Stochastic Processes and their Applications 11 (1981) 151–185
North-Holland Publishing Company

THE BASIC CONTACT PROCESSES*,**

David GRIFFEATH
Department of Mathematics, University of Wisconsin–Madison, Madison, WI 53706, U.S.A.

Received 25 September 1980

AMS (1970) Subject Class.: Primary 60 K 35
Contact process   interacting particle system
percolation   critical phenomenon

1. Introduction

Let S be the space of all subsets of the d-dimensional integer lattice Z^d. Certain continuous time Markov processes with state space S, known as interacting particle systems, have been studied extensively over the past decade (cf. [12, 24, 33, 37]). The basic (one-dimensional) contact process, introduced by Harris [17], is perhaps the simplest such interacting system which exhibits a critical phenomenon. The dynamics of the process can be described succinctly as follows. At any given time $t \geq 0$ certain sites $x \in Z$ are infected while the remainder are healthy. The set of infected sites is denoted ξ_t. Infected sites recover at constant exponential rate 1, while healthy sites are infected at an exponential rate proportional to the number of infected neighboring sites. Thus the infection rate at site x at each time t is 0, λ or 2λ depending on whether neither, one or both adjacent sites belong to ξ_t. Here λ is the infection parameter. In essence, the critical phenomenon is this: if λ is sufficiently small, infection tends to die out, whereas if λ is sufficiently large infection tends to be permanent. Interest in the model centers on the precise formulation of the dichotomy, and on detailed analysis of the ergodic properties of the process in both situations.

The present article is a more or less self-contained exposition of current knowledge concerning the basic contact process (ξ_t). In the next section we begin by constructing contact processes with the aid of Harris' graphical representations [19]. Duality equations, monotonicity and complete coupling properties are then established; these are the principal tools of the theory. To illustrate their use, we

* Supported by the National Science Foundation under grant MCS78-01241.
** Some of the results of this paper were announced at the Ninth Conference on Stochastic Processes and their Applications.

conclude Section 2 by formulating the essential qualitative aspects of the critical phenomenon.

Section 3 surveys the difficult quantitative problem: to determine the critical value λ_c below which infection dies out and above which infection persists. Lower bounds for λ_c are relatively easy to derive; we give an extremely simple argument that

$$\lambda_c \geq 1.$$

That $\lambda_c < \infty$, i.e. that infection can be permanent, is probably the deepest known result about contact processes. Three different proofs have been found. The first, due to Harris [17], gave no readily computable upper bound. The second, a remarkable argument by Holley and Liggett [21] shows that

$$\lambda_c \leq 2.$$

A third approach, less powerful but of wider applicability than the Holley–Liggett method, is due to Gray and Griffeath [8]; for the basic contact process it gives

$$\lambda_c < 7.$$

We sketch both the Holley–Liggett and Gray–Griffeath methods, since various estimates from those papers will be needed later in our study.

In Section 4 we begin analyzing the edge process, i.e. the rightmost (or leftmost) infected site when there is initially infection on $\{\ldots, -2, -1, 0\}$ ($\{0, 1, 2, \ldots\}$). Making use of results due to Durrett [5], Liggett [25] and the author [10, 11], we obtain a good understanding of the ergodic behavior of (ξ_t) at all parameter values except $\lambda = \lambda_c$. In particular, we prove a complete convergence theorem and a complete pointwise ergodic theorem.

The next two sections, 5 and 6, contain a number of new results. We have attempted to identify, insofar as possible, just what is known about the 'next level' of the ergodic theory, i.e. *rates* of convergence and mixing and velocity of the edge processes. As will be seen, the picture below λ_c is much more complete than that above λ_c. For instance, we are able to obtain exponential convergence to the state $\emptyset$ = 'all healthy' whenever $\lambda < \lambda_c$, but exponential convergence to the limiting equilibrium of permanent infection starting from Z = 'all infected' is only known for sufficiently large λ.

In Section 7 we address the critical contact process. Some partial results are discussed briefly. Then a result on convergence rates and sample path behavior is proved. As one would expect, the critical case must differ qualitatively from the subcritical case.

Approximation of contact processes by analogous models on the N-torus $S_N = \{0, 1, \ldots, N-1\}$ is the subject of Section 8. The latter models are finite Markov chains with $\emptyset$ absorbing; we consider the expected time e_N to absorption when the Nth process starts with infection everywhere on S_N. The finite approximations reflect the critical phenomenon of the limiting infinite system in terms of the growth rate of e_N as $N \to \infty$.

Section 9 deals with discrete time contact processes, and their connection with oriented percolation in the plane. In fact, discrete time contact processes were studied by Stavskaya and Piatetskii-Shapiro [35], Toom [38], Vasilev [42] and others several years before the continuous time theory emerged. Moreover, a totally equivalent model of oriented percolation in a quadrant of Z^2 was studied by Mauldon [27] and Bishir [1] almost twenty years ago. It is therefore satisfying that some of the results from Sections 4 and 6 apply equally well in discrete time to solve an open problem in oriented percolation (cf. [31]):

$$p_{\mathrm{T}} = p_{\mathrm{H}},$$

i.e. if the cluster of sites wetted by a source at the origin is finite with probability one, then the expected cluster size is finite. We conclude Section 9 by mentioning a two-dimensional growth model of Richardson [29] for which Durrett and Liggett [7] have shown that a discrete time contact process comes into play in an intriguing manner.

Finally, Section 10 addresses open problems and generalizations. We identify what seem to us the most important unresolved questions about (ξ_t). Foremost among these are the ergodicity of the critical processes and the 'explicit' evaluation of λ_c; both problems are probably quite difficult. We conclude the paper with a brief discussion of contact processes on Z^d, $d > 1$, where much less is known.

2. Graphical representation and elementary properties

Following Harris [19], we begin by constructing the basic contact processes from independent 'exponential alarm clocks' with the aid of a random graph $\mathscr{P} = \mathscr{P}(\lambda)$ called the *percolation substructure*. Start with the 'space-time diagram' $Z \times [0, \infty)$. For each $x \in Z$, draw three infinite sequences of graphical devices as follows. First draw arrows from $(x-1, \tau^1_{1,x})$ to $(x, \tau^1_{1,x})$, from $(x-1, \tau^2_{1,x})$ to $(x, \tau^2_{1,x})$, etc., where the values $\tau^1_{1,x}$, $\tau^2_{1,x} - \tau^1_{1,x}, \ldots$ are independent exponential random variables with mean λ. Second, draw arrows from $(x+1, \tau^1_{2,x})$ to $(x, \tau^1_{2,x})$, $(x+1, \tau^2_{2,x})$ to $(x, \tau^2_{2,x})$ etc., where the $\tau^n_{2,x}$ occur at rate λ. Finally, put down a sequence of δ's at $(x, \tau^1_{0,x})$, $(x, \tau^2_{0,x}), \ldots$, the $\tau^n_{0,x}$ occurring at rate 1. The arrows will transmit infection to site x if it is present at a neighboring site, while the δ's will kill infection if it is present at x. A generic realization of the graph $\mathscr{P}$ obtained in this manner is shown in Fig. 1. Say there is a path up from (y, s) to (x, t) in $\mathscr{P}$, $x, y \in Z$, $0 \leq s \leq t < \infty$, if there is a chain of upward vertical and directed horizontal edges in the resulting graph which leads from (y, s) to (x, t) without passing (vertically) through a δ. (By convention, there is a path up from (y, s) to (y, s).) Thus the δ's may be thought of as obstructions to the flow (or 'percolation') of liquid. Now define

$$\xi^A_t = \{x\text{: there is a path up from } (y, 0) \text{ to } (x, t) \text{ for some } y \in A\}, \qquad A \in S. \tag{1}$$

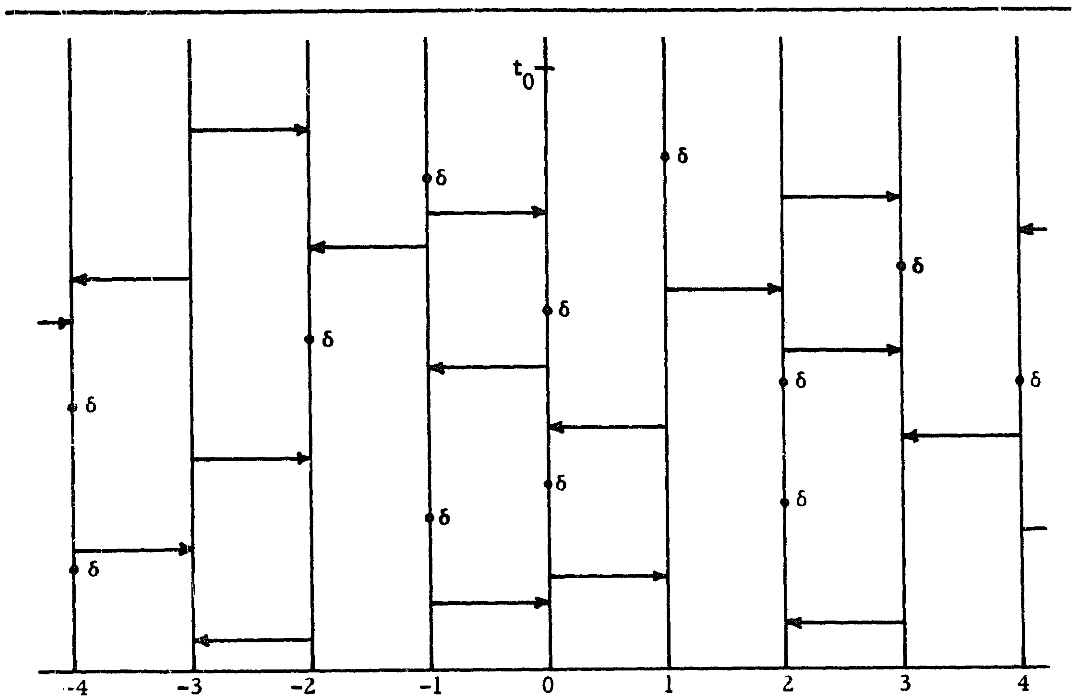

Fig. 1.

The process $(\xi_t^A)_{t\geq 0}$ is called the *basic contact process with parameter* λ *starting from* A (i.e. with infection initially on A.) The reader should check that (ξ_t^A) is Markov, and that its dynamics are precisely those described in the introduction.

Our first theorem will identify some of the key properties of contact processes which follow easily from the graphical representation. These properties will be the main tools in the analysis to be carried out. It is important to emphasize that (1) defines all the (ξ_t^A), $A \in S$, simultaneously on one probability space $(\Omega, \mathcal{F}, P)$, at least when λ is fixed. As we will soon see, processes with differing λ can also be defined on the same space by augmenting slightly the percolation substructure. Thus $\mathcal{P}$ 'couples' the evolutions of different contact processes; couplings of this sort turn out to be remarkably powerful. A few words about notation are in order here. When the role of the parameter λ is being stressed we write $\xi_t^A(\lambda)$. Also, some particularly important processes will be abbreviated:

$$\xi_t = \xi_t^Z, \qquad \xi_t^- = \xi_t^{(-\infty,0]}, \qquad \xi_t^+ = \xi_t^{[0,+\infty)}.$$

Finally, it will be convenient to write $\xi_t^x = \xi_t^{\{x\}}$ for the process starting with the single infected site x. (In general, we write x for $\{x\}$ whenever it is convenient.)

Theorem 1. *The following four properties hold*:

(i) *additivity*:

$$\xi_t^{A\cup B} = \xi_t^A \cup \xi_t^B, \quad A, B \in S, \quad t \geq 0; \tag{2}$$

(ii) *set monotonicity*:

$$\xi_t^A \subset \xi_t^B \text{ if } A \subset B, \quad t \geq 0; \tag{3}$$

(iii) λ*-monotonicity*: (*On a suitably enlarged probability space*),

$$\xi_t^A(\lambda_0) \subset \xi_t^A(\lambda) \text{ if } \lambda_0 < \lambda, \quad A \in S, \ t \geq 0; \tag{4}$$

(iv) *self-duality*:

$$\mathbf{P}(\xi_t^A \cap B \neq \emptyset) = \mathbf{P}(\xi_t^B \cap A \neq \emptyset), \quad A, B \in S, \ t \geq 0. \tag{5}$$

Proof. To get addivity simply notice that by definition (1), site x belongs to the set on either side of (2) if and only if there is a path up from $(A, 0)$ *or* $(B, 0)$ to (x, t) in $\mathscr{P}$.[1] Set monotonicity follows from (2) since, for $A \subset B$, we have

$$\xi_t^B = \xi_t^{A \cup (B-A)} = \xi_t^A \cup \xi_t^{B-A} \supset \xi_t^A.$$

A randomizing device can be used to establish λ-monotonicity as follows. Given $\lambda_0 < \lambda$, consider the graphical representation $\mathscr{P}(\lambda)$. Independently, color each arrow green with probability λ_0/λ and leave the arrow uncolored with probability $(\lambda - \lambda_0)/\lambda$. The green arrows together with the δ's in $\mathscr{P}(x)$ constitute a version of $\mathscr{P}(\lambda_0)$, so we can define $\xi_t^A(\lambda)$ by (1) and $\xi_t^A(\lambda_0)$ by

$$\xi_t^A(\lambda_0) = \{x : \exists \text{ path up from } (A, 0) \text{ to } (x, t) \text{ in } \mathscr{P}(\lambda) \text{ using only green arrows}\}.$$

From the construction, (4) is immediate. Self-duality is proved using a 'time/arrow reversal' trick. Fix λ, t, and consider the restriction of $\mathscr{P}(\lambda)$ to $Z \times [0, t]$. The key observation is that by letting time run 'down' instead of 'up', and by reversing the directions of all arrows, we get another substructure $\hat{\mathscr{P}}(\lambda)|_{Z \times [0,t]}$ with precisely the same law as $\mathscr{P}(\lambda)|_{Z \times [0,t]}$. (The piece of $\hat{\mathscr{P}}(\lambda)$ corresponding to the piece of $\mathscr{P}(\lambda)$ from Fig. 1 is shown in Fig. 2.) Thus, for $0 \leq s \leq t$ we can set

$$\hat{\xi}_s^B = \{x \in Z : \exists \text{ path down from } (B, t) \text{ to } (x, t-s) \text{ in } \hat{\mathscr{P}}(\lambda)\},$$

$B \in S$, and $(\hat{\xi}_s^B)_{0 \leq s \leq t}$ is a version of the basic contact process with parameter λ, starting in B and running up to time t. Eq. (5) now follows from the construction, since the events on both sides coincide with the event that there is a path (up or down) connecting $(A, 0)$ and (B, t) in the joint representation.

Using the properties of Theorem 1, we now indicate the qualitative nature of the critical phenomenon for contact processes. First, note that $\emptyset$ = 'all healthy' is a trap, so the measure $\delta_\emptyset$ concentrated at $\emptyset$ is invariant. The hitting time for $\emptyset$ will be denoted

$$\tau^A = \min\{t : \xi_t^A(\lambda) = \emptyset\} \quad (= \infty \text{ if no such } t \text{ exists}).$$

[1] The reader should be on the alert for hybrid space–time notation. For instance, $(A, 0)$ means $A \times \{0\}$, while $(x, \infty) \times 0$ means the spatial interval (x, ∞) cross $\{0\}$.

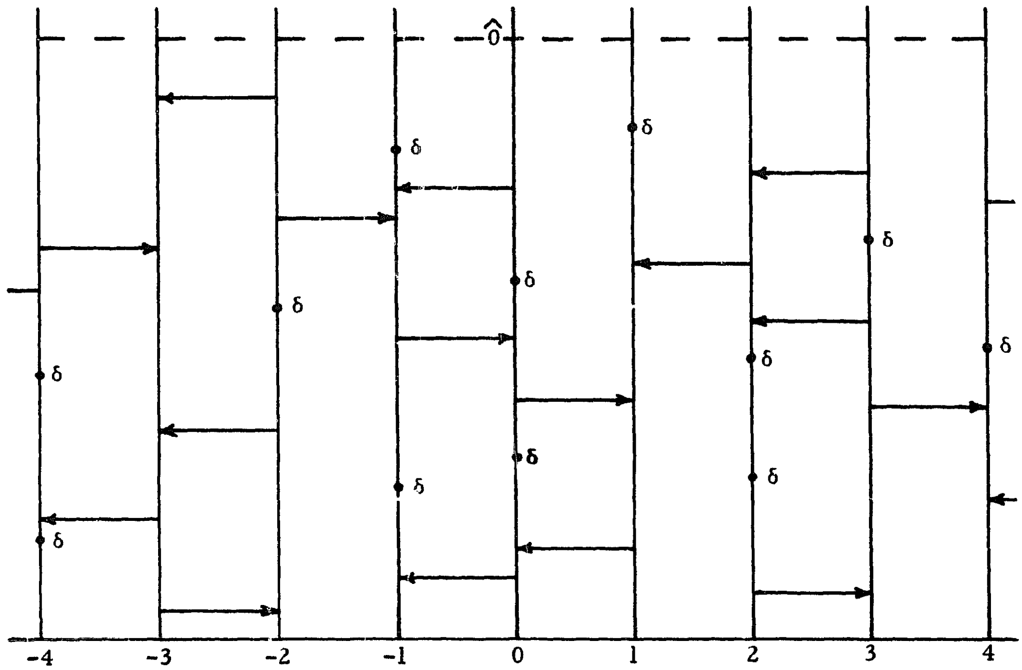

Fig. 2.

By taking $A=Z$ and $B\in S_0=\{\text{finite subsets of } Z\}$ in the self-duality equation (5), we find that

$$\mathbf{P}(\xi_t\cap B\neq\emptyset)=\mathbf{P}(\xi_t^B\neq\emptyset)=\mathbf{P}(\tau^B>t)\downarrow\mathbf{P}(\tau^B=\infty)$$

as $t\to\infty$. By inclusion–exclusion it follows that

$$\lim_{t\to\infty}\mathbf{P}(\xi_t\cap B=C)=\nu(\cdot\cap B=C),$$

$C\subset B\in S_0$, for some measure $\nu=\nu_\lambda$ on S. In terms of weak convergence,

$$\mathbf{P}(\xi_t\in\cdot)\Rightarrow\nu\quad\text{as } t\to\infty,$$

where ν is determined by

$$\nu(\cdot\cap B\neq\emptyset)=\mathbf{P}(\tau^B=\infty).$$

In particular, the density ρ of ν satisfies

$$\rho=\nu(0\in\cdot)=\mathbf{P}(\tau^0=\infty).$$

(The measures $\mathbf{P}(\xi_t\in\cdot)$ and the limit ν inherit translation invariance from $\mathscr{P}$; we will use translation invariance properties repeatedly in this paper without further comment.) If $\rho=0$, then for any $A\in S$, by set-monotonicity

$$\begin{aligned}\mathbf{P}(\xi_t^A\cap B\neq\emptyset)&\leq\mathbf{P}(\xi_t\cap B\neq\emptyset)\\&=\mathbf{P}(x\in\xi_t \text{ for some } x\in B)\\&\leq|B|\mathbf{P}(0\in\xi_t)\downarrow|B|\rho=0.\end{aligned}$$

Evidently $\nu = \delta_\emptyset$, and

$$\mathbf{P}(\xi_t^A \in \cdot) \Rightarrow \delta_\emptyset \quad \text{as } t \to \infty \ \ \forall A \in S.$$

In this case the Markov family $\{(\xi_t^A); A \in S\}$ is said to be *ergodic*, since processes with arbitrary initial state converge to the unique equilibrium $\delta_\emptyset$. Note also that

$$\mathbf{P}(\tau^B < \infty) = 1 \quad \forall B \in S_0$$

in the ergodic case. If $\rho(\lambda) > 0$, then clearly $\nu \neq \delta_\emptyset$ so there are two distinct equilibria. In this case $\{(\xi_t^A)\}$ is said to be *nonergodic*. By set-monotonicity we have

$$\mathbf{P}(\tau^B = \infty) \geqslant \mathbf{P}(\tau^0 = \infty) = \rho > 0 \quad \forall B \in S_0,\ B \neq \emptyset$$

in the nonergodic case. Finally, from λ-monotonicity it follows that

$$\mathbf{P}(\tau^0(\lambda_0) = \infty) \leqslant \mathbf{P}(\tau^0(\lambda) = \infty) \quad \text{if } \lambda_0 \leqslant \lambda,$$

so

$$\rho \text{ is an increasing function of } \lambda.$$

Thus there is a *critical value* λ_c,

$$\lambda_c = \sup\{\lambda : \rho(\lambda) = 0\},$$

such that $\{(\xi_t^A)\}$ is ergodic if $\lambda < \lambda_c$ and nonergodic if $\lambda > \lambda_c$. Note that the possibility $\lambda_c = \infty$ has not been ruled out as of yet. We summarize our findings in the form of a theorem.

Theorem 2. *There is a critical value λ_c such that*

(a) *for $\lambda < \lambda_c$ (subcritical case),*

$$\mathbf{P}(\xi_t^A \in \cdot) \Rightarrow \delta_\emptyset \quad \text{as } t \to \infty \ \ \forall A \in S,$$

$$\mathbf{P}(\tau^A < \infty) = 1 \quad \forall A \in S_0;$$

(b) *for $\lambda > \lambda_c$ (supercritical case),*

$$\mathbf{P}(\xi_t \in \cdot) \Rightarrow \nu_\lambda \neq \delta_\emptyset \quad \text{as } t \to \infty,$$

$$\mathbf{P}(\tau^A = \infty) > 0 \quad \forall A \in S_0,\ A \neq \emptyset.$$

(For a discussion of the critical case $\lambda = \lambda_c$, see Section 7 below.)

3. Numerical bounds for the critical value

In this section we discuss briefly the known lower and upper bounds for λ_c. First, let us show that for sufficiently small positive λ, the contact processes with parameter λ tend to die out, i.e. form an ergodic family. From the last section we know that ergodicity is equivalent to

$$\mathbf{P}(\tau^0 < \infty) = 1.$$

This property is quite easy to verify for $\lambda < 1$. Simply observe that if $\xi_t^0 = A \in S_0$, $|A| \geq 2$, then the diameter of ξ_t^0 increases by one at rate 2λ and decreases by *at least* one at rate 2. Thus, for $\lambda < 1$, the diameter is majorized by a random walk with mean $2(\lambda - 1) < 0$. It follows that $|\xi_t^0| = 1$ repeatedly as long as the process lives, and hence that the process dies eventually with probability one. We conclude that

$$\lambda_c \geq 1.$$

One can do better; in [12], for example, there is a simple proof that

$$\lambda_c \geq \frac{1+\sqrt{37}}{6} \approx 1.18.$$

The technique described there can be pushed further, but the computations rapidly become insurmountable.

The problem of upper bounds for λ_c is much more difficult. The best known bound is

$$\lambda_c \leq 2,$$

due to Holley and Liggett [21]. Here is a sketch of their remarkable approach. If μ is a measure on S, define

$$\theta_\mu(A) = \mu\{\cdot \cap A \neq \emptyset\}, \quad A \in S_0,$$

and note that $\theta_\mu \leq 1_{\{A \neq \emptyset\}}$. Now, for $B \in S_0$, consider the functions

$$h_B(t) = \mathbf{E}[\theta_\mu(\xi_t^B)]$$

($\mathbf{E}$ the expectation operator corresponding to P). If one can find a translation invariant measure μ with positive density such that

$$h'_B(t) \geq 0 \quad \text{for all } t \geq 0,\ B \in S_0, \tag{6}$$

then

$$\mathbf{P}(\xi_t^B \neq \emptyset) \geq h_B(t) \geq h_B(0) = \mu\{0 \in \cdot\} > 0$$

for all $t \geq 0$. Hence

$$\mathbf{P}(\tau^B = \infty) > 0, \quad B \in S_0,$$

i.e. nonergodicity holds. Using the Markov property it is easy to see that

$$h'_B(t) = \sum_{A \in S_0} \mathbf{P}(\xi_t^B = A) h'_A(0).$$

To establish (6) it therefore suffices to show

$$h'_B(0) \geq 0, \quad B \in S_0,$$

or equivalently

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathbf{E}[\mu\{\cdot \cap \xi_t^B = \emptyset\}]|_{t=0} \leq 0, \quad B \in S_0. \tag{7}$$

It turns out that there is a renewal measure which satisfies (7) if λ is large enough. A renewal measure μ_f on S is determined by a probability density $f = (f_k)_{k=1}^{\infty}$ such that $M = \sum kf_k < \infty$, by means of the basic cylinder prescriptions

$$\begin{aligned}\mu_f\{\cdot \cap [x, x+y_1+\cdots+y_n] = \{x, x+y_1, \ldots, x+y_1+\cdots+y_n\}\} \\ = M^{-1}\prod_{l=1}^{n} f_{y_l}.\end{aligned}$$

Evidently μ_f is translation invariant with density M^{-1}. The method of Holley and Liggett is to choose (f_k) so that (7) holds *with equality* in case $B = [x, y]$ for some $x \leq y$, and then to prove the inequality for arbitrary $B \in S_0$ with $\mu = \mu_f$ so chosen. The second step is difficult; see [21]. But to find the desired f, note that when B is a block, say $B = [0, m-1]$, the contact process gains one site at either end with rate λ, while an infected site $k \in [0, m-1]$ recovers at rate 1. Putting $F_m = \sum_{k=m+1}^{\infty} f_k$, a calculation therefore shows that equality in (7) for all blocks is equivalent to

$$2\lambda F_m = \sum_{k=0}^{m-1} F_k F_{m-k-1}, \quad m \geq 1 \qquad (F_0 = 1).$$

One can solve for F to get

$$F_m = \frac{(2m)!}{m!(m+1)!}(2\lambda)^{-m},$$

so F is summable for $\lambda \geq 2$. Over this parameter range

$$M = \sum_k F_k = \lambda - \sqrt{\lambda^2 - 2\lambda}.$$

The conclusion is that for $\lambda \geq 2$, the contact processes with parameter λ have a nontrivial equilibrium ν_λ, and that the density $\rho(\lambda)$ of ν_λ satisfies

$$\rho(\lambda) \geq M^{-1} = \frac{1}{2} + \sqrt{\frac{1}{4} - \frac{1}{2\lambda}}, \quad \lambda \geq 2. \tag{8}$$

More generally, when $\lambda \geq 2$ we obtain the bounds

$$\begin{aligned}\nu\{\cdot \cap B \neq \emptyset\} = \mathbf{P}(\tau^B = \infty) = \lim_{t\to\infty} \mathbf{P}(\xi_t^B \neq \emptyset) \\ \geq h_B(0) = \mu_f\{\cdot \cap B \neq \emptyset\}.\end{aligned} \tag{9}$$

We will see later in the paper that for blocks B, these inequalities turn out to be useful.

A second method of proving nonergodicity for large λ is based on what is known as the contour method. As we have seen, permanence of infection is equivalent to the condition

$$\mathbf{P}(\tau^{A} < \infty) < 1 \quad \text{for some } A \in S_0,$$

so we seek upper bounds for these extinction probabilities. To keep matters simple we begin by taking $A = \{0\}$; the contour method produces the desired bound by analyzing the boundary of the region infected by (0, 0) in the space–time diagram for (ξ_t^0). Here is an outline of the application to basic contact processes. For more details and a more general setting see Gray and Griffeath [8]. First, embed $Z \times [0, \infty)$ in $R \times [0, \infty)$, introduce

$$E = \{(y, t) \colon |y - x| \leq \tfrac{1}{2} \text{ for some } x \in \xi_t^0,\ t \geq 0\},$$

and let $\tilde{E}$ be E with all its holes filled in. On $\{\tau^0 < \infty\}$, $\tilde{E}$ is a bounded set; we denote by Γ the boundary of $\tilde{E}$ oriented clockwise. On $\{\tau^0 = \infty\}$ it is convenient to define $\Gamma = \emptyset$. A representative picture of the *contour* Γ is shown in Fig. 3. With probability one, Γ consists of $4n$ alternating vertical and horizontal edges for some $n \geq 1$, which we encode as follows. A curve with $4n$ edges will be described by a *direction vector*

$$D = (D_1, \ldots, D_{2n})$$

and a length vector

$$L = (L_1, \ldots, L_{2n}).$$

The D_i are one of the seven triples:

dld, drd, dru, ulu, uru, urd, dlu.

Here d, u, l and r stand for down, up, left and right respectively. To determine the direction vector for a contour Γ we start at $(-\frac{1}{2}, 0)$ and proceed clockwise around the

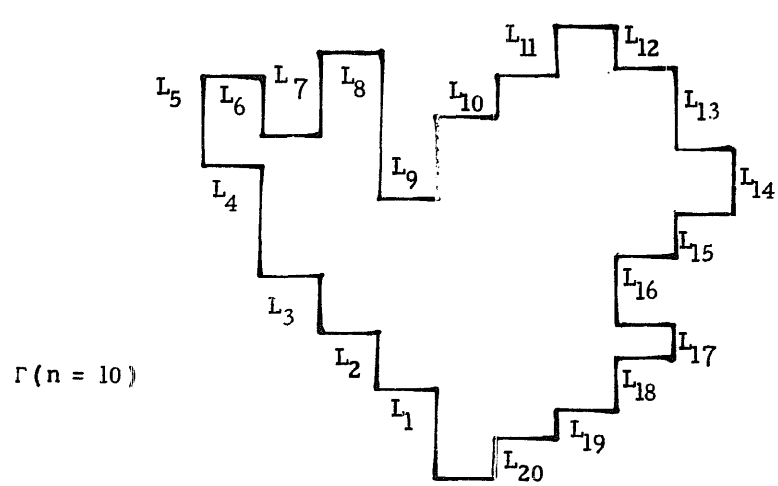

Fig. 3.

curve, reading off the sequence of directions of Γ in triples starting with successive vertical directions. Thus the direction vector for the Γ of Fig. 3 is:

ulu, ulu, ulu, ulu, urd, dru, urd, dru, uru, uru,

urd, drd, drd, dld, dld, drd, dld, dld, dld, dlu.

Note that *uld* cannot occur in Γ, and that *dlu* occurs only as the value of D_{2n}. The value L_i is the length of the vertical segment corresponding to the *first* direction in the triple D_i; the L_i are shown for the Γ of Fig. 3. (Of course the horizontal edges of Γ are all of length one with probability one.) Let N_1 through N_7 be the respective numbers of direction vectors of the above types in Γ, $4N$ the total number of edges. It is not hard to see that (P – a.s.)

$$N_1+N_4=N-1, \qquad N_2+N_3+N_5+N_6=N, \qquad N_6=N_3+1,$$

so the 6-tuple $(N_1,\ldots,N_6)$ is determined by N_1, N_2 and N_3.

Now the key observation is that the shape of Γ is intimately connected with the behavior of the independent exponential alarm clocks at nearby space–time points. This gives rise to an upper estimate on the "density" of Γ. Namely, if l is a $2n$-vector of possible vertical lengths, then

$$\mathbf{P}(D(\Gamma), L(\Gamma)\in dl, (N_1(\Gamma),\ldots,N_6(\Gamma))=(n_1,\ldots,n_6))$$

$$\leq \mathrm{e}^{-(1+\lambda)\sum_{i\in I_d} l_i}\,\mathrm{e}^{-\lambda\sum_{i\in I_u} l_i}\lambda^{n_1+n_4}\prod_{i=1}^{2n-1} dl_i \tag{10}$$

where $I_d=\{i\colon \text{the first letter of } D_i \text{ is } d\}$, $I_u=\{i\colon \text{the first letter of } D_i \text{ is } u\}$. Briefly, the argument for (10) is as follows. Along each downward edge of Γ we know that no right directed arrow emanates from the site immediately to the left and no δ occurs there over a time interval of length l_i; the probability of this is $\mathrm{e}^{-(1+\lambda)l_i}$. Similarly, along each upward edge no left directed arrow emanates from the site immediately to the right over a time interval of length l_i; the probability of this is $\mathrm{e}^{-\lambda l_i}$. Also, a right directed arrow arrives at the site immediately to the left of each vertex of Γ, where the direction changes from down to left *except* for the $4n$th vertex, and a left directed arrow arrives at the site just to the left of each vertex where the change is from up to left. Thus there are n_1+n_4 events of these sorts, each with 'probability' $\lambda\, dl_i$. Finally, just to the right of each vertex where Γ turns to the right a δ occurs; there are $n_2+n_3+n_5+n_6$ such events, each with probability dl_i. Moreover, the independence properties of the exponential variables involved imply that all of the above events are independent. Note that we cannot make use of the lack of δ's immediately to the right of up edges in Γ, since in certain cases this coincides with the lack of δ's to the left of down edges. Avoiding this dependence, however, we arrive at the 'multiplicative' estimate (10). For more details, see [8]. Next, we change variables by replacing the final l_i, $i \in I_u$, with $h = \sum_{i \in I_u} l_i$, and integrate over the region $\{\sum_{i \in I_d, i \neq 2n} l_i < h\}$ to get

$$\begin{aligned}
&\mathbf{P}(D(\Gamma); N_1 = n_1, N_2 = n_2, N_3 = n_3) \\
&\quad \leq \lambda^{n-1} \int_0^\infty e^{-(1+2\lambda)h} \frac{h^{|I_u|-1}}{(|I_u|-1)!} \frac{h^{|I_d|-1}}{(|I_d|-1)!} \, dh \\
&\quad = \lambda^{n-1} \binom{2n-2}{|I_d|-1} (1+2\lambda)^{-(2n-1)} \leq C_0 \binom{2n-1}{n_1+n_2+n_3} A_0^n,
\end{aligned}$$

where

$$C_0 = \left[\frac{1+2\lambda}{\lambda}\right], \qquad A_0 = \frac{\lambda}{(1+2\lambda)^2}.$$

Hence, putting

$$\#(n_1, n_2, n_3) = |\{D : N_1(D) = n_1, N_2(D) = n_2, N_3(D) = n_3\}|,$$

we have

$$\mathbf{P}(\Gamma \neq \emptyset) \leq \sum_{n=1}^{\infty} \sum_{n_1, n_2, n_3} \#(n_1, n_2, n_3) \cdot C_0 \binom{2n-1}{n_1+n_2+n_3} A_0^n.$$

Observe next that

$$\#(n_1, n_2, n_3) \leq \frac{(n_1+n_2+n_3)!}{n_1! n_2! n_3!} \frac{(n_4+n_5+n_6)!}{n_4! n_5! n_6!},$$

since any direction vector D is completely determined by the ordering of its triples beginning with u and the ordering of its triples beginning with d. Thus

$$\mathbf{P}(\Gamma \neq \emptyset) \leq C_0 \sum_{n=1}^{\infty} A_0^n \left[\sum_{n_1, n_2, n_3} \frac{(2n-1)!}{n_1! n_2! n_3! n_4! n_5! n_6!} \right].$$

Now the bracketed sum is at most $\frac{1}{4} \cdot 32^n$, since it can be written as

$$\begin{aligned}
&\sum_{n_1, n_2, n_3} \binom{2n-1}{n_1+n_4} \binom{n_1+n_4}{n_1} \binom{n_2+n_3+n_5+n_6}{n_3+n_6} \binom{n_2+n_5}{n_2} \binom{n_3+n_6}{n_3} \\
&\quad = \binom{2n-1}{n-1} \left[\sum_{n_1} \binom{n-1}{n_1} \right] \left[\sum_{n_3} \left\{ \sum_{n_2} \binom{n-2n_3-1}{n_2} \right\} \binom{n}{2n_3+1} \binom{2n_3+1}{n_3} \right] \\
&\quad \leq 2^{2n-1} \cdot 2^{n-1} \left[\sum_{n_3} 2^{n-2n_3-1} \binom{n}{2n_3+1} 2^{2n_3+1} \right] \\
&\quad \leq 2^{2n-1} \cdot 2^{n-1} \cdot 2^{2n}.
\end{aligned}$$

Setting

$$A_1 = 32 A_0 = \frac{32\lambda}{(1+2\lambda)^2},$$

we conclude that if $A_1 < \frac{1}{2}$, then

$$\mathbf{P}(\tau^0 < \infty) = \mathbf{P}(\Gamma \neq \emptyset) \leq \frac{A_1}{1 - A_1} < 1,$$

so nonergodicity holds. Better bounds are obtained by doing analogous computations starting from the blocks $A = [0, m-1]$: we get

$$\varphi(m) = \mathbf{P}(\tau^{[0,m-1]} < \infty) \leq C(\lambda) A_1^m, \tag{11}$$

for some constant $C(\lambda)$. By taking m large we see that nonergodicity holds whenever $A_1 < 1$, so that

$$\lambda_c \leq \tfrac{7}{2} + 2\sqrt{3} < 7.$$

The bound (11) is the contour method counterpart of (9).

4. The edge process and its applications

The simple proof of ergodicity for $\lambda < 1$ given at the beginning of the last section was based on the observation that the rightmost infected site r_t^0 of ξ_t^0 tends to drift to the left, and by symmetry the leftmost site l_t^0 drifts right. Consideration of the extreme sites

$$l_t^A = \min\{x : x \in \xi_t^A\}, \qquad r_t^A = \max\{x : x \in \xi_t^A\},$$

turns out to be a very effective approach to the detailed ergodic theory of contact processes. An analysis along these lines was initiated in papers by Harris [19], Griffeath [10] and Liggett [25], and most fully realized in a recent paper by Durrett [5]. This section is essentially an overview of Durrett's theorems for the edge processes, and the resulting ergodic theorems. The reader is referred to [5] for several of the proofs, which are nice but rather involved.

Recall that $\xi_t = \xi_t^Z$, $\xi_t^- = \xi_t^{(-\infty,0]}$, $\xi_t^+ = \xi_t^{[0,\infty)}$. Denote by r_t^- and l_t^+ the rightmost site of ξ_t^- and the leftmost site of ξ_t^- respectively. It is more convenient to study r_t^- and l_t^+ than r_t^0 and l_t^0, since the former processes are defined for all $t \geq 0$ whereas the latter are only defined on $\{\tau^0 > t\}$. The first result of this section, which we call the complete coupling property, shows that ξ_t^0 is intimately connected with ξ_t, ξ_t^- and ξ_t^+.

Theorem 3. *On* $\{\tau^0 > t\}$,

$$\xi_t^0 = \xi_t^A \cap [l_t^0, r_t^0], \qquad 0 \in A \subset Z, \tag{12}$$

and

$$l_t^+ = l_t^0, \qquad r_t^- = r_t^0. \tag{13}$$

Thus,

$$\xi_t^0 = \xi_t \cap [l_t^0, r_t^0] = \xi_t^+ \cap \xi_t^- \quad \text{on } \{\tau^0 > t\}. \tag{14}$$

Proof. If $\tau^0 > t$, then there is a path up from $(0, 0)$ to (l_t^0, t) and a path from $(0, 0)$ to (r_t^0, t). If $x \in \xi_t^A \cap [l_t^0, r_t^0]$, then there is a path up from $(A, 0)$ to (x, t). This last path must intersect one of the first two, and so there is a 'composite path' from $(0, 0)$ to (x, t). Therefore $x \in \xi_t^0$, and so $\xi_t^A \cap [l_t^0, r_t^0] \subset \xi_t^0$. The reverse inclusion holds by set-monotonicity. This proves (12). The argument for (13) is similar. Namely, if $r_t^0 = x < r_t^-$, then there is a path from $(0, 0)$ to (x, t) and a path from $(-\infty, 0) \times 0$ to $(x, \infty) \times t$. But this yields a composite path from $(0, 0)$ to $(x, \infty) \times t$, contradicting the definition of r_t^0. Thus $r_t^- \leqslant r_t^0$. The reverse inequality again follows from set-monotonicity, so $r_t^- = r_t^0$. The proof for the left edge is analogous. Now the first equation of (14) is the special case of (12), where $A = \mathbb{Z}$, and the second follows from (12) and (13):

$$\begin{aligned}\xi_t^0 &= (\xi_t^- \cap [l_t^0, r_t^0]) \cap (\xi_t^+ \cap [l_t^0, r_t^0]) \\ &= \xi_t^- \cap \xi_t^+ \cap [l_t^0, r_t^0] = \xi_t^- \cap \xi_t^+ \cap [l_t^+, r_t^-] = \xi_t^- \cap \xi_t^+.\end{aligned}$$

Using Theorem 3, one can analyse ergodicity of contact processes in terms of the behavior of r_t^-. Roughly speaking, if r_t^- drifts left ergodicity holds, whereas nonergodicity holds if r_t^- drifts right. Thus we introduce the expected displacements

$$\alpha_t = \mathbf{E}[r_t^-] \quad (= -\mathbf{E}[l_t^+]).$$

A key property of α_t is *subadditivity*:

$$\alpha_{t+u} < \alpha_t + \alpha_u, \qquad t, u > 0. \tag{15}$$

In fact, the process $(r_t^-)_{t \geqslant 0}$ is itself *subadditive* in the following sense: for any $t, u \geqslant 0$ there is a random variable s_{tu}^-, independent of r_t^- and with the same distribution as r_u^-, such that

$$r_{t+u}^- \leqslant r_t^- + s_{tu}^- \quad \text{a.s. for each } t, u \geqslant 0. \tag{16}$$

To see this, consider

$$s_{tu}^- = \max\{x : \exists \text{ path from } (-\infty, r_t^-] \times \{t\} \text{ to } (r_t^- + x, t+u)\}.$$

By translation invariance, s_{tu}^- is r_u^--distributed. Moreover, if γ is a path up to $(r_{t+u}^-, t+u)$ in ξ^-, then $\gamma_t \leqslant r_t^-$, so there is a path from $(-\infty, r_t^-] \times \{t\}$ to $(r_{t+u}^-, t+u)$, i.e. (16) holds. If $t, u > 0$, then since there is a positive probability of a path from $(-\infty, r_t^-) \times \{t\}$ to $(r_{t+u}, \infty) \times \{t+u\}$, by taking expectations in (16) we get (15). Now by a well-known theorem on subadditive sequences,

$$\alpha \equiv \inf_{M>0} \frac{\alpha_M}{M} = \lim_{t\to\infty} \frac{\alpha_t}{t}.$$

(Note that $\alpha = -\infty$ is a possibility.) We call α the *asymptotic velocity of the right edge*. By using the stochastic subadditivity (16) and related ideas, Durrett [5] has proved two useful properties concerning $\alpha = \alpha(\lambda)$:

$$\frac{r_t^-}{t} \to \alpha \quad \text{a.s. (and in } L^1) \text{ as } t \to \infty, \tag{17}$$

and

$$\alpha_t(\lambda+\delta) \geq \alpha_t(\lambda) + \delta t \quad \lambda, \delta, t \geq 0,$$

so that

$$\alpha(\lambda+\delta) \geq \alpha(\lambda) + \delta, \quad \lambda, \delta \geq 0. \tag{18}$$

(Here $-\infty + \delta \equiv -\infty$.) The result (17) asserts that the right edge settles down to its asymptotic speed with probability one. It is not hard to see that

$$\limsup_{t\to\infty} \frac{r_t^-}{t} \leq \alpha \quad \text{a.s. (and in } L^1\text{);}$$

Durrett proves the much more difficult inequality

$$\liminf_{t\to\infty} \frac{r_t^-}{t} \geq \alpha$$

by constructing a stationary ergodic process with mean α which 'lies to the left' of r_t^-. Inequality (18) states that the speed is a *strictly* increasing function of the infection parameter λ as soon as $\alpha(\lambda) > -\infty$. This is derived by a clever coupling argument. We refer the reader to [5] for the proofs of both theorems.

With the aid of Durrett's results, we now show that the critical value λ_c for the contact processes is precisely the value of λ at which the asymptotic velocity of the right edge changes from negative to positive.

Theorem 4. $\lambda_c = \sup\{\lambda : \alpha(\lambda) < 0\} = \sup\{\lambda : \alpha(\lambda) \leq 0\}$.

Proof. Put $\lambda_M = \sup\{\lambda : \alpha_M(\lambda) < 0\}$, $\lambda_* = \sup_M \lambda_M$. It follows easily from the definition of $\alpha(\lambda)$ that

$$\lambda_* = \sup\{\lambda : \alpha(\lambda) < 0\}.$$

Thus we will prove that $\lambda_* = \sup\{\lambda : \alpha(\lambda) \leq 0\} = \lambda_c$. If $\lambda > \lambda_*$, choose $\lambda' \in (\lambda_*, \lambda)$; then $\alpha_M(\lambda') > 0$ for all M, and so $\alpha(\lambda') \geq 0$. By (18), $\alpha(\lambda) > 0$, and hence $\lambda_* = \sup\{\lambda : \alpha(\lambda) \leq 0\}$. Moreover, (17) yields

$$r_t^-(\lambda) \to +\infty \quad \text{a.s.}$$

and by symmetry,

$$l_t^+(\lambda) \to -\infty \quad \text{a.s.}$$

Hence we can choose N large enough that

$$\mathbf{P}(r_t^- \text{ ever} < -N) = \mathbf{P}(l_t^+ \text{ ever} > N) = \varepsilon < \tfrac{1}{2}.$$

Now (14) generalizes easily to

$$\begin{aligned} \xi_t^{[-N,N]} &= \xi_t \cap [l_t^{[-N,\infty]}, r_t^{(-\infty,N]}] \\ &= \xi_t^{(-\infty,N]} \cap \xi_t^{[-N,\infty)} \quad \text{on } \{\tau^{[-N,N]} > t\}, \end{aligned} \tag{19}$$

and from (19) it is not hard to deduce that

$$\tau^{[-N,N]} = \min\{t: r_t^{(-\infty,N]} < l_t^{[-N,\infty]}\} \tag{20}$$

($=\infty$ if no such t exists). Using (20) and translation invariance we get

$$\begin{aligned}\mathbf{P}(\tau_0^{[-N,N]} < \infty) &\leq \mathbf{P}(r_t^{(-\infty,N]} \text{ ever} < 0) + \mathbf{P}(l_t^{[-N,\infty)} \text{ ever} > 0)\\ &< 2\varepsilon < 1.\end{aligned}$$

By set additivity and translation invariance again,

$$\rho = \mathbf{P}(\tau^0 = \infty) \geq \frac{1}{2N+1}\mathbf{P}(\tau^{[-N,N]} = \infty) > 0.$$

We have therefore proved that if $\lambda > \lambda_*$, then $\lambda > \lambda_c$. In other words, $\lambda_c \leq \lambda_*$.

To get the opposite inequality we argue as follows. If $\lambda < \lambda_*$, then $\alpha_M(\lambda) < 0$ for some integer M. Fix such an M, and let $(R_n^M; n = 0, 1, \ldots)$ be the random walk with negative mean α_M having displacement distribution $\mathbf{P}(r_M^- \in \cdot)$, (L_n^M) the random walk with positive mean $-\alpha_M$ having displacement distribution $\mathbf{P}(l_M^+ \in \cdot)$. Then copies of (R_n^M) and (L_n^M) can be defined on our percolation substructure $\mathscr{P}$ in such a way that $r_{nM}^- \leq R_n^M$ and $l_{nM}^+ \geq L_n^M$ for all n P-a.s. This is accomplished by 'filling in the holes' of ξ_t^- and ξ_t^+ when t is a multiple of M, just as in the argument for (16). By the law of large numbers,

$$R_n^M \to \infty \quad \text{and} \quad L_n^M \to -\infty \quad \text{a.s.}$$

This forces

$$\mathbf{P}(r_{nM}^- - l_{nM}^- < 0 \text{ for some } n) = 1. \tag{21}$$

Together, (21) and (20) with $N = 1$ imply that $\rho = \mathbf{P}(\tau^0 = \infty) = 0$. Thus, if $\lambda < \lambda_*$, then $\lambda \leq \lambda_c$, i.e. $\lambda_* \leq \lambda_c$. The proof is finished.

Remark. A straightforward argument shows that $\alpha_M(\lambda)$ is continuous for each M. By (18) α_M is strictly increasing, so λ_M is the unique root of $\alpha_M(\lambda) = 0$. Thus $\alpha_M(\lambda_*) \geq 0$ for each M, and hence $\alpha(\lambda_*) \geq 0$. Since $\lambda_* = \lambda_c$, we conclude that

$$\alpha(\lambda_c) \geq 0. \tag{22}$$

(In fact, $\alpha(\lambda_c) = 0$. See Section 7.)

We are now prepared to present two fundamental limit theorems for nonergodic contact processes. Since the proofs have already appeared elsewhere, we will only sketch them. The first result is known as the complete convergence theorem.[2]

[2] One way to define (ξ_t^μ) for nondeterministic initial μ is to enlarge the underlying probability space to support an independent μ-distributed random subset A of S, and to set $\xi_t^\mu = \xi_t^A$ on $\{A = A\}$.

Theorem 5. *If $\lambda > \lambda_c$, μ is any probability measure on S, and if (ξ_t^μ) is the contact process with parameter λ and initial distribution μ, then*

$$\mathbf{P}(\xi_t^\mu \in \cdot) \Rightarrow \mathbf{P}(\tau^\mu < \infty)\delta_\emptyset + \mathbf{P}(\tau^\mu = \infty)\nu \quad \text{as } t \to \infty. \tag{23}$$

(τ^μ is the hitting time for $\emptyset$ of (ξ_t^μ).) In particular, any invariant measure for $\{(\xi_t^A)\}$ is a mixture of $\delta_\emptyset$ and ν.

Sketch of proof. For simplicity, consider $\mu = \delta_{\{0\}}$. To get (23) in this case, it suffices to show that

$$\lim_{t \to \infty} \mathbf{P}(\xi_t^0 \cap A \neq \emptyset \mid \tau^0 = \infty) = \nu(\cdot \cap A \neq \emptyset), \quad A \in S_0;$$

weak convergence then follows by inclusion–exclusion. Since $\lambda > \lambda_c$, $\alpha(\lambda) > 0$ by Theorem 4. Hence,

$$l_t^+ \to -\infty \quad \text{and} \quad r_t^- \to \infty \quad \text{a.s.}$$

by (17). Thus, for all sufficiently large t,

$$A \subset [l_t^+, r_t^-],$$

so that by (14), a.s. on $\{\tau^0 = \infty\}$ we have

$$\xi_t^0 \cap A = \xi_t \cap [l_t^+, r_t^-] \cap A = \xi_t \cap A \tag{24}$$

for all sufficiently large t. It is therefore enough to show that the distribution of $\xi_t \mid \{\tau^0 = \infty\}$ converges to ν. Now we know that the unconditioned distribution of ξ_t converges to ν; using only this fact and set-monotonicity, it is not hard to finish the proof for $\mu = \delta_{\{0\}}$. The extensions, first to $\mu = \delta_A$ ($A \in S_0$), then to $\mu = \delta_A$ ($A \in S$), and finally to general μ, are all straightforward. Details may be found in [5] or [12].

We close this section with a second limit theorem for the nonergodic case, the complete pointwise ergodic theorem. This result states, for instance, that the proportion of time in $[0, t]$ when the origin is infected, given that infection survives forever, converges to ρ as $t \to \infty$ with probability one.

Theorem 6. *If $\lambda > \lambda_c$, if μ is any probability measure on S, and if f is any continuous function on S, then*

$$\frac{1}{t}\int_0^t f(\xi_s^\mu)\,\mathrm{d}s \to \begin{cases} f(\emptyset) & \text{as } t \to \infty \text{ a.s. on } \{\tau^\mu < \infty\}, \\ \displaystyle\int_S f\,\mathrm{d}\nu & \text{as } t \to \infty \text{ a.s. on } \{\tau^\mu = \infty\}. \end{cases} \tag{25}$$

Sketch of proof. The first line of (25) is obvious. To prove the second, one combines Birkhoff's ergodic theorem with a coupling argument. Namely, it follows from the complete convergence (23) and the Feller property of contact processes that the stationary process (ξ_t^ν) is Birkhoff ergodic (in fact, mixing), so that

$$\frac{1}{t}\int_0^t f(\xi_s^\nu)\,\mathrm{d}s \to \int_S f\,\mathrm{d}\nu \quad \text{a.s. } f\in L^1(\nu). \tag{26}$$

Also, assuming $\lambda > \lambda_c$, property (24) generalizes to $A \in S$, $\Lambda \in S_0$:

$$\xi_t^A \cap \Lambda = \xi_t \cap \Lambda \quad \text{for all large } t \text{ a.s. on } \{\tau^A = \infty\}. \tag{27}$$

Clearly $\tau^A = \infty$ a.s. for each $A \in S_\infty = S - S_0$, and it is easy to check that $\nu(S_\infty) = 1$. Hence

$$\xi_t^\nu \cap \Lambda = \xi_t \cap \Lambda \quad \text{for all large } t \text{ a.s.} \tag{28}$$

Combining (27) and (28) we get

$$\xi_t^A \cap \Lambda = \xi_t^\nu \cap \Lambda \quad \text{for all large } t \text{ a.s. on } \{\tau^A = \infty\}. \tag{29}$$

For $\mu = \delta_A$ and f depending only on sites in the finite set Λ, the second line of (25) follows easily from (26) and (29). The extensions to general μ and continuous f are routine. See [11] or [12] for more details.

5. Convergence rates in the subcritical case

In this section we prove a theorem giving rates of ergodicity for subcritical contact processes. Our result asserts that exponential ergodicity takes place for all $\lambda < \lambda_c$. As a sample path consequence, we find that infection dies out not only in the weak sense, but in fact it eventually disappears forever from any finite set Λ.

Theorem 7. *If $\lambda < \lambda_c$, then there are positive constants K_0 and K_1, depending only on λ, such that for each $\Lambda \in S_0$,*

$$\sup_{A\in S} \|\mathbf{P}(\xi_t^A \cap \Lambda \in \cdot\,) - \delta_\emptyset|_\Lambda\| \le K_0|\Lambda|\,\mathrm{e}^{-K_1 t}, \quad t \ge 0. \tag{30}$$

Thus, for any $A \in S$, $\Lambda \in S_0$,

$$\mathbf{P}(\xi_t^A \cap \Lambda = \emptyset \textit{ for all sufficiently large } t) = 1. \tag{31}$$

Proof. If $\lambda < \lambda_c$, then from the proof of Theorem 4 we know that $\alpha_M(\lambda) < 0$ for some integer M, and hence that there are random walks R_n^M and L_n^M with means $\alpha_M(\lambda)$ and $-\alpha_M(\lambda)$ respectively, such that

$$\mathbf{P}(\tau^0 > n) \le \mathbf{P}(R_n^M - L_n^M \ge 0) \le \mathbf{P}(R_n^M \ge 0 \textit{ or } L_n^M \le 0).$$

By a standard large deviations result for random walks, the last term is at most $2K\,\mathrm{e}^{-K_1 n}$, for some constants K, K_1 depending only on λ. By monotonicity,

$$\mathbf{P}(\tau^0 > t) \le 2K\,\mathrm{e}^{K_2}\,\mathrm{e}^{-K_2 t}, \quad t \ge 0. \tag{32}$$

Next, for any $A \in S$, $\Lambda \in S_0$, we can write

$$\{\xi_t^A \cap \Lambda \in \cdot\} = \bigcup_n \{\xi_t^A \cap \Lambda = \Lambda_n\}$$

for some finite collection of disjoint subsets Λ_n of Λ. For $\Lambda_n \neq \emptyset$, choose $x_n \in \Lambda_n$. Then

$$\mathbf{P}(\xi_t^A \cap \Lambda \in \cdot) \leq \mathbf{P}(\xi_t^A \cap \Lambda = \emptyset) + \mathbf{P}\Big(\bigcup_n \{x_n \in \xi_t^A\}\Big).$$

Therefore, by set monotonicity, translation invariance, duality and (32),

$$\begin{aligned}
&\|\mathbf{P}(\xi_t^A \cap \Lambda \in \cdot) - \delta_\emptyset|_\Lambda\| \\
&\quad \leq 2\mathbf{P}\Big(\bigcup_{x\in\Lambda} \{x \in \xi_t^A\}\Big) \leq 2\mathbf{P}\Big(\bigcup_{x\in\Lambda} \{x \in \xi_t\}\Big) = 2|\Lambda|\mathbf{P}(0 \in \xi_t) \\
&\quad = 2|\Lambda|\mathbf{P}(\tau^0 > t) \leq 4K\, \mathrm{e}^{K_2}\, \mathrm{e}^{-K_2 t}.
\end{aligned}$$

Thus (30) holds with $K_1 = 4K\,\mathrm{e}^{K_2}$. The claim (31) follows easily from (30) by a Borel–Cantelli argument.

As a consequence of Theorem 7, we get the following result for the subcritical asymptotic velocities.

Theorem 8. *If $\lambda < \lambda_c$, then $\alpha(\lambda) = -\infty$.*

Proof. We will prove a stronger assertion: if $\lambda < \lambda_c$, then there is a $C = C(\lambda) > 0$ such that

$$\mathbf{P}(r_t^- > -\mathrm{e}^{Ct}) \to 0 \quad \text{exponentially as } t \to \infty. \tag{33}$$

Thus the right edge of (ξ_t^-) runs off to $-\infty$ exponentially fast in the subcritical case. To see this, simply note that

$$\begin{aligned}
\mathbf{P}(r_t^- > -\mathrm{e}^{Ct}) &= \mathbf{P}(r_t^- > \mathrm{e}^{Ct}) + \mathbf{P}(r_t^- \in (-\mathrm{e}^{Ct}, \mathrm{e}^{Ct}]) \\
&\leq \mathbf{P}(r_t^- > \mathrm{e}^{Ct}) + \mathbf{P}(\xi_t^- \cup (-\mathrm{e}^{Ct}, \mathrm{e}^{Ct}] \neq \emptyset).
\end{aligned}$$

The first probability on the left side tends to 0 exponentially fast for any $C > 0$ since r_t^- moves right with speed at most λ. According to Theorem 7, the second probability is majorized by

$$(2K_0\,\mathrm{e}^{Ct})\,\mathrm{e}^{-K_1 t},$$

with K_0 and K_1 as in (30). Thus (33) holds for any $C < K_1$.

As noted previously (cf. (22)), it is easy to see that $\alpha(\lambda_c) \geq 0$. In fact, as a consequence of a result from the next section, $\alpha(\lambda_c) = 0$. Thus α is discontinuous at $\lambda = \lambda_c$. For $\lambda \geq \lambda_c$ we know from (18) that α is strictly increasing. Being also the infimum of the continuous functions α_M/M, α is right continuous. Presumably α is concave, and hence continuous, on $[\lambda_c, \infty)$, but these properties are not known.

In the nonergodic case one wants upper and lower bounds for α. Since r_t^- moves right one unit at rate λ and moves left *at least* one unit at rate 1, the easy upper bound is

$$\alpha(\lambda) \leq \lambda - 1, \quad \lambda \geq 1.$$

From (8) and (18) one gets the lower bound

$$\alpha(\lambda) \geq \lambda - 2, \quad \lambda \geq 2.$$

6. Convergence and mixing rates in the nonergodic case

In contrast to the subcritical case, exponential convergence rates are not known for all values $\lambda > \lambda_c$. This undoubtedly reflects the inadequacy of available techniques rather than the presence of slower rates just above the critical value. In any case, we will have to be content in this section with results which hold for sufficiently large λ.

Two key quantities for our purposes are

$$\psi_\lambda(t) = \mathbf{P}(t < \tau^0(\lambda) < \infty), \quad t \geq 0,$$

and (cf. (11))

$$\varphi_\lambda(m) = \mathbf{P}(\tau^{[0,m-1]}(\lambda) < \infty), \quad m \geq 1.$$

If we define

$$\lambda_\psi = \inf\{\lambda > \lambda_c : \psi_\lambda(t) \to 0 \text{ exponentially in } t\},$$

and

$$\lambda_\varphi = \inf\{\lambda : \varphi_\lambda(m) \to 0 \text{ exponentially in } m\},$$

then presumably $\lambda_\psi = \lambda_\varphi = \lambda_c$. The best rigorous bounds, however, are

$$\lambda_\psi < 7, \tag{34}$$

$$\lambda_\varphi \leq 2. \tag{35}$$

To get (34) we use the contour calculus. Adopting the notation of Section 3, we note that if $t < \tau < \infty$, then a contour $\Gamma \neq \emptyset$ such that $h = \sum_{i \in I_u} l_i > t$ occurs. Hence we arrive at the estimate

$$\begin{aligned}
&\mathbf{P}(t < \tau^0 < \infty) \\
&\leq \sum_{n=1}^{\infty} \frac{32^n}{4} \frac{\lambda^{n-1}}{(2n-1)!} \int_t^\infty e^{-(1+2\lambda)h} h^{2n-2}\, dh \\
&= \frac{e^{-(1+2\lambda)t}}{4\lambda} \sum_{n=1}^{\infty} \frac{(32\lambda)^n}{(2n-1)!} \int_0^\infty e^{-(1+2\lambda)h} (t+h)^{2n-2}\, dh
\end{aligned}$$

$$= \frac{e^{-(1+2\lambda)t}}{4\lambda} \sum_{n=1}^{\infty} \frac{(32\lambda)^n}{(2n-1)!} \sum_{k=0}^{2n-2} \binom{2n-2}{k} t^k \frac{(2n-2-k)!}{(1+2\lambda)^{2n-1-k}}$$

$$\leq \frac{e^{-(1+2\lambda)t}}{4\lambda} \sum_{k=0}^{\infty} \frac{[(1+2\lambda)t]^k}{k!} \sum_{k/2+1}^{\infty} A_1^n$$

$$\leq \frac{A_1}{4\lambda(1-A_1)} \exp\{-(1+2\lambda)(1-\sqrt{A_1})t\}$$

$$\to 0 \quad \text{exponentially in } t,$$

provided that $A_1 = 32\lambda/(1+2\lambda)^2 < 1$. Thus (34) holds. As noted in (11), the contour method also shows that $\lambda_\varphi < 7$, but in this case the Holley–Liggett approach yields the better bound (35). Suppose $\lambda > 2$, and consider the renewal measure μ_f constructed in Section 3. According to (9),

$$\varphi_\lambda(m) = \lim_{t\to\infty} \mathbf{P}(\xi_t^{[0,m-1]} = \emptyset) \leq \mu_f\{\cdot \cap [0, m-1] = \emptyset\}.$$

By the construction and translation invariance of μ_f, the rightmost term equals

$$\sum_{k=m}^{\infty} \mu_f\{1 \text{ at } 0, \text{ all 0's on } [1, k]\}$$

$$= \frac{1}{M} \sum_{k=m}^{\infty} F_k \leq \sum_{k=m}^{\infty} \left(\frac{2}{\lambda}\right)^k = \frac{\lambda}{\lambda-2} \left(\frac{2}{\lambda}\right)^m$$

$$\to 0 \quad \text{exponentially in } m.$$

Hence (35) holds.

We now discuss applications of the exponential bounds on $\psi(t)$ and $\varphi(t)$. First, we easily derive exponential convergence to ν for the nonergodic contact processes with $\lambda \geq 7$ and initial state Z.

Theorem 9. *If $\lambda > \lambda_\varphi$, then for each $A \in S_0$ there are constants $K_0 = K_0(\lambda)$ and $K_1 = K_1(\lambda)$, such that*

$$\|\mathbf{P}(\xi_t \cap A \in \cdot) - \nu|_A\| \leq K_0|A| e^{-K_1 t}.$$

Proof. By inclusion–exclusion, it suffices to show that

$$|\mathbf{P}(x \in \xi_t) - \rho(x)| \to 0 \quad \text{exponentially in } t.$$

Using self-duality, note that the above difference is in fact *equal* to $\psi(t)$.

An exponential convergence theorem allowing a more general class of initial states requires some work. A good deal of this work has been done by Harris, as we now explain.

Remark. In [19], Harris has used graphical methods to obtain exponential estimates for nonergodic contact processes. Because our continuous time contour calculus was not available when [19] was written, he resorted to certain discrete approximations. Consequently, his results hold only for extremely large λ. However his ideas can now be combined with continuous contour techniques to prove results such as:

If $\lambda > \lambda_\varphi$, then there is a $C = C(\lambda) < \infty$ such that

$$\mathbf{P}(|\{x \in [0, m-1]: \tau^x = \infty\}| < Cm) \to 0$$

exponentially in m,

and

If $\lambda \geq 14$, then there is a $C' = C'(\lambda) < \infty$ such that

$$\mathbf{P}(|\xi_t^0| < C't, \tau^0 = \infty) \to 0 \tag{36}$$

exponentially in t.

By Borel–Cantelli and (36), one gets Harris' Growth Theorem for $\lambda \geq 14$:

$$\mathbf{P}\left(\liminf_{t\to\infty} \frac{|\xi_t^0|}{t} > 0 \,|\, \tau^0 = \infty\right) = 1.$$

Next, let μ_θ be Bernoulli product measure with density $\theta > 0$. Using self-duality, it is not hard to see that

$$|\mathbf{P}(x \in \xi_t) - \mathbf{P}(x \in \xi_t^{\mu_\theta})| = \mathbf{E}[(1-\theta)^{|\xi_t^x|}, \tau^x > t].$$

For $\lambda \geq 14$, one can apply (36) to check that the right side tends to 0 exponentially in t, and hence that Theorem 9 holds for $(\xi_t^{\mu_\theta})$, $\theta > 0$.

We now turn our attention to the *spatial* dependence structure of the invariant measure ν. In [5], Durrett gives a simple proof that each ν_λ, $\lambda > \lambda_c$, is ergodic. The next result asserts that the ν_λ are asymptotically uncorrelated, and that the correlations decay exponentially for $\lambda > \lambda_\varphi$.

Theorem 10. *Introduce the correlation function* $\varphi_\lambda(A) = \nu_\lambda(\cdot \cap A = \emptyset)$, $\varphi_\lambda(m) = \varphi_\lambda([0, m-1])$. *Then for each* $\lambda > \lambda_c$,

$$0 \leq \varphi(A \cup B) - \varphi(A)\varphi(B) \leq 2|B|\varphi(d(A, B)), \tag{37}$$

$A \in S$, $B \in S_0$, *where* $d(A, B) = \min\{|x-y|: x \in A, y \in B\}$ *is the distance between* A *and* B. *Thus the correlations decay exponentially if* $\lambda > \lambda_\varphi$.

Proof. It suffices to prove (37) for arbitrary finite A and finite B, both of which we fix in the argument which follows. If we use *independent* substructures $\mathcal{P}_1$ and $\mathcal{P}_2$ to define the subfamily $\{(\xi_t^{A_0}); A_0 \subset A\}$ and $\{(\xi_t^{B_0}); B_0 \subset B\}$, and if we define

$$\bar{\tau}^{A_0B_0} = \min\{t: \xi_t^{A_0} \cup \xi_t^{B_0} = \emptyset\}. \tag{38}$$

then copies of the processes $(\xi_t^{A_0\cup B_0})$ can be defined in terms of $\mathscr{P}_1$ and $\mathscr{P}_2$ so that the following properties hold:

$$\varphi(A_0\cup B_0)=\mathbf{P}(\tau^{A_0\cup B_0}<\infty), \tag{39}$$

$$\varphi(A_0)\varphi(B_0)=\mathbf{P}(\bar{\tau}^{A_0B_0}<\infty), \tag{40}$$

and

$$\xi_t^{A_0\cup B_0}\begin{cases}=\xi_t^{A_0}\cup\xi_t^{B_0} & t\le\tau(A_0,B_0)=\min\{t:\ \xi_t^{A_0}\cap\xi_t^{B_0}\ne\emptyset\},\\ \subset\xi_t^{A_0}\cup\xi_t^{B_0} & \text{for all } t.\end{cases} \tag{41}$$

For more on this construction, see [12].

Combining (38)–(41), we have

$$\begin{aligned}\varphi(A\cup B)-\varphi(A)\varphi(B)&=\mathbf{P}(\tau^{A\cup B}<\infty,\ \bar{\tau}^{AB}=\infty)\\ &\le P\Big(\bigcup_{y\in B}\{\exists x\in A:\ \tau(x,y)\le\tau^{\{x,y\}}<\infty\}\Big).\end{aligned} \tag{42}$$

At this point it is convenient to return to the canonical representation of $\{(\xi_t^A);\ A\subset Z\}$ in terms of a single substructure $\mathscr{P}$. Introduce:

$$x_l^y=\max\{x<y:\ x\in A\},\qquad x_r^y=\min\{x>y:\ x\in A\}.$$

Setting

$$E_y=\{\forall x\in A:\ \tau(x,y)\le\tau^{\{x,y\}}<\infty\},$$

we claim that $E_y\subset\{\tau^{[x_l^y,y]}<\infty\}\cup\{\tau^{[y,x_r^y]}<\infty\}$. This is so because on E_y the contour of one of the processes $(\xi^{[x_l^y,y]})$ or $(\xi^{[y,x_r^y]})$ is 'enclosed' by the contour of some $(\xi_t^{\{x,y\}})$, $x\in A$. Thus, from (42) we have

$$\begin{aligned}\varphi(A\cup B)-\varphi(A)\varphi(B)&\le|B|\sup_{y\in B}\mathbf{P}(E_y)\\ &\le|B|\sup_{y\in B}[\mathbf{P}(\tau^{[x_l^y,y]}<\infty)+\mathbf{P}(\tau^{[y,x_r^y]}<\infty)].\end{aligned}$$

Since $[x_l^y,y]$ and $[y,x_r^y]$ are blocks of length at least $d(A,B)$, the desired result (37) follows by translation invariance. By self-duality $\lim_{m\to\infty}\varphi(m)=\nu(\{\emptyset\})$. From the complete convergence theorem (Theorem 5), or from more elementary monotonicity considerations, we know that ν and δ are extreme. Hence $\nu(\{\emptyset\})=0$ if $\lambda>\lambda_c$, and so the correlations decrease to 0. By definition of λ_φ, the rate is exponential if $\lambda>\lambda_\varphi$. The proof is finished.

Remark. General limit theorems for random fields (see e.g. [26]) yield a central limit theorem whenever the correlations decay exponentially. Thus, letting ξ_∞ be ν_λ-distributed, if we write

$$S_m=|\xi_\infty\cup[-m,m]|,\qquad\sigma_m^2=var(S_m),$$

then for $\lambda \geq 2$ it follows that

$$\lim_{m\to\infty} \nu_\lambda\left\{\frac{S_m - \mathbf{E}[S_m]}{\sigma_m} \leq a\right\} = \int_{-\infty}^{a} (2\pi)^{-1/2} e^{-u^2/2}\, du.$$

Whether ν_λ obeys the central limit theorem for all $\lambda > \lambda_c$ remains an open problem.

7. The critical contact process

The problem of ergodicity at $\lambda = \lambda_c$ may be viewed as a question about the equilibrium density function $\rho(\lambda)$. Since $\rho(\lambda) = 0$ for $\lambda < \lambda_c$ and $\rho(\lambda) > 0$ for $\lambda > \lambda_c$, the family of critical contact processes is ergodic if and only if ρ is continuous at λ_c. Before confronting the critical case, we note that it is relatively easy to prove continuity of ρ away from λ_c.

Theorem 11. *$\rho(\lambda)$ is continuous on* (λ_c, ∞).

Proof. Here is an argument which we learned from Larry Gray. Compactness considerations show that for any λ_0 there are sequences $\lambda' \uparrow \lambda_0$ and $\lambda'' \downarrow \lambda_0$, and invariant measures ν^- and ν^+ for $\{(\xi_t^A(\lambda_0))\}$, such that

$$\nu_{\lambda'} \Rightarrow \nu^-, \qquad \nu_{\lambda''} \Rightarrow \nu^+.$$

In particular letting ρ^- and ρ^+ denote the respective densities of ν^- and ν^+,

$$\rho(\lambda') \uparrow \rho^-, \qquad \rho(\lambda'') \downarrow \rho^+.$$

Since

$$\begin{aligned}
\nu^+\{\cdot \cap A \neq \emptyset\} &= \lim_{t\to\infty} \mathbf{P}(\xi_t^{\nu^+}(\lambda_0) \cap A \neq \emptyset) \\
&\leq \lim_{t\to\infty} \mathbf{P}(\xi_t(\lambda_0) \cap A \neq \emptyset) \\
&= \nu_{\lambda_0}\{\cdot \cap A \neq \emptyset\} \\
&\leq \lim_{\lambda''\downarrow\lambda_0} \nu_{\lambda''}\{\cdot \cap A \neq \emptyset\} = \nu^+\{\cdot \cap A \neq \emptyset\},
\end{aligned}$$

necessarily $\nu^+ = \nu_{\lambda_0}$. Moreover, for $\lambda_0 > \lambda_c$,

$$\rho^- \geq \lim_{\lambda'\uparrow\lambda_0} \rho(\lambda') > 0.$$

Thus if ρ were discontinuous at λ_0, then we would have

$$0 < \rho^- < \rho^+ = \rho(\lambda_0).$$

By Theorem 5, ν^- would have to be a mixture of $\delta_\emptyset$ and ν_{λ_0}, evidently nontrivial, forcing $\nu^-\{\emptyset\}>0$. But, as noted in the proof of Theorem 9, $\nu_\lambda\{\emptyset\}=0$ for all $\lambda>\lambda_c$, so by λ-monotonicity,

$$\nu^-\{\emptyset\}\leq \lim_{\lambda'\uparrow\lambda_0} \nu_\lambda\{\emptyset\}=0.$$

This contradiction implies the desired continuity at λ_0.

Let us now turn to the critical contact processes $\{\xi_t^A(\lambda_c)\}$. Current research has led to a couple of partial results. The analysis is quite involved, so we will not go into it here. Instead we simply mention the progress we have made so far. We have succeeded in proving:

(a) if $\alpha(\lambda)>0$, then $\lambda>\lambda_c$, and
(b) if $\rho(\lambda)>0$, then $r_t^-\to\infty$ in probability as $t\to\infty$ [3].

Result (a) shows that $\alpha(\lambda_c)=0$. It is based on a percolation idea which we learned from H. Kesten (cf. the important paper [23] in this connection). Result (b) says, in spirit, that if the right edge process visits the negative half-line infinitely often with positive probability, then $(\xi_t^0(\lambda))$ dies out with probability one. This assertion should be plausible in light of Theorem 3; the rigorous proof involves 0–1 law considerations. One can show using (b) that if the critical processes are nonergodic, then (23) holds at $\lambda=\lambda_c$.

Of course (a) and (b) are both consistent with either ergodicity or nonergodicity, and there is really no compelling evidence one way or the other. Based on very loose analogies with other systems one is inclined to suspect that ergodicity occurs. This is surely the premiere open problem in the theory.

Our next result gives a maximal convergence rate for the critical contact process (ξ_t), and a sample path consequence. In contrast to the subcritical case, weak convergence to $\delta_\emptyset$ must occur slowly, and the infected set ξ_t includes each finite subset of Z at arbitrarily large times with probability one. (This result is only of interest if the critical processes turn out to be ergodic.)

Theorem 12. *There is a positive constant K such that the critical contact process $(\xi_t(\lambda_c))$ satisfies*

$$\mathbf{P}(x\in\xi_t(\lambda_c))\geq K(1+t)^{-1}. \tag{43}$$

Moreover, for any $A\in S_0$,

$$\mathbf{P}(A\subset\xi_t(\lambda_c)\ \textit{for arbitrarily larg times}\ t)=1. \tag{44}$$

Proof. By $\alpha(\lambda_c)\geq 0$ and (17), we can choose $k>0$ large enough that

$$\mathbf{P}(r_t^-\ \text{ever}<-t-k)=\mathbf{P}(l_t^+\ \text{ever}>t+k)=p<\tfrac{1}{2}.$$

Making use of (20) and translation invariance, for any n it follows that

$$\mathbf{P}(\tau^{[-n-k,n+k]} > n) \geq \mathbf{P}(r_t^{(-\infty,n+k]} \geq 0 \text{ and } l_t^{[-k-n,\infty)} \leq 0 \ \forall t \leq n)$$
$$\geq 1 - 2p = \varepsilon > 0.$$

Now set additivity gives

$$\mathbf{P}(\tau^0 > n) \geq \frac{\mathbf{P}(\tau^{[-n-k,n+k]} > n)}{2(n+k)+1} \geq \frac{\varepsilon}{2(n+k)+1}.$$

Finally, by monotonicity, (43) holds with $K = \frac{1}{2}\varepsilon/(k+1)$. The argument for (44) is based on one in [12], but is simple enough that we can give it here. First we note that it suffices to prove (44) for $A = \{0\}$; since

$$\mathbf{P}(A \subset \xi_1^0) > 0 \quad \text{for any } A \in S_0,$$

the general case will then follow by set monotonicity and a standard application of the Markov property. Define

$$\sigma_t = \min\{s \geq t : 0 \in \xi_s\} \quad (-\infty \text{ if no such } s \text{ exists}),$$

and note that

$$\{0 \in \xi_s \text{ for arbitrarily large } s\} = \lim_{t\to\infty} \lim_{u\to\infty} \{\sigma_t \in [t, u]\}.$$

Thus we need only show that

$$\lim_{u\to\infty} \mathbf{P}(\sigma_t \in [t, u]) = 1 \quad \text{for each } t \geq 0. \tag{45}$$

Introduce

$$e_{t,u}^A = \mathbf{E}[\text{time in } [t, u]: 0 \in \xi_\cdot^A], \qquad e_{t,u} = e_{t,u}^Z.$$

Applying the strong Markov property and set monotonicity, we get

$$e_{t,u} = \int_t^u \int_S \mathbf{P}(\sigma_t \in \mathrm{d}r, \xi_{\sigma_t} \in \mathrm{d}A) e_{0,u-r}^A$$
$$\leq \int_t^u \int_S \mathbf{P}(\sigma_t \in \mathrm{d}r, \xi_{\sigma_t} \in \mathrm{d}A) e_{0,u}$$
$$= \mathbf{P}(\sigma_t \in [t, u]) e_{0,u}.$$

Thus,

$$\mathbf{P}(\sigma_t \in [t, u]) \geq \frac{e_{tu}}{e_{0u}} = 1 - \frac{e_{0t}}{e_{0u}}. \tag{46}$$

By (43),

$$e_{\mathrm{C}u} \geq \int_0^u K(1+r)^{-1}\,\mathrm{d}r \to \infty \quad \text{as } u \to \infty. \tag{47}$$

Use (46) and (47) to check (45), and the proof is finished.

8. Approximation by finite systems

Let $\{{}_N\xi_t^A(\lambda));\ A \subset [0, N)\}$ be the contact system with parameter λ on the torus of sites $[0, N) = \{0, 1, \ldots, N-1\}$, where 0 and $N-1$ are neighbors. For each N, this family of finite Markov chains has the single absorbing state $\emptyset$, and hence

$$e_N(\lambda) = \mathbf{E}[{}_N\tau^{[0,N)}] < \infty.$$

The critical phenomenon which occurs in the infinite systems is reflected in the finite systems by the growth rate of e_N, as $N \to \infty$. Namely, for small λ e_N grows logarithmically, whereas for large λ the growth is exponential. We now establish this phenomenon, which was noted for related discrete time systems by Stavskaya and Piatetskii–Shapiro [35] and Toom [38], as an application of techniques developed in earlier sections.

Theorem 13. *If $\lambda < \lambda_c$, then*

$$0 < \liminf_{N\to\infty} \frac{e_N(\lambda)}{\log N} \leq \limsup_{N\to\infty} \frac{e_N(\lambda)}{\log N} < \infty. \tag{48}$$

If $\lambda > 7$, then there are constants $1 < c_\lambda \leq C_\lambda < \infty$ such that

$$\liminf_{N\to\infty} c_\lambda^{-N} e_N(\lambda) > 0, \tag{49}$$

$$\limsup_{N\to\infty} C_\lambda^{-N} e_N(\lambda) < \infty. \tag{50}$$

Proof. Abbreviate $p_N(t) = \mathbf{P}({}_N\tau^{[0,N]} > t)$. All of the inequalities are derived by first estimating $p_N(t)$, and then using the tail formula

$$e_N = \int_0^\infty p_N(t)\,\mathrm{d}t. \tag{51}$$

The left inequality of (48) is based on comparison with a system where infected sites recover independently at rate 1, but infection cannot occur. Let τ_x, $0 \leq x \leq N-1$, be the time until a δ appears at site x in the graphical representation of $({}_N\xi_t^{[0,N)})$. Clearly ${}_N\tau^{[0,N)} \geq \max_x \tau_x$, so

$$p_N(t) \geq 1 - [\mathbf{P}(\tau_0 \leq t)]^N = 1 - (1 - \mathrm{e}^{-t})^N.$$

Using (51),

$$e_N \geq \log N\, p_N(\log N) \geq \log N\left[1-\left(1-\frac{1}{N}\right)^N\right].$$

We conclude that

$$\liminf_{N\to\infty} \frac{e_N(\lambda)}{\log N} \geq 1-\mathrm{e}^{-1} > 0.$$

The right inequality of (48) is, of course, more involved. To begin, define a comparison process $(\bar{\xi}_t^{[0,N)})$ on $[0, N)$ in terms of the contact process $(\xi_t^{[0,N)})$ on Z by:

$$x \in \bar{\xi}_t^{[0,N)} \text{ iff } x+kN \in \xi_t^{[0,N)} \quad \text{for some } k \in Z.$$

It is not hard to see that $({}_N\xi_t^{[0,N)})$ can be constructed on the same probability space as $(\bar{\xi}_t^{[0,N)})$ in such a way that

$${}_N\xi_t^{[0,N)} \subset \bar{\xi}_t^{[0,N)} \quad \text{for all } t.$$

Then

$$\{r_t^{[0,N)} < 0 \text{ and } l_t^{[0,N)} \geq N\}$$
$$\subset \{\xi_t^{[0,N)} = \emptyset\} \subset \{\bar{\xi}_t^{[0,N)} = \emptyset\} = \{{}_N\tau^{[0,N)} \leq t\}.$$

Hence we arrive at the estimate

$$p_N(t) \leq 2\mathbf{P}(r_t^- > -N).$$

Now assume $\lambda < \lambda_c$, and let C be as in (43). Writing $N(t) = \mathrm{e}^{Ct}$, for $t \geq C^{-1}\log N$ we have

$$p_N(t) \leq p_{N(t)}(t) \leq 2P(r_t^- > -N(t)) \leq 2C_1\,\mathrm{e}^{-C_2 t}$$

for some positive C_1 and C_2. Hence, from (51),

$$e_N \leq C^{-1}\log N + 2C_1 \int_{C^{-1}\log N}^{\infty} \mathrm{e}^{-C_2 t}\,\mathrm{d}t.$$

We conclude that

$$\limsup_{N\to\infty} \frac{e_N(\lambda)}{N} \leq C^{-1} < \infty,$$

as desired.

The derivation of (50) is based on comparison with a system where infected sites recover at rate 1 and healthy sites become infected at rate 2λ, independently of all other sites. Using this comparison, it is easy to see that for any $A \subset [0, N)$,

$$\mathbf{P}({}_N\tau^A \leq 1) \geq \left[\frac{1}{1+2\lambda}(1-\mathrm{e}^{-(1+2\lambda)})\right]^N,$$

the right side being the probability that the independent system starting from $[0, N)$ is in state $\emptyset$ at time 1. Thus, by a standard Markov chain argument,

$$p_N(n) \leq \left\{1 - \left[\frac{1}{1+2\lambda}(1-e^{-(1+2\lambda)})\right]^N\right\}^n.$$

By (51),

$$e_N \leq \sum_{n=0}^{\infty} p_N(n) \leq \left[\frac{1}{1+2\lambda}(1-e^{-(1+2\lambda)})\right]^{-N}.$$

Thus (50) holds with $C_\lambda = (1+2\lambda)/(1-e^{-(1+2\lambda)})$. For the remaining inequality (49), we make use of contours. In the graphical representation of $({}_N\xi_t^{[0,N)})$, the event $\{{}_N\tau^{[0,N)} \leq t\}$ implies that there is a 'contour' starting at some $(0, s)$, $s \leq t$, labelled with a δ, and then wrapping around the torus. The estimate (11) applies when $\lambda \geq 7$ to yield

$$p_N(t) \geq 1 - \int_0^t C(\lambda)A_1^N \, ds = 1 - C(\lambda)A_1^N t.$$

Thus, from (51),

$$e_N(\lambda) \geq \int_0^{C^{-1}(\lambda)A_1^{-N}} [1 - C(\lambda)A_1^N t] \, dt = \frac{2}{C(\lambda)} A_1^{-N}.$$

We conclude that (49) holds with $c_\lambda = A_1^{-1}$.

Remark. At $\lambda = \lambda_c$ one can use ideas from Theorem 12 to show that

$$\liminf_{N\to\infty} \frac{e_N(\lambda)}{N} > 0.$$

Presumably N is the correct critical growth rate, and presumably $e_N(\lambda)$ grows exponentially for all $\lambda > \lambda_c$.

9. Discrete time, oriented percolation and growth models

There are two discrete models, essentially equivalent, which are amenable to much of the analysis we have carried out. The first is the system of one-sided discrete time contact processes, often called the 'Russian lamps' [35, 38, 41, 42, 43]. Given $p \in [0, 1]$, for each $A \subset Z$ a discrete time S-valued Markov process $(\xi_n^A(p))$ is defined inductively as follows. To begin, $\xi_0^A(p) = A$. At time $n \geq 1$, a coin with probability p of heads is flipped independently at each site $x \in Z$. If $\xi_{n-1}^A \cap \{x-1, x\} \neq \emptyset$, and if the coin turns up heads, then $x \in \xi_n^A$. Otherwise $x \notin \xi_n^A$.

The second model is called oriented site percolation in the plane [1, 4, 9, 14, 27, 31]. Here we let μ_p be Bernoulli product measure on Z^2 with density $p \in [0, 1]$, and introduce the oriented graph structure:

$$(x, y) \text{ has neighbors } (x+1, y) \text{ and } (x, y+1), \quad (x, y) \in Z^2.$$

Let η be a μ_p-distributed random field of 0's and 1's on Z^2. Define random subsets C^A of Z^2, $A \subset Z^2$, by

$$(x, y) \in C^A \text{ iff there is a nearest neighbor path from some site of } A \text{ to } (x, y) \text{ arriving only at 1's in } \eta.$$

(By convention, $A \subset C^A$.)

The two models are equivalent in the sense that a representation of $\{(\xi_n^A(p))\}$ is given by

$$x \in \xi_n^A(p) \text{ iff } (x, n-x) \in C^B(p), \quad \text{where } B = \{(y, -y) \colon y \in A\}.$$

In particular,

$$x \in \xi_n^0(p) \text{ iff } (x, n-x) \in C^0(p).$$

The central objects of study in both models are

$$\rho(p) = \mathbf{P}(\xi_n^0 \neq \emptyset \text{ for all } n) = \mathbf{P}(|C^0| = \infty),$$

and the critical constants

$$p_{\mathrm{H}} = \sup\{p \colon \rho(p) = 0\}, \qquad p_{\mathrm{T}} = \sup\{p \colon \mathbf{E}[|C^0|] < \infty\}.$$

As for continuous contact systems, the edge processes

$$r_n^- = \max\{x \in \xi_n^{(-\infty,0]}\}, \qquad l_n^+ = \min\{x \in \xi_n^{[0,\infty)}\},$$

and the expected displacement at time n:

$$\alpha_n(p) - \beta_n(p) = \mathbf{E}[r_n^-(p) - l_n^+(p)]$$

are central to the analysis. The discrete and continuous theories have a great deal in common, though each enjoys many features not shared by the other.

We now present solutions to two outstanding problems in oriented percolation, both of which are easy in light of results and techniques from previous sections.

Theorem 14. (a) *If p_n is the unique solution of $\alpha_n(p) = 0$, then*

$$p_n \to p_{\mathrm{H}} \quad \text{as } n \to \infty.$$

(b) $p_{\mathrm{T}} = p_{\mathrm{H}}$. *In fact,*

$$\mathbf{E}[|C^0|] \begin{cases} < \infty, & p < p_{\mathrm{H}} \\ = \infty, & p \geq p_{\mathrm{H}}. \end{cases}$$

Proof. (a) Just as in the proof of Theorem 4 one can show that $\sup_n p_n = p_H$. Moreover, the argument given there applies equally well to any subsequence (n'), since

$$\alpha(p) = \lim_{n\to\infty} \frac{\alpha_n(p)}{n} = \inf_{n>0} \frac{\alpha_n(p)}{n}.$$

Thus $\sup_{n'} p_{n'} = p_H$ for any (n'), which proves (a).

(b) Arguments similar to ones in Section 5 and 8 show that for $p < p_H$,

$$\mathbf{P}(r_\cdot^0 - l_\cdot^0 \text{ ever} > n) \to 0 \quad \text{exponentially fast as } n \to \infty, \tag{52}$$

$$\mathbf{P}(\xi_n^0 \neq \emptyset) \to 0 \quad \text{exponentially fast as } n \to \infty, \tag{53}$$

whereas for $p \geq p_H$,

$$\mathbf{P}(\tau^0 > n) \geq c(1+n)^{-1} \quad \text{for some } c > 0. \tag{54}$$

Using (52) and (53), if $p < p_H$, then

$$\begin{aligned}\mathbf{P}(|C^0| > n^2) &\leq \mathbf{P}(r_\cdot^0 - l_\cdot^0 \text{ ever} > n \text{ or } \tau^0 > n)\\ &\to 0 \quad \text{exponentially fast as } n \to 0.\end{aligned}$$

Thus

$$\mathbf{E}[|C^0|] \leq \sum_{n=0}^{\infty} (2n+1)\mathbf{P}(|C^0| > n^2) < \infty.$$

(Similarly, all moments of $|C^0|$ are finite for $p < p_H$.)

If $p \geq p_H$, then since $\{\xi_n^0 \neq \emptyset\} \subset \{|C^0| > n\}$, by (51),

$$\mathbf{E}[|C^0|] \geq \sum_{n=0}^{\infty} \mathbf{P}(\xi_n^0 \neq \emptyset) = \infty.$$

Remarks. In the discrete case the α_n are polynomials, so the p_n are computable as long as one has the patience. Thus $p_1 = \frac{2}{3}$, $p_2 \geq 0.672$, $p_3 \geq 0.676$. Probably the p_n are increasing, but this is not obvious. In the introduction to [27] it was implied that our (a) would be proved, but the proof was not given. While it is comforting to know that the p_n converge to the critical constant, the method of 'ceilings' in [9] gives better bounds for small n. Thus it is known (cf. [9]) that

$$p_H \geq 0.688.$$

Stochastic growth models have been widely studied, especially since the pioneering paper by Richardson [29]. We refer the reader to [29] for a handful of different growth models. Here we will mention only one of his processes, for which the discrete time contact systems come into play in a olorful way. The so-called Gp model we have in mind evolves as follows. At time 0 a single infected site occupies the origin of Z^2. At time $n \geq 1$, if site $(x, y) \in Z^2$ is healthy and at least one of its four neighbors is infected, then (x, y) becomes infected at time $n+1$ independently with probability $p \in [0, 1]$, and remains healthy with probability $1-p$. A site which is infected remains so forever. Let $A_n \subset R^2$ be the set of infected sites at time n, together with a unit square centered at each such infected site. Clearly A_n grows over time. Richardson [29] has proved that there is a norm f_p on R^2 such that for any $\varepsilon > 0$,

$$\lim_{n\to\infty} \mathbf{P}\left(\{f_p \le 1-\varepsilon\} \subset \frac{A_n(p)}{n} \subset \{f_p \le 1+\varepsilon\}\right) = 1.$$

Thus A_n has an asymptotic shape:

$$B_p = \{f_p \le 1\}.$$

Recently, Durrett and Liggett [7] have made the nice observation that

$$\xi_n^0 = \{x \in Z: (x, n-x) \in A_n(p)\}$$

defines a one-sided contact process with the same parameter p. If $p > p_{\mathrm{H}}$, then with positive probability (ξ_n^0) grows linearly in diameter forever, which means that the boundary of B_p intersects the line $x+y=1$ in an interval around $(\frac{1}{2}, \frac{1}{2})$. Moreover, by using the discrete time version of Theorem 7, Durrett and Liggett are able to show that if $p < p_{\mathrm{H}}$, then $B_p \cap \{x+y=1\} = \emptyset$. Thus the critical value for 'flat edges' in the asymptotic shape of the Gp growth model is precisely p_{H}. The reader is referred to [7] for details.

10. Open problems; generalizations

In this final section we indicate some possible directions for further research on contact processes. By now the reader should be aware that the theory of the basic contact process on Z is fairly complete. Nevertheless, a number of open problems have emerged during the exposition. Here are 10 of the more important ones; most are probably rather difficult.

Open problems:

(1) Is $\rho(\lambda_c) = 0$?
(2) Can λ_c be 'explicitly' computed?
(3) Is there an 'explicit' sequence of upper bounds for λ_c which converges to λ_c?
(4) Is $\alpha(\lambda)$ continuous (concave) for $\lambda > \lambda_c$?
(5) Is $\lambda_\varphi = \lambda_c$?
(6) Is $\lambda_\psi = \lambda_c$?
(7) Does Harris' Growth Theorem hold for all $\lambda > \lambda_c$?
(8) Does the central limit theorem hold for ν_λ, for all $\lambda > \lambda_c$?
(9) Is the density of $\xi_t(\lambda_c)$ of order $1/t$?
(10) Does $e_N(\lambda)$ grow exponentially in N for all $\lambda > \lambda_c$?

Another possible avenue for further work is the analysis of more general contact processes. The basic system which we have discussed can be generalized in various ways. As long as the recovery rate is constant and the rate of infection is 0 in a totally healthy local environment, then (ξ_t^A) is still called a contact process [17]. One can stick to the one dimensional nearest neighbor case, but consider a more general class of infection rates. A great many of the methods of this paper will continue to apply as long as the infection rates are additive [19]; some methods only require monotonicity of the rates. As soon as one moves to the non-nearest neighbor case, however, most of the techniques we have used fail to apply. While the main theorems undoubtedly continue to hold in such cases, new ideas, as yet undiscovered, are needed. A recent paper by Bramson and Gray [2] studies an infinite range contact process, and raises some intriguing questions.

Finally, one can consider systems in several dimensions. Of these, the simplest is the *basic* (nearest neighbor) *contact process on* Z^d, for which infection takes place at a rate proportional to the number of infected neighbors. As for $d=1$, the proportionality parameter is λ, and the recovery rate is 1. Again, the methods of this paper fail to apply for the most part. The existence of a critical $\lambda_c^{(d)}$ can be proved just as in the one dimensional case, but the ergodic theory is much less developed for $d>1$. The lower bounds $\lambda_c^{(d)} \geqslant 1/(2d-1)$ are proved in [18] and [20]. The best known upper bound in two dimensions, $\lambda_c^{(2)} \leqslant 1$, is in [21]. Recently, Holley and Liggett [22] have obtained upper bounds for $d \geqslant 3$ which are asymptotically equal to $1/2d$ as $d \to \infty$. Thus $\lim_{d\to\infty} d\lambda_c^{(d)} = \frac{1}{2}$; conceptually, interaction disappears in the limit. Virtually all of the known convergence results in several dimensions are due to Harris [18], [19]; unfortunately, he requires regularity assumptions on the initial state. A forthcoming paper of Durrett and Griffeath [6] will establish stronger theorems for the basic nonergodic processes with sufficiently large λ. For example, it will be shown that complete convergence (23) holds in any dimension d provided that $\lambda > \lambda_c^{(1)}$. More research along these lines would be welcome.

Added in proof

New results of Durrett and the author answer questions 5, 6, 7, 8 and 10 in the affirmative.

References

[1] J. Bishir, A lower bound for the critical probability in the one-quadrant oriented-atom percolation process, J. Roy. Statist. Soc. 25 (1963) 401–404.
[2] M. Bramson and L. Gray, A note on the survival of the long-range contact process, preprint (1980).
[3] M. Bramson, L. Gray and D. Griffeath, unpublished (1980).
[4] S. Broadbent and J. Hammersley, Percolation processes I. Crystals and mazes, Proc. Cambridge Philos. Soc. 53 (1957) 629–645.

[5] R. Durrett, On the growth of one dimensional contact processes. Ann. Probability 8 (1980) 890–907.
[6] R. Durrett and D. Griffeath, Contact processes in several dimensions, preprint (1980).
[7] R. Durrett and T. Liggett, The shape of the limit set in Richardson's growth model, preprint (1980).
[8] L. Gray and D. Griffeath, A stability criterion for attractive nearest neighbor spin systems on Z, Ann. Probability, to appear.
[9] L. Gray, R. Smythe and J. Wierman, Lower bounds for the critical probability in percolation models with oriented bonds, preprint (1980).
[10] D. Griffeath, Limit theorems for nonergodic set-valued Markov processes, Ann. Probability 8, (1978) 379–387.
[11] D. Griffeath, Pointwise ergodicity of the basic contact process, Ann. Probability 7 (1979) 139–143.
[12] D. Griffeath, Additive and Cancellative Interacting Particle Systems, Lecture Notes in Mathematics 724 (Springer, New York, 1979).
[13] R. Griffiths, The Peierls argument for the existence of phase transitions, in: J.C.T. Pool, Ed., Mathematical Aspects of Statistical Mechanics, SIAM–AMS Proceedings, Vol. 5 (Amer. Math. Soc., Providence, 1972) 13–26.
[14] J. Hammersley, Bornes supérieures de la probabilité critique dans un processus de filtration, in: Le Calcul des Probabilitiés et ses Applications (Centre National de la Recherche Scientifique, Paris, 1959) 17–37.
[15] T.E. Harris, Lower bound for the critical probability in a certain percolation process, Proc. Cambridge Philos. Soc. 56 (1960) 13–20.
[16] T.E. Harris, Nearest neighbor Markov interaction processes on multi-dimensional lattices, Advances in Math. 9 (1972) 66–89.
[17] T.E. Harris, Contact interactions on a lattice, Ann. Probability 2 (1974) 969–988.
[18] T.E. Harris, On a class of set-valued Markov processes, Ann. Probability 4 (1976) 175–194.
[19] T.E. Harris, Additive set-valued Markov processes and graphical methods, Ann. Probability 6 (1978) 355–378.
[20] R. Holley and T.M. Liggett, Ergodic theorems for weakly interacting infinite systems and the voter model, Ann. Probability 3 (1975) 643–663.
[21] R. Holley and T.M. Liggett, The survival of contact processes, Ann. Probability 6 (1978) 198–206.
[22] R. Holley and T.M. Liggett, Generalized potlatch and smoothing processes, preprint (1980).
[23] H. Kesten, The critical probability of bond percolation on the square lattice equals 1/2, Commun. Math. Phys. 74 (1980) 41–59.
[24] T.M. Liggett, the stochastic evolution of infinite systems of interacting particles, Lecture Notes in Mathematics 598 (Springer, New York, 1977) 187–248.
[25] T.M. Liggett, Attractive nearest neighbor spin systems on the integers, Ann. Probability 6 (1978) 629–636.
[26] V.A. Malyšev, The central limit theorem for Gibbsian random fields, Soviet Math. Dokl. 16 (1975) 1141–1145.
[27] J.G. Mauldon, Assymetric oriented percolation on a plane, Proc. 4th Berkeley Symposium, Vol. 2 (1961) 337–345.
[28] C. Neaderhouser, Some limit theorems for random fields, Comm. Math. Phys. 61 (1978) 293–305.
[29] D. Richardson, Random growth in a tesselation, Proc. Cambridge Philos. Soc. 74 (1973) 515–528.
[30] V.K. Shante and S. Kirkpatrick, Introduction to percolation theory, Advances in Physics 20 (1971) 325–357.
[31] R.T. Smythe and J.C. Wierman, First Passage Percolation on the Square Lattice Lecture Notes in Mathematics 671, (Springer, New York, 1978).
[32] F. Spitzer, Interaction of Markov processes, Advances in Math. 5 (1970) 246–290.
[33] F. Spitzer, Random fields and interacting particle systems M.A.A. Summer Seminar Notes, Williamstown, MA (1971).
[34] F. Spitzer, Introduction aux processus de Markov à paramètres dans Z_ν, Lecture Notes in Mathematics 390 (Springer, New York, 1974).
[35] O.N. Stavskaya and I.I. Piatetskii-Shapirio, On homogeneous networks of spontaneously active elements, Problemy Kibernet. 20 (1968) 91–106.
[36] W.F. Stout, Almost Sure Convergence (Academic Press, New York, 1974).

[37] D. Stroock, Lectures on Infinite Interacting Systems, Kyoto University Lectures in Mathematics 11 (1978).
[38] A.L. Toom, A family of uniform nets of formal systems, Soviet Mathematics 9 (1968) 1338–1341.
[39] A.L. Toom, Nonergodic multidimensional systems of automata, Problemy Peredachi Informatsii 10 (1974) 239–246.
[40] L.N. Vasershtein, Markov processes over denumerable products of spaces, describing large systems of automata, Problemy Peredachi Informatsii 5 (3) (1969) 64–73.
[41] L.N. Vasershtein and A.M. Leontovich, Invariant measures of certain Markov operators describing a homogeneous random medium, Problemy Pereadachi Informatsii 6 (1) (1970) 71–80.
[42] N.B. Vasilev, Limit behavior of one random medium, Problemy Peredachi Informatsii 5 (4) (1969) 68–74.
[43] N.B. Vasil'ev, L.G. Mityushin, I.I. Pyatetskii-Shapiro and A.L. Toom, Stavskaya Operators (In Russian), preprint no. 12, Institute of Applied Math., Academy of Sciences, U.S.S.R., Moscow (1973).