive approximation to $P_0$ through the following calculation. Recall that

$$P_0 = \frac{1}{1 + \sum_{k=1}^{n} ar^{k-1}(n-1)!/kc^k(n-k)!}$$
$$= \frac{c}{c + a(n-1)! \sum_{k=1}^{n} (r/c)^{k-1}/k(n-k)!}. \tag{3}$$

Writing the sum in Eq. (3) as

$$f(x) = \sum_{i=1}^{n} \frac{x^{i-1}}{i(n-i)!} = \sum_{i=0}^{n-1} \frac{x^{n-i-1}}{(n-i)i!},$$

we recognize this as an integral (suppressing the constant)

$$\int f(x)\,\mathrm{d}x = \sum_{i=0}^{n-1} \frac{x^{n-i}}{i!} \sim x^n \mathrm{e}^{1/x},$$

so differentiating both sides gives

$$f(x) \sim x^{n-1}\mathrm{e}^{1/x}(1+x). \tag{4}$$

Plugging 4 into 3 we have

$$P_0 \sim \frac{c}{c + a(n-1)!(\frac{r}{c})^n \mathrm{e}^{c/r}(\frac{r}{c}+1)}. \tag{5}$$

## 3. Analysis

To obtain an approximation of the typical number of infected computers, we want to evaluate a measure of central tendency of the limiting distribution. Since the expression for the mean obtained from Mathematica in terms of generalized hypergeometric functions does not contribute to our understanding, and the median may only be computed numerically, we determine the mode of the limiting distribution, which is given by a relatively simple formula that is easily interpreted. However, note that further study, later in this section, shows that the distribution is well approximated, for some parameter ranges, by Poisson and Normal distributions, for which the mode and mean (and, for the Normal distribution, the median as well) coincide.

### 3.1. The mode

Using the standard approach for finding the mode of the binomial and Poisson distributions, we consider the ratio of two successive probabilities:

$$\frac{P_{k+1}}{P_k} = \frac{(a/(k+1))(r^k/c^{k+1})((n-1)!/(n-k-1)!)}{(a/k)(r^{k-1}/c^k)((n-1)!/(n-k)!)},$$
$$= \frac{kr(n-k)}{(k+1)c}.$$

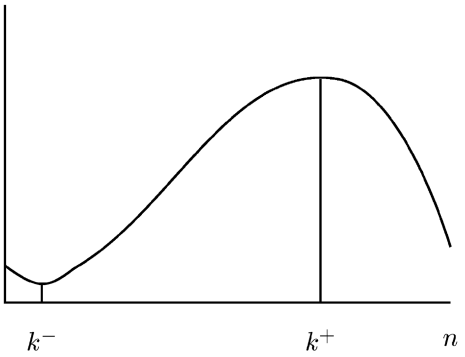


Fig. 1. Sketch showing the presumed shape of the distribution.

To determine when the ratio is $=, >$, or $< 1$, we solve for $k$ in the equation

$$\begin{aligned}
&\frac{kr(n-k)}{(k+1)c} = 1 \\
&\quad \Leftrightarrow rnk - rk^2 = ck + c \\
&\quad \Leftrightarrow rk^2 + (c - rn)k + c = 0 \\
&\quad \Leftrightarrow k^{\pm} = \frac{-(c-rn) \pm \sqrt{(c-rn)^2 - 4rc}}{2r}. \qquad (6)
\end{aligned}$$

The ratio is greater than one for values of $k$ in the interval between the two solutions, so the probabilities $P_k$ are nondecreasing from $\lceil k^- \rceil$ to $\lceil k^+ \rceil$. The probabilities decrease for $k < \lceil k^- \rceil$ and $k > \lceil k^+ \rceil$. Thus, the mode is either 0 or $\lceil k^+ \rceil$. By choosing the reintroduction rate parameter sufficiently large, the probability of being in state zero can be made as small as desired, so in this case the mode is $\lceil k^+ \rceil$. Note that $k^{\pm}$ are independent of the reintroduction rate $a$.

Considering the terms that make up $k^{\pm}$, we note that

$$\frac{-(c-rn)}{2r} = \frac{n}{2} - \frac{c}{2r},$$

is somewhat less than half the population size, while

$$\frac{\sqrt{(c-rn)^2 - 4rc}}{2r}$$

is slightly less than the first term. Thus, the interval of increasing probabilities covers nearly the entire range of population sizes, if $r$ and $c$ are fixed, in which case the mode is near $n$, the population size. If $c$ grows linearly with $n$, then the mode can be significantly less than $n$ (see Fig. 1), which means the infection becomes endemic, with a stable proportion of the population infected.

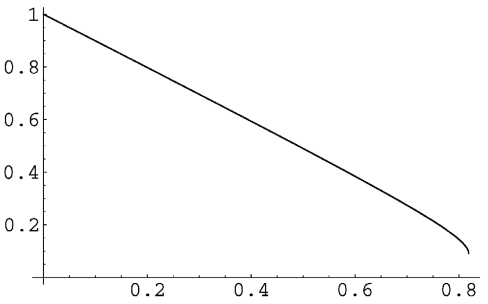


Fig. 2. The mode calculated using Eq. (6) with $r = 0.01$ and $n = 100$ as a function of $c$. The proportion of infected machines is plotted on the $y$-axis.

As can be seen in Fig. 2, the mode of the number of infected machines is roughly linear in the cure rate, for fixed $r$ and $n$.

### 3.2. Probability distribution approximations

As seen in Eq. (3), the stationary distribution of the birth-and-death process modeling the number of infected computers depends on the cure and infection rates only through their ratio $c/r$. (This is intuitively clear because $c$ and $r$ are both rates in the Markov process setting, and rescaling time would not change the stationary distribution.) Since $c$ is the cure rate per individual while $r$ is the infection rate from each infected individual, it is appropriate to consider $c/r$ as a function of $n$ to achieve a balance between the total infection and total cure rates for the population. In this section, we will discuss three approximations for the stationary distribution, which are valid in four different ranges of $c/r$. While we sketch the ideas of the proofs here, the details will be deferred to an Appendix.

#### 3.2.1. Small $c/r$ (Poisson approximation)

If the cure rate is sufficiently small relative to the infection rate, intuitively one would expect nearly all the population to be infected most of the time. This suggests that a Poisson distribution, often used to model occurrences of rare events, may be appropriate for the number of uninfected computers. The following result makes this precise.

Let $X_n$ and $Y_n = n - X_n$ denote the number of infected and uninfected computers respectively, when the process is in the invariant distribution $\{P_{i,n}\colon\ i = 0, 1, \ldots, n\}$. (We use the subscript $n$ to indicate the possible dependence upon the population size $n$.)

**Lemma 1.** *If* $\lim_{n\to\infty} c/r = b > 0$, *then* $\lim_{n\to\infty} P[Y_n = i] = (b^i e^{-b}/i!)\ \forall i = 0, 1, 2, \ldots$, *i.e.* $Y_n$ *asymptotically has a Poisson*$(b)$ *distribution.*

**Proof** (Sketch). Denoting the distribution of $Y_n$ by $\{Q_{i,n}: i = 0, 1, \ldots, n\}$, we have

$$Q_{i,n} = P_{n-i,n} = \frac{1}{Z_n} \frac{a}{c} \frac{1}{n-i} \left(\frac{r}{c}\right)^{n-i-1} \frac{(n-i)!}{i!}$$

for $i = 0, 1, \ldots, n$, with $Q_{n,n} = P_{0,n} = 1/Z_n$, where $Z_n$ is a normalizing factor. It is useful to express the other $Q_{i,n}$ as multiples of $Q_{0,n}$ as

$$Q_{i,n} = \frac{n}{n-i} \left(\frac{c}{r}\right)^{i} \frac{1}{i!} Q_{0,n}, \tag{7}$$

where $Q_{0,n}$ now plays the role of a normalizing factor. Without the factor $n/(n-i)$ we would have $Q_{0,n} = \mathrm{e}^{-c/r}$ exactly, giving us the Poisson$(c/r)$ distribution. Since $n/(n-i) \to 1$ as $n \to \infty$ for each fixed $i = 0, 1, \ldots$, and $c/r \to b$, it suffices to show that $Q_{0,n} \to \mathrm{e}^{-b}$. This is accomplished in the Appendix, using a geometric series to bound the tail probability and thus the normalizing factor. □

The approximation by a Poisson distribution leads to the interpretation that it is a rare event that a particular computer is not infected, when $c$ and $r$ are fixed. This would correspond to a particularly virulent disease, and does not correspond well to observations of computer virus infections, in which a small proportion of the population is infected at a given time.

Given the Poisson approximation result when $c$ and $r$ are constant, it is clear that we must consider cases where $c$ and $r$, or at least the ratio $c/r$, depends on the population size $n$, to achieve a suitable balance between infection and curing processes. This is intuitive if one recalls that the cure rate $c$ is per individual, while the total infection rate per individual is a multiple of $r$.

Remark: For completeness, we note that if $c/r \to 0$ as $n \to \infty$, a simpler analysis than above shows that the probability that all computers are infected converges to one.

### *3.2.2. Moderate c/r (Normal approximation)*

To achieve a better balance between the cure rate and total infection rate, we consider the case where $\lambda_n = c/r \to \infty$ as $n \to \infty$. The parameter $b$ appeared as the mean of the approximating Poisson distribution in the case where $c/r$ was asymptotically constant. Since the Poisson distribution becomes asymptotically Normal as the mean converges to infinity, it is natural to expect that the number of uninfected computers among $n$ total computers, denoted here by $Y_n$, has an approximate Normal distribution for some rates of $\lambda_n = c/r \to \infty$. The following proof is based on converting the Poisson approximation into a Normal approximation, under the condition that $c/r = \mathrm{o}(n)$ as $n \to \infty$.

**Lemma 2.** *If* $\lambda_n = \mathrm{o}(n)$, *then* $\lim_{n\to\infty} P[a < (Y_n - \lambda_n)/\sqrt{\lambda_n} < b] = \Phi(b) - \Phi(a)$ *for all* $-\infty < a < b < \infty$, *where* $\Phi$ *denotes the standard Normal cumulative distribution function.*

**Proof** (Sketch). For convenience, we denote the frequency function of the Poisson($\lambda_n = c/r$) distribution by $\{\pi_{i,n}: i = 0, 1, 2, \ldots\}$. We may then calculate

$$P\left[a < \frac{Y_n - \lambda_n}{\sqrt{\lambda_n}} < b\right] = P[\lambda_n + a\sqrt{\lambda_n} < Y_n < \lambda_n + b\sqrt{\lambda_n}]$$

$$= \sum_{i=\lambda_n + a\sqrt{\lambda_n}}^{\lambda_n + b\sqrt{\lambda_n}} Q_{i,n}$$

$$= \sum_{i=\lambda_n + a\sqrt{\lambda_n}}^{\lambda_n + b\sqrt{\lambda_n}} \frac{n}{n-i} \left(\frac{c}{r}\right)^i \frac{1}{i!} Q_{0,n}.$$

Under our hypothesis, however, $Q_{0,n} \to 0$, but we can again bound the sum of probabilities to show that

$$e^{c/r} Q_{0,n} = 1 + o(1).$$

Thus, we have

$$P\left[a < \frac{Y_n - \lambda_n}{\sqrt{\lambda_n}} < b\right] = (1 + o(1)) \sum_{i=\lambda_n + a\sqrt{\lambda_n}}^{\lambda_n + b\sqrt{\lambda_n}} \frac{n}{n-i} \pi_{i,n}.$$

Since $\lambda_n + b\sqrt{\lambda_n} = o(n)$, then $1 \leqslant n/(n-i) \leqslant 1 + o(1)$ for all $i \leqslant \lambda_n + b\sqrt{\lambda_n}$, so

$$P\left[a < \frac{Y_n - \lambda_n}{\sqrt{\lambda_n}} < b\right] = (1 + o(1)) \sum_{i=\lambda_n + a\sqrt{\lambda_n}}^{\lambda_n + b\sqrt{\lambda_n}} \pi_{i,n}.$$

However, as $n \to \infty$, the Poisson($\lambda_n = c/r$) distribution, normalized, is asymptotically Normal, so

$$\sum_{i=\lambda_n + a\sqrt{\lambda_n}}^{\lambda_n + b\sqrt{\lambda_n}} \pi_{i,n} \to P \quad [a < W < b],$$

where $W$ has a standard Normal distribution. □

This result shows that the number of infected computers is approximately Normal with mean $n - c/r$ and variance $c/r$, when $c/r \to \infty$, $c/r = o(n)$. However, in this case, we still have the proportion of the population that is infected, $(n - c/r)/n$, converging to 1.

### *3.2.3. Large c/r (Normal and logarithmic limit)*

We now consider the case when the cure rate increases roughly linearly with the population size. From the analysis of the mode, we see that typically a stable proportion of the population will be infected. In this situation, with approximately $\alpha n$ infected computers, $0 < \alpha < 1$, the population cure rate is approximately $c\alpha n$, while the population infection rate is approximately $r\alpha(1 - \alpha)n$. Stability can be achieved when these are balanced.

We consider $c/r = nd + o(n)$ as $n \to \infty$, and find that the distribution has quite different behavior when $d < 1$ than when $d > 1$.

**Lemma 3.** *If* $\lambda_n = nd + o(n)$, $d < 1$, *then* $\lim_{n\to\infty} P[a < (Y_n - \lambda_n)/\sqrt{\lambda_n} < b] = \Phi(b) - \Phi(a)$.

**Proof** (Sketch). As in the proofs of Lemmas 1 and 2, we consider the number of uninfected computers, viewing its distribution as closely related to the Poisson($\lambda_n$) distribution

$$Q_{i,n} = \frac{n}{n-i}\,\pi_{i,n}K_n,$$

where $K_n$ normalizes the sum to produce a probability distribution. For $d < 1$, we can choose $\varepsilon > 0$ such that $\varepsilon < d$ and $d + \varepsilon < 1$, and use Chernoff bounds to show that

$$P\left[|Y_n - nd| > \varepsilon n\right] \leqslant e^{-\rho n}$$

for some $\rho > 0$. Thus, nearly all the probability is concentrated near $nd$, where the factor $n/(n-i)$ is near $1/(1-d)$. This can be used to establish that the normalizing constant is approximately $1 - d$, so

$$P\left[a < \frac{Y_n - \lambda_n}{\sqrt{\lambda_n}} < b\right] = (1 + o(1)) \sum_{i=\lambda_n + a\sqrt{\lambda_n}}^{\lambda_n + b\sqrt{\lambda_n}} \pi_{i,n} \to \Phi(b) - \Phi(a)$$

as $\lambda_n \to \infty$ as in the proof of Lemma 2. □

**Lemma 4.** *If* $\lambda_n = nd + o(n)$, $d > 1$, *then*

$$\lim_{n\to\infty} P[X_n = k] = \frac{a/kd^{k-1}}{1 - \log(1 - 1/d)}, \quad k = 0, 1, \ldots .$$

**Proof** (Sketch). Note that

$$P_{k,n} = \frac{a}{k}\left(\frac{r}{c}\right)^{k-1} \frac{(n-1)!}{(n-k)!} \frac{1}{{}_3F_1(\{1,1,n-1\},\{2\},-r/c)}.$$

Under our hypothesis,

$$\left(\frac{r}{c}\right)^{k-1} \frac{(n-1)!}{(n-k)!} \to \left(\frac{1}{d}\right)^{k-1}.$$

Note that ${}_3F_1(\{1,1,n-1\},\{2\},-r/c)$ is a normalizing factor which is independent of $k$. Since the series of limiting probabilities converges in $k$ geometrically, $\lim_{n\to\infty} {}_3F_1(\{1,1,n-1\},\{2\},-r/c)$ is the normalizing constant for the limiting distribution, which can be computed by standard series convergence methods from calculus. □

This limiting distribution differs from the logarithmic distribution,

$$P[X = k] = \frac{\theta^k}{-k\log(1-\theta)} \quad k = 1, 2, \ldots ,$$

where $0 < \theta < 1$, due to the atom at zero (and necessary renormalization). The logarithmic distribution is discussed in Johnson et al., 1992, Chapter 7. It arises in a

model strongly related to ours, being the steady-state distribution of a birth and death process with rates $\lambda_i = \lambda i$, $i \leqslant 1$, and $\mu_i = \mu i$, $i > 1$, but $\mu_1 = 0$, a process appearing in Caraco (1979) in the context of animal group dynamics. We are not aware of these distributions arising in previous analyses of computer virus epidemic models.

## 4. Simulation

The National Computer Security Association reports a computer virus infection rate of approximately 35 infections per 1000 computers per month. See the web page at

www.webmastersecurity.com/ncsa97virusprevalencesurveya.htm

for the 1997 report. The ICSA computer virus prevalence report, available at

www.trusecure.com/html/tspub/pdf/vps20001.pdf,

reports slightly different, but similar, rates (see Table 1). The Sophos computer virus lab

www.sophos.com/virusinfo/whitepapers/prevention.html

reports that in the second quarter of 2000, approximately 800 new viruses were introduced each month, with over 50,000 known viruses in existence. Telcordia Technologies (www.netsizer.com) reports the number of machines on the Internet to be 127 million as of October 2001. Other sources of statistics on virus prevalence are available at http://www.securitystats.com/virusstats.asp and http://www.virusbtn.com/.

These reports, along with information on the propagation methods of known viruses, allow one to produce ballpark estimates for $n$, $r$ and $a$. One then would be interested in studying the properties of the system as a function of $c$. Numbers are available for various of these parameters for different intervals, allowing some determination of trends and rates of change.

Kephart and White (1993) report some virus prevalence data from 1991 which shows a relatively low level of virus infections (see Fig. 3). It should be noted that the spike in the middle is interpreted to be the result of a higher level of vigilance and reporting, as opposed to an actual increase in the number of viruses.

Some simulations can illustrate the model. In order to simulate this process, we take a two step approach. First, we draw a random time, $t \sim \exp(r + c)$, corresponding to the time until the next event. Then we flip a (biased) coin to determine if the event was an infection or a cure: we draw a uniform random variable $y$ and test if $y < R_{n,m}/(R_{n,m}+C_{\mathrm{m}})$ (infection) or $y > R_{n,m}/(R_{n,m}+C_{\mathrm{m}})$ (cure), where $R_{n,m}=r(n-m)m$ and $C_{\mathrm{m}} = cm$. In the event that $m = 0$, we set $m$ to 1 at time $t \sim \exp a$ later.

Table 1
Rate of infection per 1000 computers in the first two months of each year

| Year | Rate of infection |
|---|---|
| 1996 | 10 |
| 1997 | 21 |
| 1998 | 32 |
| 1999 | 80 |
| 2000 | 91 |

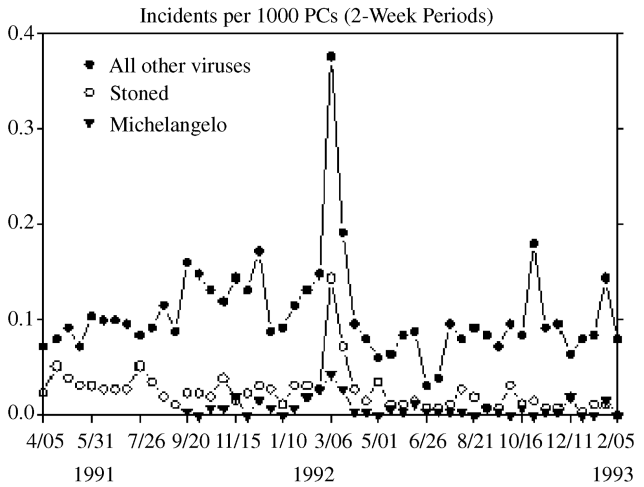


Fig. 3. Number of virus incidents reported per 1000 PCs for two specific viruses, Stoned (open circles) and Michelangelo (triangles), and all other viruses (closed circles) during two week periods ending with the indicated date. From Kephart and White (1993), with permission (© 1993 IEEE).

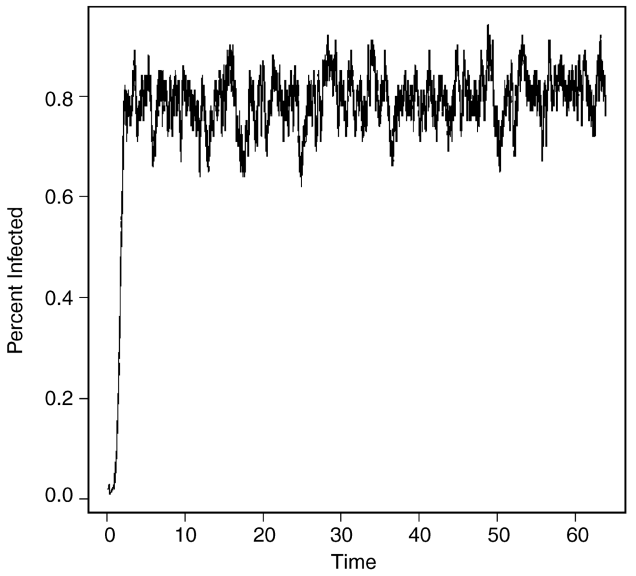


Fig. 4. A simulation run with parameters $a = 1$, $r = 0.05$, $c = 1$ and $n = 100$.

Set $a = 1$, $r = 0.05$, $c = 1$ and $n = 100$. In this case, Eq. (6) gives an estimate for the mode of 79.75 infected computers. Note that since $k$ is an integer, this says that the mode falls at 80 for these parameters. Fig. 4 shows the time progression of the simulation. For these parameters Mathematica gives a value of $P_0 = 7 \times 10^{-34}$, while

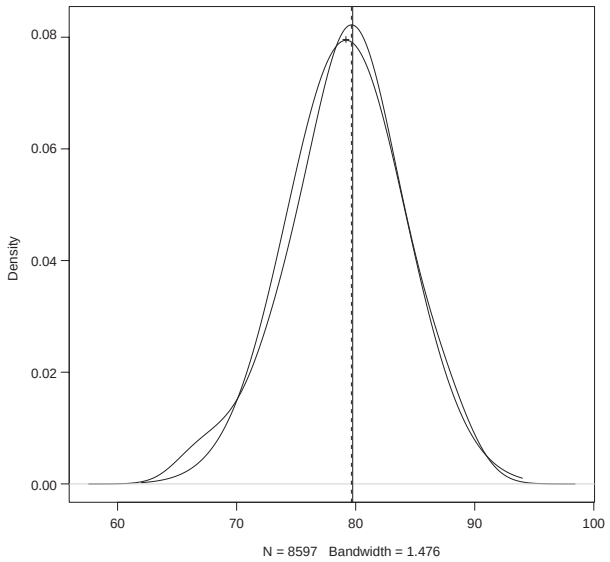


Fig. 5. A kernel estimator fit to the final 8536 observations of the simulation depicted in Fig. 4. The estimate for the mode given in Eq. (6) is depicted by a solid line, with the mode of the kernel estimator indicated by a dotted line. A Normal density with parameters equal to the sample mean and standard deviation for these data is depicted as a dotted curve. A "+" is drawn at the mode of the Normal density.

Eq. (5) gives a value of $1.7 \times 10^{-34}$. Thus the approximation agrees to an order of magnitude with the "exact" value provided by Mathematica (note that even in this "exact" calculation, the transcendental functions must be approximated numerically). Note that from a practical standpoint, these values indicate that the time to extinction for these values of the parameters is so large as to be of no practical consequence.

In order to estimate the mode of the limiting distribution, we consider the time steps from $t = 10$ on. A kernel estimator is fit to these data and plotted in Fig. 5. The mode of the kernel estimator is at 79.20, in agreement with the analysis.

Solving Eq. (2) for a maximum, using Mathematica's numerical optimization solver, results in a value of 80.25, in agreement with the above analysis (remembering that Mathematica is treating the function as a continuous one, while in reality it is discrete).

The mean and variance of the distribution are also available, either in closed form using Mathematica or numerically. For the parameters of the simulation we obtain a mean 79.75 (compared to the simulation value of 79.20) and a variance of 20.322 (compared to the simulation value of 25.213). Thus the theory is in close agreement with the simulation.

To illustrate the large $c$ behavior, Figs. 6–8 show the results of a simulation with $a = 1, r = 0.008$ and $c = 1$. This shows the tendency of the epidemic to die out with large $c$. The reintroduction rate is relatively large as well, causing the infection to start up again quickly after an extinction. Note that as many as 15% of the computers become infected with these rates. Fig. 8 shows the logarithmic distribution for the number of infected machines.

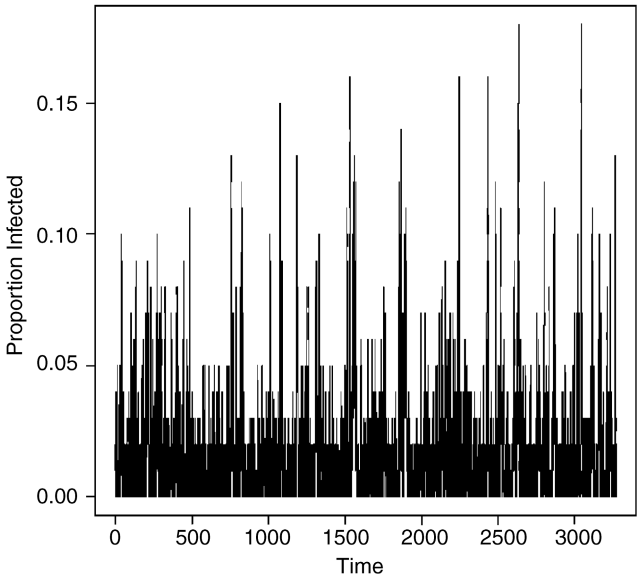


Fig. 6. A similar simulation to that in Fig. 4 with $a = 1,\ c = 1,\ r = 0.008$ and $n = 100$.

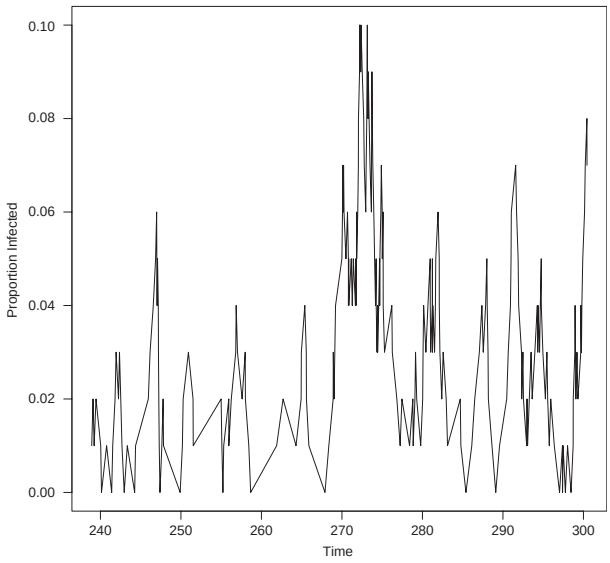


Fig. 7. A subset of the simulation of Fig. 6.

For small $c$ behavior, Figs. 9 and 10 show close agreement between the theoretical and simulation values for $Q_{i,n}$ of Eq. (7). This demonstrates the closeness of the Poisson approximation even for moderate values of $n$.

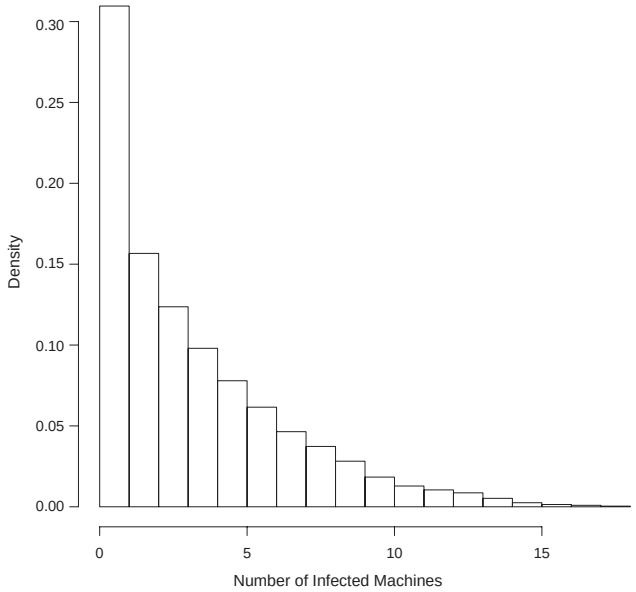


Fig. 8. A histogram of the number of infected machines parameters $c = 1$, $r = 0.008$, $a = 1$ and $n = 100$.

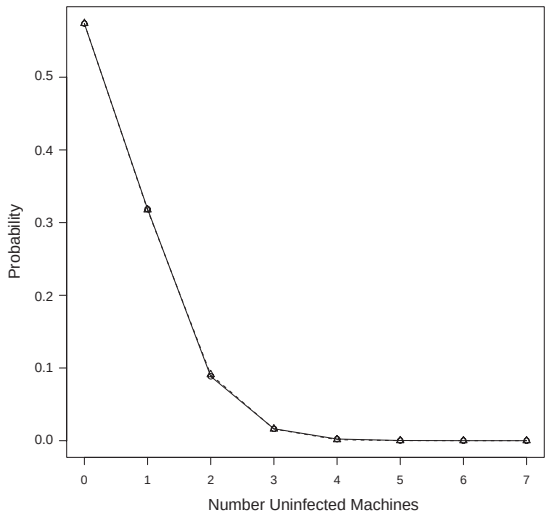


Fig. 9. $Q_{i,1000}$ for $a = 30$, $c = 0.01$ and $r = 0.018$. The theoretical values from Eq. (7) are depicted as a solid curve, with the simulation values depicted as a dotted curve. The curves essentially overlap in this figure.

## 5. Discussion

We proposed an epidemic model for the prevalence of computer viruses in a homogeneous closed population of computers. We determined the invariant distribution

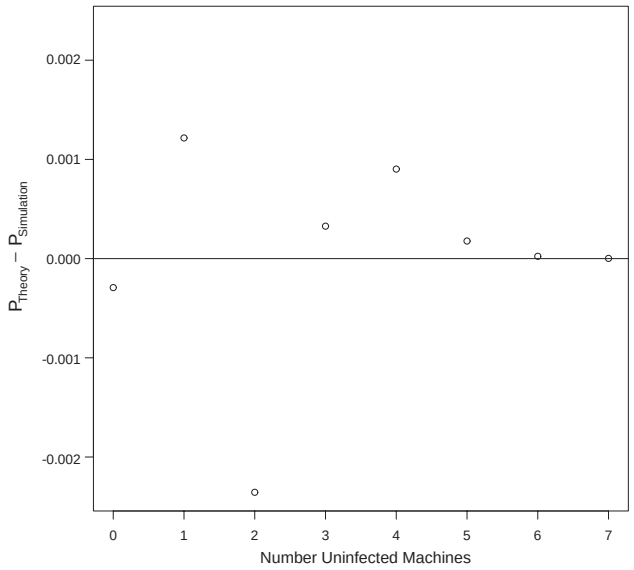


Fig. 10. The difference between the theoretical $Q_{i,1000}$ (Eq. (7)) and simulation values, from the simulation of Fig. 9. As in that simulation, $a = 0.03$, $c = 0.001$ and $r = 0.01$.

of the number of infected computers, and found approximations to this distribution for different ranges of parameters of the model. A key feature of the model is the existence of a threshold value of $c/r$ at which a dramatic change occurs in the distribution, from approximately logarithmic (which has a finite mean regardless of population size) to asymptotically normal (when properly standardized), corresponding to the mean increasing without bound as the population size increases. The threshold may be expressed as $c/r = n$. This implies that the ratio of cure rate to infection rate must be much larger to prevent a widespread infection in a large population than in a small population.

Due to the homogeneity assumed in the model, it is more applicable to local networks or clusters of computers, where reintroduction may correspond to infection by a virus from outside the local network. A model where the population increases in time might be more appropriate for the entire Internet.

A more realistic model might have different levels of subgroups of computers, which communicate among themselves at different rates. For example, a three level model might consist of departments within a company, companies, and the entire Internet. Different computers (or groups of computers, such as those with specific operating systems), might have different probabilities of becoming infected when exposed. Preventative action ("vaccination" or the installation of anti-virus software) could be taken into account. In the literature on epidemics, each of these effects has been considered in some model, but not all combined in any model for which rigorous mathematical analysis has been done.

## Acknowledgements

Professor Wierman gratefully acknowledges research support through the Navy-American Society for Engineering Education sabbatical program and the Acheson J. Duncan Fund for the Advancement of Research in Statistics. Dr. Marchette was supported through the Naval Surface Warfare Center Seeds program.

## Appendix

**Proofs.** To finish the proof of Lemma 1 we need to fill in the details of the Poisson convergence. To prove that $Q$ converges to a Poisson distribution, however, we must show that the $Q_{i,n}$ do not all converge to zero as $n \to \infty$ because they contain the normalizing factor $Z_n$.

Choose $I_n > c/r$ sufficiently large that $I_n! > n$. Then we may bound the tail probability by a geometric series to obtain

$$\sum_{i=I_n}^{n-1} Q_{i,n} \leqslant \left(\frac{c}{rI_n}\right)^{I_n} \frac{1}{1-(c/nI_n)} \frac{n}{I_n!}\, Q_{0,n}(1+\mathrm{o}(1)).$$

Note also that $Q_{n,n}/Q_{0,n} \to 0$. Thus,

$$\begin{aligned} Q_{0,n}\left[\sum_{i=0}^{I_n}\left(\frac{c}{r}\right)^i \frac{1}{i!}\right] &\leqslant 1 = \sum_{i=0}^{n} Q_{i,n} \\ &\leqslant Q_{0,n}\left[\frac{n}{n-I_n}\sum_{i=0}^{I_n}\left(\frac{c}{r}\right)^i \frac{1}{i!} + \mathrm{o}(1)\right] \\ &\leqslant Q_{0,n}\left[\frac{n}{n-I_n}\,\mathrm{e}^{c/r} + \mathrm{o}(1)\right]. \end{aligned}$$

Since $I_n! > n$ requires that $I_n \to \infty$, and we may choose $I_n$ to be $\mathrm{o}(n)$, asymptotically we have

$$\mathrm{e}^{b-\varepsilon}\limsup_{n\to\infty} Q_{0,n} \leqslant \mathrm{e}^{c/r}\limsup_{n\to\infty} Q_{0,n} \leqslant 1 \leqslant \mathrm{e}^{c/r}\liminf_{n\to\infty} Q_{0,n} \leqslant \mathrm{e}^{b+\varepsilon}\liminf_{n\to\infty} Q_{0,n}$$

for each $\varepsilon > 0$, so $\lim_{n\to\infty} Q_{0,n}$ exists and is equal to $\mathrm{e}^{-b}$. This implies that $\lim_{n\to\infty} Q_{i,n}$ exists and is equal to

$$b^i \frac{1}{i!}\, e^{-b} \quad \text{for } i = 1,2,\ldots\ . \qquad \square$$

To complete the proof of Lemma 2, we now verify our assumption, $Q_{o,n}/\mathrm{e}^{-c/r} = 1 + \mathrm{o}(1)$. First, note that

$$1 = \sum_{i=0}^{n} Q_{i,n} \geqslant Q_{0,n} \left[ \sum_{i=0}^{n-1} \left(\frac{c}{r}\right)^i \frac{1}{i!} \right],$$

so

$$\begin{aligned} Q_{o,n} &\leqslant \frac{1}{\sum_{i=0}^{n-1} (c/r)^i \frac{1}{i!}} \\ &= \frac{1}{\mathrm{e}^{c/r} - \sum_{i=n}^{\infty} (c/r)^i 1/i!} \\ &\leqslant \frac{1}{\mathrm{e}^{c/r} - \frac{(c/r)^n}{n!}\mathrm{e}^{c/r}} \\ &= \frac{\mathrm{e}^{-c/r}}{1 - \frac{(c/r)^n}{n!}}. \end{aligned}$$

By Stirling's Formula

$$\frac{(c/r)^n}{n!} \approx \frac{(c/r)^n}{\sqrt{2\pi} n^{n+\frac{1}{2}} \mathrm{e}^{-n}} = \frac{1}{\sqrt{2\pi}} \left(\frac{ce}{r}\right)^n n^{-n-\frac{1}{2}} = \mathrm{O}(n^{-\frac{1}{2}}),$$

if $ce/r < n$, or, equivalently $c/r \leqslant n/e$. Then

$$\frac{Q_{0,n}}{\mathrm{e}^{-c/r}} \leqslant \frac{1}{1 - B/\sqrt{n}} \leqslant 1 + \frac{A}{\sqrt{n}}, \tag{8}$$

for some $A, B > 0$.

For a bound in the other direction

$$\begin{aligned} 1 &= \sum_{i=0}^{n} Q_{i,n} \\ &= \sum_{i=0}^{I_n} (c/r)^i \frac{1}{i!} Q_{0,n} + \sum_{i=I_n+1}^{n-1} Q_{i,n} + Q_{n,n} \end{aligned} \tag{9}$$

for some $I_n > c/r$. For $I_n = \mathrm{o}(n)$,

$$\begin{aligned} \sum_{i=0}^{I_n} \left(\frac{c}{r}\right)^i \frac{1}{i!} Q_{0,n} &\leqslant \left(\frac{n}{n - I_n}\right) Q_{0,n} \sum_{i=0}^{I_n} \left(\frac{c}{r}\right)^i \frac{1}{i!} \\ &= (1 + \mathrm{o}(1)) Q_{0,n} \sum_{i=0}^{\infty} \left(\frac{c}{r}\right)^i \frac{1}{i!} \\ &= (1 + \mathrm{o}(1)) Q_{0,n} \mathrm{e}^{c/r}. \end{aligned}$$

For the second term in Eq. (9),

$$\sum_{i=I_n+1}^{n-1} Q_{i,n} = \sum_{i=I_n+1}^{n-1} \frac{n}{n-1} \left(\frac{c}{r}\right)^i \frac{1}{i!} Q_{0,n}.$$

Note that for $i > I_n > c/r$, the factor $(c/r)^i/i!$ is decreasing.

As before, use Stirling's Formula

$$\frac{(c/r)^{I_n}}{I_n!} \approx \frac{(c/r)^{I_n}}{\sqrt{2\pi} I_n^{I_n+\frac{1}{2}} \mathrm{e}^{-I_n}} = \frac{1}{\sqrt{2\pi}} \left(\frac{ce}{r}\right)^{I_n} I_n^{-I_n-\frac{1}{2}}.$$

Choose $I_n \approx 2e(c/r)$, so $c/r \approx (1/2e)I_n$, then

$$\frac{(c/r)^{I_n}}{I_n!} \approx \frac{1}{\sqrt{2\pi}} \left(\frac{1}{2} I_n\right)^{I_n} I_n^{I_n-\frac{1}{2}} \leqslant \frac{D}{2^{I_n}}.$$

Then

$$\begin{aligned}
\sum_{i=I_n+1}^{n-1} Q_{i,n} &= Q_{0,n} \sum_{i=I_n+1}^{n-1} \frac{n}{n-1} \left(\frac{c}{r}\right)^i \frac{1}{i!} \\
&\leqslant Q_{0,n} (n - I_n)^n \frac{D}{2^{I_n}} \\
&\leqslant Q_{0,n} D n^2 2^{-2\mathrm{e}(c/r)} \\
&\leqslant Q_{0,n} E \mathrm{e}^{-2(c/r)}.
\end{aligned}$$

From Eq. (5), we have

$$P_0 = Q_{n,n} = \mathrm{o}(\mathrm{e}^{-c/r}).$$

Combining, we have

$$\begin{aligned}
1 &\leqslant (1 + \mathrm{o}(1)) Q_{0,n} \mathrm{e}^{c/r} + Q_{0,n} E \mathrm{e}^{-2c/r} + \mathrm{o}(\mathrm{e}^{-c/r}) \\
&\leqslant (1 + \mathrm{o}(1)) Q_{0,n} \mathrm{e}^{c/r},
\end{aligned}$$

so, from this and Eq. (8),

$$1 - \mathrm{o}(1) \leqslant \frac{Q_{0,n}}{\mathrm{e}^{-c/r}} \leqslant 1 + \frac{A}{\sqrt{n}}.$$

Thus we obtain asymptotic normality when $c/r \to \infty$ and $c/r = \mathrm{o}(n)$. □

For Lemma 4, we need to compute the normalizing constant. To evaluate this normalizing constant, we write the probabilities $(ad/r)(1/k)(r/d)^k$ as $(ad/r)(1/k)x^k$ using $x = r/d$, and sum $(1/k)x^k$ as follows. Notice that for $|x| < 1$,

$$\left(\sum_{k=1}^{\infty} \frac{1}{k} x^k\right)' = \sum_{k=1}^{\infty} x^{k-1} = \frac{1}{1-x} = (-\log(1-x))'.$$

So,

$$\sum_{k=1}^{\infty} \frac{1}{k} x^k = -\log(1-x) + C$$

and evaluating at $x = 0$ shows that $C = 1$. Thus

$$\lim_{n\to\infty} (c + aF_n) = \frac{ad}{r}\left(1 - \log\left(1 - \frac{r}{d}\right)\right)$$

if $r < d$, and

$$\lim_{n\to\infty} P_k = \frac{(1/k)(\frac{r}{d})^k}{1 - \log(1 - r/d)}.$$

The mean infected population size in the limiting distribution is

$$\frac{\sum_{k=1}^{\infty} (r/d)^{k-1}}{1 - \log(1 - r/d)} = \frac{1}{(1 - r/d)(1 - \log(1 - r/d))}$$

$$= \frac{d}{(d - r)(1 - \log(1 - r/d))}, \tag{10}$$

which is finite for $r < d$. Similarly, all moments are finite when $r < d$. However, for $r > d$, the mean infected population size tends to infinity as $n \to \infty$. □

## References


Abreu, E.M., 2001. Computer virus costs reach $10.7b this year. The Washington Post, Sept 1, 2001. available at http://www.washtech.com/news/netarch/12267-1.html.

Ball, F., 1983a. The threshold behavior of epidemic models. J. Appl. Probab. 20, 227–241.

Ball, F., 1983b. A threshold theorem for the Reed-Frost chain-binomial epidemic. J. Appl. Probab. 20, 153–157.

Ball, F., 1999. Stochastic and deterministic models for SIS epidemics among a population partitioned into households. Math. Biosci. 156, 41–67.

Ball, F., Donnelly, P., 1995. Strong approximations for epidemic models. Stoch. Proc. Appl. 55, 1–21.

Bartholomew, D.J., 1976. Continuous time diffusion models with random duration of interest. J. Math. Sociology 4, 187–199.

Caraco, T., 1979. Ecological response of animal group size frequencies. International Co-operative Publishing House, Fairland, MD, pp. 371–386.

Cavender, J.A., 1978. Qausi-stationary distributions of birth-and-death processes. Adv. Appl. Probab. 10, 570–586.

Darroch, J.N., Seneta, E., 1967. On quasi-stationary distributions in absorbing continuous-time finite markov chains. J. Appl. Probab. 4, 192–196.

Jacquez, J.A., Simon, C.P., 1993. The stochastic SI model with recruitment and deaths. I. comparison with the closed SIS model. Math. Biosci. 117, 77–125.

Johnson, N.L., Katz, S., Kemp, A.W., 1992. Univariate Discrete Distributions. Wiley, New York.

Kendall, D.G., 1956. Deterministic and stochastic epidemics in closed populations. In: J. Neyman (Ed.), Third Berkeley Symposium on Mathematical Statistics and Probability, Vol. 4. University of California Press, California, pp. 149–165.

Kephart, J.O., White, S.R., 1991. Directed–graph epidemiological models of computer viruses. In: 1991 IEEE Computer Society Symposium on Research in Security and Privacy, Oakland, California, pp. 343–359.

Kephart, J.O., White, S.R., 1993. Measuring and modeling computer virus prevalence. In: 1993 IEEE Computer Society Symposium on Research in Security and Privacy, Oakland, California, pp. 2–15.

Kephart, J.O., White, S.R., Chess, D.M., 1993. Computers and epidemiology. IEEE Spectrum 30, 20–26.
Kryscio, R.J., Lefèvre, C., 1989. On the extinction of the SIS stochastic logistic epidemic. J. Appl. Probab. 27, 685–694.
Martin-Löf, A., 1986. Symmetric sampling procedures, general epidemic processes, and their threshold limit theorems. J. Appl. Probab. 23, 265–282.
Murray, W., 1988. The application of epidemiology to computer viruses. Comput. Security 7, 139–150.
Nåsell, I., 1996. The quasi-stationary distribution of the closed endemic SIS model. Adv. Appl. Probab. 28, 895–932.
Nåsell, I., 1999. On the time to extinction in recurrent epidemics. J. R. Statist. Soc. B 61, 309–330.
Oppenheim, I., Shuler, K.E., Weiss, G.H., 1977. Stochastic theory of nonlinear rate processes with multiple stationary states. Physica A 88, 191–214.
Ross, R., 1915. Some a Priori Pathometric Equations. Br. Med. J. 1, 546.
Ross, S., 1996. Stochastic Processes. Wiley, New York.
Scalia-Tomba, G., 1985. Asymptotic final size distribution for some chain-binomial processes. Adv. Appl. Probab. 17, 477–495.
von Bahr, B., Martin-Löf, A., 1980. Threshold limit theorems for some epidemic processes. Adv. Appl. Probab. 12, 319–349.
Weiss, G.H., Dishon, M., 1971. On the asymptotic behavior of the stochastic and deterministic models of an epidemic. Math. Biosci. 11, 261–265.
Wolfram, S., 1996. The Mathematica Book, 3rd Edition. Cambridge University Press, New York.