

CRITICAL SCALING FOR THE SIS STOCHASTIC EPIDEMIC

R. G. DOLGOARSHINNYKH,* *Columbia University*

STEVEN P. LALLEY,** *University of Chicago*

Abstract

We exhibit a scaling law for the critical SIS stochastic epidemic. If at time 0 the population consists of \sqrt{N} infected and $N - \sqrt{N}$ susceptible individuals, then when the time and the number currently infected are both scaled by \sqrt{N} , the resulting process converges, as $N \rightarrow \infty$, to a diffusion process related to the Feller diffusion by a change of drift. As a consequence, the rescaled size of the epidemic has a limit law that coincides with that of a first passage time for the standard Ornstein–Uhlenbeck process. These results are the analogs for the SIS epidemic of results of Martin-Löf (1998) and Aldous (1997) for the simple SIR epidemic.

Keywords: Stochastic epidemic model; SIS; SIR; Feller diffusion; Ornstein–Uhlenbeck process

2000 Mathematics Subject Classification: Primary 60K30
Secondary 92D30

1. Introduction

Among the most thoroughly studied stochastic epidemic models are the simple SIR and SIS epidemics (see [13] for the origin of the SIS model); among the many problems associated with these models, perhaps the most basic and most interesting are to do with the duration and size of the epidemic. When the epidemic is either subcritical or supercritical, the large-population behavior of the duration and size is reasonably well understood: in the subcritical case the epidemic is stochastically dominated by a subcritical Galton–Watson process (see below), and in the supercritical case the epidemic may become *endemic* (see [7] and [12]). For critical epidemics, large-population asymptotics are more delicate. It was shown in [9] (see also [1]) that the size $S = S_N$ has an interesting and nontrivial asymptotic behavior as the population size N tends to ∞ . If the number X_0 of individuals initially infected is of order $bN^{1/3}$, then the total infection time S_N (i.e. the sum of the total infection times over the whole population) has a limit distribution

$$\frac{S_N}{N^{2/3}} \xrightarrow{D} T_b^*, \quad (1)$$

where T_b^* is the first passage time of $W(t) + t^2/2$ to the level b , $W(t)$ is a standard Wiener process, and ‘ \xrightarrow{D} ’ denotes convergence in distribution. Furthermore, the critical exponent is equal to $\frac{1}{3}$, i.e. the quadratic drift is not felt if X_0 is much smaller than $N^{1/3}$. If $X(0) \sim bN^\alpha$

Received 17 January 2006.

* Postal address: Department of Statistics, Columbia University, New York, NY 10027, USA.

Email address: regina@stat.columbia.edu

** Postal address: Department of Statistics, University of Chicago, Chicago, IL 60637, USA.

Email address: lalley@galton.uchicago.edu

for some $\alpha < \frac{1}{3}$ then $S_N/N^{2\alpha}$ converges in law to the first passage time of $W(t)$ to the level b . (This is neither proved nor stated in [9], but it can be deduced from the methods used there.)

The purpose of this note is to establish an analogous scaling law for the critical SIS epidemic. If the number X_0 of individuals initially infected is of order $b\sqrt{N}$ then, as $N \rightarrow \infty$,

$$\frac{S_N}{N} \xrightarrow{D} \tau_b,$$

where τ_b is the time of first passage to 0 of a standard Ornstein–Uhlenbeck process started at b . Our approach has a rather different character than those in [1] and [9]. We shall establish that the SIS epidemic process itself, suitably rescaled, converges in law to a diffusion process we call the *attenuated Feller diffusion*. This has a law absolutely continuous with respect to, but not equal to, that of the standard Feller diffusion. (See (4), below, for the stochastic differential equation governing the Feller diffusion; see (7), below, for the attenuated Feller diffusion.) Furthermore, we will show that the critical exponent is equal to $\frac{1}{2}$ in the following sense. If $X(0) \sim bN^\alpha$ for some $\alpha < \frac{1}{2}$ then the rescaled SIS process converges in law to a standard Feller diffusion with *no* drift. It will follow that the duration of the epidemic, rescaled by $\sqrt{X(0)}$, converges in law as $N \rightarrow \infty$ to the first passage time to 0 of the corresponding Feller diffusion.

Our analysis will show that the critical *scaling window* for the transmissivity parameter β (see below) is of order $1/\sqrt{N}$. This scaling window has also been observed – but in a different context – in [10] and [11], where it was shown that the quasistationary distribution of the SIS epidemic undergoes a scaling transition when the transmissivity parameter varies from below $1 - O(N^{-1/2})$ to above $1 + O(N^{-1/2})$. This phenomenon does not seem to be directly linked to the critical scaling in our Theorem 1.

2. The SIS epidemic and its branching envelope

2.1. The SIS model

The SIS epidemic is a continuous-time birth–death Markov chain $X_t = X(t)$ on the state space $[N] := \{0, 1, 2, \dots, N\}$ whose infinitesimal transition probabilities are as follows:

$$\begin{aligned} P\{X(t + \delta t) = x + 1 \mid X(t) = x\} &= \beta x \left(1 - \frac{x}{N}\right) \delta t + o(x\delta t), \\ P\{X(t + \delta t) = x - 1 \mid X(t) = x\} &= x\delta t + o(x\delta t), \\ P\{X(t + \delta t) = x \mid X(t) = x\} &= 1 - \beta x \left(1 - \frac{x}{N}\right) \delta t - x\delta t + o(x\delta t). \end{aligned} \tag{2}$$

These probabilities describe a population of N individuals in which X_t are infected and the remainder, $N - X_t$, are susceptible to infection at time t . Infected individuals recover at rate 1, after which they once again become susceptible to (re-)infection, and susceptible individuals become infected at rate $\beta X_t/N$ proportional to the number of infected individuals in the population. The epidemic ends at the first time $T = T_N = t$ when $X_t = 0$ (note that state 0 is absorbing). The epidemic is said to be *critical* when $\beta = 1$, and *nearly critical* when $\beta - 1 = O(\sqrt{N})$.

2.2. The branching envelope

When the number of individuals infected is small compared to the population size, the epidemic evolves approximately as a continuous-time branching process $Z(t) = Z_t$ with

infinitesimal transition probabilities

$$\begin{aligned}
 P\{Z(t + \delta t) = z + 1 \mid Z(t) = z\} &= \beta z \delta t + o(z \delta t), \\
 P\{Z(t + \delta t) = z - 1 \mid Z(t) = z\} &= z \delta t + o(z \delta t), \\
 P\{Z(t + \delta t) = z \mid Z(t) = z\} &= 1 - (\beta + 1)z \delta t + o(z \delta t).
 \end{aligned}
 \tag{3}$$

We shall refer to this process as the *branching envelope* of the SIS process. Observe that the death rate x is the same as for the SIS epidemic, but that the birth rate βx dominates the birth rate $\beta x(1 - x/N)$ of the SIS process; the difference, $\beta x^2/N$, will be called the *attenuation* or *attrition rate*. It is possible, by a standard construction, to build the SIS process $X(t)$ and its branching envelope $Z(t)$ on the same probability space in such a way that $X(0) = Z(0)$ and $X(t) \leq Z(t)$ for all $t \geq 0$. Thus, the size and duration of the SIS epidemic are stochastically dominated by the total progeny and extinction time of the branching envelope.

2.3. Critical scaling for the branching envelope

It was proved in [4] that a critical branching process, when properly renormalized, behaves approximately as a *Feller diffusion* with drift λY_t , that is, a solution to the stochastic differential equation

$$dY_t = \lambda Y_t dt + \sqrt{Y_t} dW_t, \tag{4}$$

where W_t is a standard Wiener process. (Equivalently, the Feller diffusion with drift parameter λ may be described as the diffusion process on $[0, \infty)$ with infinitesimal generator $\mathcal{G}^\lambda = \lambda x \partial_x + x \partial_{xx}^2/2$.) Feller’s theorem (see [6] or [8] for a proof) asserts that if $Z^m(t)$ is a sequence of branching processes satisfying (3) with $\beta = \beta_m = 1 + \lambda/m$ and with $Z^m(0) \sim bm$ for some $b > 0$ and $\lambda \in \mathbb{R}$, then

$$\frac{Z^m(mt)}{m} \xrightarrow{D} Y_t, \tag{5}$$

where Y_t is the Feller diffusion with drift parameter λ and initial value $Y_0 = b$.

2.4. Critical scaling for the SIS process

Because the branching envelope stochastically dominates the SIS process, the scaling law (5) limits the duration and growth of the critical and nearly critical SIS process. Since time is scaled by the factor m , where $Z^m(0) \sim bm$, it follows that the corresponding SIS process started with $X(0) \sim bN^\alpha$ infected individuals cannot have duration longer than $O_P(N^\alpha)$ time units. Consequently, we should expect that if the attenuation rate, divided by the scale factor N^α and integrated to time N^α , is $o_P(1)$ then the limiting behavior of the rescaled SIS process $X(N^\alpha t)/N^\alpha$ should be no different from that of the branching envelope $Z(N^\alpha t)/N^\alpha$. An easy calculation shows that this will be the case when $\alpha < \frac{1}{2}$. When $\alpha = \frac{1}{2}$, the accumulated attrition over the duration of the branching envelope will be of the same order of magnitude as the fluctuations, and so the rescaled SIS process should have a genuinely different asymptotic behavior from the branching envelope. Our main result makes this precise.

Theorem 1. *Assume that the process $X(t) = X^N(t)$ has infinitesimal transition probabilities defined in (2). If, for some constants $\alpha \leq \frac{1}{2}$ and $b > 0$, the number of individuals initially infected satisfies $X^N(0) \sim bN^\alpha$ and if the birth rate (2) satisfies $\beta = \beta_N = 1 + \lambda/N^\alpha$, then, as $N \rightarrow \infty$,*

$$\frac{X^N(N^\alpha t)}{N^\alpha} \xrightarrow{D} Y_t, \tag{6}$$

where

- (a) if $\alpha < \frac{1}{2}$ then Y_t is a Feller diffusion with drift λ and initial state $Y_0 = b$,
- (b) if $\alpha = \frac{1}{2}$ then Y_t is an attenuated Feller diffusion with drift λ and initial state $Y_0 = b$; that is, Y_t is a solution to the stochastic differential equation

$$dY_t = (\lambda Y_t - Y_t^2) dt + \sqrt{Y_t} dW_t, \tag{7}$$

where W_t is a standard Wiener process.

Note that the *attenuation term* ‘ $-Y_t^2$ ’ in the drift of the limiting process (7) can be guessed from the form of the attrition $\beta x^2/N$. The proof of Theorem 1 is given in Section 4.

3. Size of the epidemic

The size of an epidemic is usually defined to be the total number ξ of new infections during its entire course. An asymptotically equivalent quantity is the *total infection time*, which is defined as follows:

$$S = S_N = \int_0^T X_t dt. \tag{8}$$

Although the two definitions are not the same, it can be shown that the two quantities have the same asymptotic behavior for large N , that is, $\xi_N \sim S_N$. (This follows from the fact that the lengths of the infection periods for infected individuals are independent and identically distributed unit exponential random variables.) Because the integral (8) is a continuous functional of the path X_t (relative to the Skorokhod topology), Theorem 1 implies that if $X(0) \sim b\sqrt{N}$ and $\beta = 1 + \lambda/\sqrt{N}$ then

$$\frac{S_N}{N} \xrightarrow{D} \int_0^{\tau(0)} Y_t dt,$$

where Y_t is the attenuated Feller diffusion (7) with initial state $Y_0 = b$ and τ_0 is the first passage time to 0 of Y_t .

By an odd bit of luck, the instantaneous rate $Y_t dt$ at which infection time accrues coincides with the rate of change in accumulated quadratic variation of the semimartingale Y_t . (In fact this is really no accident, but rather an artifact of the fundamental connection between Galton–Watson processes and random walks via the ‘depth-first search’ algorithm; see [1] for more details.) This suggests making the natural time change to the diffusion Y_t so as to make the instantaneous quadratic variation constant. The new time scale $s = s(t)$ and the old one, t , are related by

$$ds = Y_t dt,$$

and so $\int Y_t dt = \int ds$ is the limit of the rescaled total infection time S_N/N . The time-changed process $V_s = Y_{t(s)}$ satisfies the stochastic differential equation

$$dV_s = (\lambda - V_s) ds + d\tilde{W}_s,$$

where \tilde{W}_s is a standard Wiener process. Setting $U_s = V_s - \lambda$, we obtain the stochastic differential equation for the standard Ornstein–Uhlenbeck process as follows:

$$dU_s = -U_s ds + d\tilde{W}_s.$$

This proves the following corollary.

Corollary 1. *If $X(0) \sim b\sqrt{N}$ and $\beta = 1 + \lambda/\sqrt{N}$ then*

$$\frac{S_N}{N} \xrightarrow{D} \tau(b - \lambda; -\lambda),$$

where $\tau(x; y)$ is the time of first passage to y of a standard Ornstein–Uhlenbeck process started at x .

The Laplace transforms of the distributions of $\tau(x; y)$ can be expressed in terms of parabolic cylinder (i.e. Weber) functions; see [2]. These do not invert easily. However, in the special case $\lambda = 0$ (i.e. the case corresponding to the critical SIS epidemic), the distribution of $\tau(b; 0)$ has the following simple closed form:

$$P\{\tau(b; 0) > s\} = P^b\{U_s > 0\} - P^b\{U_s < 0\}.$$

This can be obtained from a reflection principle, using the symmetry of the Ornstein–Uhlenbeck process about the origin.

4. Proof of Theorem 1

We prove Theorem 1 using the weak convergence machinery developed in [3], which reduces the problem to checking convergence, in an appropriate sense, of infinitesimal generators. Let

$$Y_t^N = \frac{X^N(N^\alpha t)}{N^\alpha}$$

be the rescaled epidemic process and denote by E_y^N the corresponding expectation operator under the initial condition $Y_0^N = y$. Let $\hat{C}[0, \infty)$ be the space of bounded continuous functions on the time interval $[0, \infty)$, let $C^2(0, \infty)$ be the space of twice continuously differentiable functions on $(0, \infty)$, and let $C_c^\infty(0, \infty)$ be the space of infinitely differentiable functions with compact support on $(0, \infty)$. For $f \in \hat{C}[0, \infty)$, define

$$\mathcal{G}^N f(y) = \lim_{h \rightarrow 0} \frac{E_y^N[f(Y_h) - f(y)]}{h}$$

and

$$\mathcal{G}f(y) = \begin{cases} (\lambda y) \frac{\partial f}{\partial y}(y) + y \frac{\partial^2 f}{\partial y^2}(y) & \text{if } \alpha < \frac{1}{2}, \\ (\lambda y - y^2) \frac{\partial f}{\partial y}(y) + y \frac{\partial^2 f}{\partial y^2}(y) & \text{if } \alpha = \frac{1}{2}. \end{cases}$$

By [3, Section 11.2, Corollary 1.2] the operator \mathcal{G} restricted to $\hat{C}[0, \infty) \cap C^2(0, \infty)$ generates a Feller semigroup on $\hat{C}[0, \infty)$, and [3, Section 1.5, Proposition 3.3] implies that $C_c^\infty(0, \infty)$ is a core for the generator. (An easy calculation, which we omit, shows that 0 is an exit boundary and ∞ is a natural boundary in both cases.) Moreover, the Markov processes determined by these Feller semigroups can be constructed so as to satisfy the stochastic differential equations (4) and (7), respectively. By [3, Theorem 2.5, p. 167] and [3, Theorem 6.1, p. 28], to prove convergence of (6) it is enough to show that for each f in the core of \mathcal{G} the generators converge in the sense of the following lemma.

Lemma 1. *Let $f \in C_c^\infty(0, \infty)$. Then we obtain*

$$\lim_{N \rightarrow \infty} \sup_{y \in [N]/N^\alpha} |\mathcal{G}^N f(y) - \mathcal{G}f(y)| = 0.$$

Proof. Consider first the case $\alpha = \frac{1}{2}$. The first step is to calculate $\mathcal{G}^N f$ for $f \in C_c^\infty[0, \infty)$. Using the infinitesimal transition probabilities (3), we have (with $x = yN^{1/2}$ and $h = tN^{1/2}$)

$$\begin{aligned} E_y^N[f(Y_h^N) - f(y)] &= [f(1 + N^{1/2}) - f(y)][(1 + \lambda N^{-1/2})yN^{1/2}(1 - yN^{-1/2})hN^{1/2}] \\ &\quad + [f(1 - N^{-1/2}) - f(y)][yN^{1/2}hN^{1/2}] + o(Nhy) \\ &= [f(1 + N^{-1/2}) - f(y)][(yN^{1/2} + \lambda y - y^2 - \lambda y^2 N^{-1/2})hN^{1/2}] \\ &\quad + [f(1 - N^{-1/2}) - f(y)][yN^{1/2}h\sqrt{N}] + o(Nhy). \end{aligned}$$

The error term $o(Nhy)$ is uniform in y because f is assumed to have compact support. Taking the limit of this expression as $h \rightarrow 0$ yields

$$\begin{aligned} \mathcal{G}^N f(y) &= [f(1 + N^{-1/2}) - f(y)][N^{1/2}(\lambda y - y^2 - \lambda y^2 N^{-1/2})] \\ &\quad + [(f(1 - N^{-1/2}) - f(y)) - (f(y) - f(y - N^{-1/2}))]Ny. \end{aligned}$$

Since $f \in C_c^\infty[0, \infty)$, there exists a constant $C > 0$ such that f and all its partial derivatives vanish for $y > C$. Therefore, uniformly in all $y \in [N]/\sqrt{N}$,

$$\lim_{N \rightarrow \infty} \mathcal{G}^N f(y) = \frac{\partial f}{\partial y}(y)(\lambda y - y^2) + \frac{\partial^2 f}{\partial y^2}(y)y = \mathcal{G}f(y).$$

A similar calculation establishes convergence of generators when $\alpha < \frac{1}{2}$.

5. The SIR epidemic revisited

The continuous-time SIR epidemic differs from the SIS epidemic in that individuals may only be infected once; upon recovery, individuals are effectively removed from the population. Thus, the state at any time t is determined by two variables, the number currently infected ($I(t) = I^N(t)$) and the number removed ($R(t) = R^N(t)$). These take values in the set of nonnegative integer pairs (i, r) such that $0 \leq i + r \leq N$, where N is the (original) population size. The instantaneous transition rates are as follows:

$$\begin{aligned} (i, r) &\mapsto (i - 1, r + 1) \quad \text{at rate } idt, \\ (i, r) &\mapsto (i + 1, r) \quad \text{at rate } \beta i(N - i - r) \frac{dt}{N}. \end{aligned} \tag{9}$$

All states (i, r) with $i = 0$ are absorbing; the epidemic ends the first time one of these states is visited.

As for the SIS epidemic, if the numbers of infected and removed individuals are small compared to the total population size N , then the second transition rate in (9) reduces to βidt , and so the process $I(t)$ evolves approximately as the branching process (3). Therefore, by the same logic as in Section 2.4, the limiting behavior of the epidemic can be deduced by examination of the accumulated attrition over the duration of the branching process. The result is as follows.

Theorem 2. *Assume that $(I^N(t), R^N(t))$ has instantaneous transition rates (9), and assume that $R^N(0) = 0$. If, for some $\alpha \leq \frac{1}{3}$ and $b > 0$, the number $I^{N(0)}$ of individuals initially infected satisfies $I^N(0) \sim bN^\alpha$ and if the birth rate β satisfies $\beta = 1 + \lambda/N^\alpha$, then, as $N \rightarrow \infty$,*

$$\begin{pmatrix} N^{-\alpha} I^N(t) \\ N^{-2\alpha} R^N(t) \end{pmatrix} \xrightarrow{D} \begin{pmatrix} I(t) \\ R(t) \end{pmatrix}, \tag{10}$$

where the limit process $(I(t), R(t))$ satisfies

$$dI(t) = \lambda I(t) dt + \sqrt{I(t)} dW_t,$$

$$dR(t) = I(t) dt,$$

if $\alpha = \frac{1}{3}$ and

$$dI(t) = (\lambda I(t) - I(t)R(t)) dt + \sqrt{I(t)} dW_t,$$

$$dR(t) = I(t) dt,$$

if $\alpha = \frac{1}{3}$.

Equation (1) can be easily recovered from Theorem 2 by the same device as used in Section 3. Define the new time scale s by $ds = I_t dt$ and the corresponding time-changed process $dJ(s) = dI(t)$. Then the total size of the epidemic is just the integral $\int ds$ up to the time of first passage to 0 of $J(s)$. But $J(s)$ is just the Wiener process with a quadratic drift, and so (1) follows.

Theorem 2 can be proved either by martingale methods or by use of the Ethier–Kurtz machinery. The latter approach is mildly complicated by the fact that the generator $\mathcal{G} = (\lambda i - ir)\partial_i + \sqrt{i}\partial_{ii} + i\partial_r$ is not elliptic, but rather parabolic, and singular along the $i = 0$ axis. The singularity at $i = 0$ can be handled by truncating the state space. To prove (10), it suffices to prove weak convergence for the processes $(I^N(t \wedge \tau_\varepsilon), R^N(t \wedge \tau_\varepsilon))$, where τ_ε is the time of first passage to the level $i = \varepsilon$. Nonellipticity of the generator may be handled by using standard existence results from the theory of parabolic partial differential equations (see [5]) to verify the hypotheses of the Hille–Yosida theorem (see [3, Theorem 2.2]). Weak convergence of the truncated processes may then be proved by checking convergence of generators; this is another routine calculation similar to that carried out for the SIS epidemic in Section 4.

Acknowledgement

S. Lalley was supported by the NSF (grant number DMS-0405102).

References

- [1] ALDOUS, D. (1997). Brownian excursions, critical random graphs and the multiplicative coalescent. *Ann. Prob.* **25**, 812–854.
- [2] DARLING, D. A. AND SIEGERT, A. J. F. (1953). The first passage problem for a continuous Markov process. *Ann. Math. Statist.* **24**, 624–639.
- [3] ETHIER, S. N. AND KURTZ, T. G. (1986). *Markov Processes. Characterization and Convergence*. John Wiley, New York.
- [4] FELLER, W. (1951). Diffusion processes in genetics. In *Proc. 2nd Berkeley Symp. Math. Statist. Prob., 1950*, University of California Press, Berkeley and Los Angeles, pp. 227–246.
- [5] FRIEDMAN, A. (1964). *Partial Differential Equations of Parabolic Type*. Prentice-Hall, Englewood Cliffs, NJ.
- [6] JIŘINA, M. (1969). On Feller’s branching diffusion processes. *Časopis Pěst. Mat.* **94**, 84–90, 107.
- [7] KRYSZCIO, R. J. AND LEFÈVRE, C. (1989). On the extinction of the S-I-S stochastic logistic epidemic. *J. Appl. Prob.* **26**, 685–694.
- [8] LINDVALL, T. (1974). On Feller’s branching diffusion processes. *Adv. Appl. Prob.* **6**, 309–321.
- [9] MARTIN-LÖF, A. (1998). The final size of a nearly critical epidemic, and the first passage time of a Wiener process to a parabolic barrier. *J. Appl. Prob.* **35**, 671–682.
- [10] NASELL, I. (1996). The quasi-stationary distribution of the closed endemic SIS model. *Adv. Appl. Prob.* **28**, 895–932.
- [11] NASELL, I. (1999). On the time to extinction in recurrent epidemics. *J. R. Statist. Soc. B* **61**, 309–330.
- [12] NORDEN, R. H. (1982). On the distribution of the time to extinction in the stochastic logistic population model. *Adv. Appl. Prob.* **14**, 687–708.
- [13] WEISS, G. AND DISHON, M. (1971). On the asymptotic behavior of the stochastic and deterministic models of an epidemic. *Math. Biosci.* **11**, 261–265.