EFFECT OF SCALE ON LONG-RANGE RANDOM GRAPHS AND CHROMOSOMAL INVERSIONS

BY NATHANAËL BERESTYCKI AND RICHARD PYMAR

University of Cambridge

We consider bond percolation on n vertices on a circle where edges are permitted between vertices whose spacing is at most some number $L = L(n)$. We show that the resulting random graph gets a giant component when $L \gg (\log n)^2$ (when the mean degree exceeds 1) but not when $L \ll \log n$. The proof uses comparisons to branching random walks. We also consider a related process of random transpositions of n particles on a circle, where transpositions only occur again if the spacing is at most L. Then the process exhibits the mean-field behavior described by Berestycki and Durrett if and only if $L(n)$ tends to infinity, no matter how slowly. Thus there are regimes where the random graph has no giant component but the random walk nevertheless has a phase transition. We discuss possible relevance of these results for a dataset coming from D. repleta and D. melanogaster and for the typical length of chromosomal inversions.

1. Introduction and results.

1.1. *Random graphs results.* Let $n \geq 1$ and let $L = L(n) \geq 1$. Define vertex set $V = \{1, \ldots, n\}$ and edge set $\mathcal{R}_L = \{(i, j) \in V^2, \|i - j\| \leq L\}$, where $\|i - j\|$ denotes the cyclical distance between i and j, that is, $\|u\| = \min(|u|, n - |u|)$ for $u \in V$. In this paper we consider bond percolation on V where each edge in $\mathcal{R}_L$ is open with probability p. Equivalently, let $(G(t), t \geq 0)$ be the random graph process where a uniformly chosen edge of $\mathcal{R}_L$ is opened in continuous time, at rate 1. Let $\Lambda^1(t) \geq \Lambda^2(t) \geq \cdots$ denote the ordered component sizes of $G(t)$. At a fixed time t this corresponds to the above model with $p = 1 - \exp\{-t/(nL)\}$. When $L = 1$, this is the usual bond percolation model on the cycle of length n, while for $L(n) = n/2$, we find that $G(t)$ is a realization of the much studied random graph model of Erdős and Renyi [see Bollobás (1985) and Durrett (2010) for background]. Hence, our random graph model interpolates between these two cases.

In this paper we are interested in the properties of the connected components of $G(t)$, particularly those related to the possible emergence of a giant component when the average degree exceeds 1. The main result of this paper shows that this

depends on the scale $L(n)$. To state our results, we let $c > 0$ and consider $t = cn/2$, so that the expected degree of a given vertex in $G(t)$ converges to c when $n \to \infty$. Let $\Lambda^1(t) \geq \Lambda^2(t) \geq \cdots$ denote the ordered component sizes of $G(t)$.

THEOREM 1. *Let $t = cn/2$, where $c > 0$ is fixed as $n \to \infty$.*

(i) *If $c < 1$, then there exists $C < \infty$ depending only on c such that $\Lambda^1(t) \leq C \log n$ with high probability as $n \to \infty$.*

(ii) *If $c > 1$ and there exists $\xi > 0$ such that $L(n) \geq (\log n)^{2+\xi}$, then there is a unique giant component; more precisely,*

$$(1) \qquad \frac{\Lambda^1(t)}{n} \to \theta(c)$$

in probability as $n \to \infty$, where $\theta(c)$ is the survival probability of a Poisson(c) *Galton–Watson tree. Moreover, $\Lambda^2(t)/n \to 0$ in probability.*

(iii) *However, if $c > 1$ and $L = o(\log n)$, then for all $a > 0$,*

$$(2) \qquad \frac{\Lambda^1(t)}{n^a} \to 0$$

in probability as $n \to \infty$. In particular there are no giant components.

Statement (i) is fairly easy to prove using the standard technique of approximating the size of a component in the graph by the total progeny of a branching process. The result follows since in the case $c < 1$ we know that the total progeny of the branching process is almost surely finite and has exponential tails.

Part (ii) is the most challenging. We start by noting that the exploration of the component containing a given vertex v may be well-approximated by the trace of a branching random walk where the step distribution is uniform on $\{-L, \ldots, L\}$. This approximation is valid so long as the local density of the part of the component already explored stays small. Thus, showing the existence of a giant component requires a balancing act; we need to ensure that the local density of what we explore stays small enough to ignore self-intersections, but large enough for global connections to occur. Careful estimates on survival probabilities of killed branching random walks are used to achieve this.

Part (iii) is the easiest to prove, and requires showing the existence of many "blocking" intervals of size L which consist just of vertices with degree 0. When there are many such intervals, no giant component can exist.

1.2. *Long-range random transpositions.* Theorem 1 was originally motivated by the study of a question concerning long-range transpositions, which may itself be rephrased as a question in computational biology. We now discuss the question on long-range random transpositions and delay the applications to comparative genomics until Section 2.

Recall the definitions of V and $\mathcal{R}_L$ in Section 1.1. Consider a random process $(\sigma_t, t \geq 0)$ with values in the symmetric group $\mathcal{S}_n$, which evolves as follows. Initially, $\sigma_0 = e$ is the identity permutation. Let $(i_1, j_1), (i_2, j_2), \ldots$ be an i.i.d. infinite sequence of pairs of elements of V, where each pair is uniformly distributed on $\mathcal{R}_L$. Then we put

$$\sigma_t = \tau_{N_t} \circ \cdots \circ \tau_1, \tag{3}$$

where for each $k \geq 1$ we let τ_k denote the transposition (i_k, j_k), $(N_t, t \geq 0)$ is an independent Poisson process with rate 1 and $\circ$ the composition of two permutations. That is, informally, if we view the permutation σ_t as describing the positions on the circle of n particles labeled by V [with $\sigma_t(i)$ denoting the position of particle $i \in V$], then in continuous time at rate 1, a pair of positions (i, j) is sampled uniformly at random from $\mathcal{R}_L$ and the two particles at positions i and j are swapped. Thus the case where $L(n) \geq n/2$ corresponds to the well-known random transposition process (i.e., the composition of uniform random transpositions), whereas the case where $L(n) = 1$ corresponds to the case of random adjacent transpositions on the circle.

Our interest consists of describing the time-evolution of $\delta(\sigma_t)$, where for all $\sigma \in S_n$ we set $\delta(\sigma) = n - |\sigma|$ and $|\sigma|$ to be the number of cycles of σ. By a well-known result of Cayley, this is the length of a minimal decomposition of σ into a product of *any* transpositions (i.e., whose range is not necessarily restricted to $\mathcal{R}_L$). The reason for this choice will become apparent in subsequent sections and is motivated by the applications to comparative genomics.

For $c > 0$, define a function

$$u(c) = 1 - \sum_{k=1}^{\infty} \frac{1}{c} \frac{k^{k-2}}{k!} (ce^{-c})^k. \tag{4}$$

It is known that $u(c) = c/2$ for $c \leq 1$ but $u(c) < c/2$ for $c > 1$ [see, e.g., Bollobás (1985), Theorem 5.12]. The function u is continuously differentiable but has no second derivative at $c = 1$. We shall prove the following results.

THEOREM 2. *Assume $L(n) \to \infty$ as $n \to \infty$. Then we have the following convergence in probability as $n \to \infty$: for all $c > 0$,*

$$\frac{1}{n} \delta(\sigma_{cn/2}) \to u(c). \tag{5}$$

In this result the distance between the two points being transposed at every transposition is uniform within $\{1, \ldots, L(n)\}$. We will prove in Theorem 6 given in Section 5 a more general version of this result, where this length is allowed to be some arbitrary distribution subject to the condition that there are no "atoms in the limit," which is the equivalent of requiring here $L(n) \to \infty$.

By contrast, the microscopic regime (where L is assumed to be constant or to have a limit) shows a remarkably different behavior.

THEOREM 3. *Assume* $\lim_{n\to\infty} L(n)$ *exists. Then we have convergence in probability*; *for all* $c > 0$,

$$\frac{1}{n}\delta(\sigma_{cn/2}) \to v(c)$$

as $n \to \infty$, *for some* C^2 *function* $v(c)$ *which satisfies* $0 < v(c) < c/2$ *for all* $c > 0$.

As we will describe in greater detail later on, there is a connection between the long-range random transposition process and the random graph process of Theorem 1. Roughly speaking, when $L(n)$ is bounded, we expect $v(c) < c/2$ because each new edge has a positive probability of having its two endpoints in the same connected component. Alternatively, the branching random walk which is used to explore the connected component of a vertex has a positive probability of making a self-intersection at every new step.

The mean-field case where $L(n) = n/2$ recovers Theorem 4 of Berestycki and Durrett (2006). Theorem 2 above relies on a coupling with the random graph $G(t)$ of Theorem 1; this coupling is similar to the coupling with the Erdős–Renyi random graph introduced in Berestycki and Durrett (2006). In that paper, the emergence of the giant component in the Erdős–Renyi random graph was a crucial aspect of the proofs. As a result, one might suspect that the phase transition of $\delta(\sigma_t)$ is a direct consequence of the emergence of a giant component in the random graph. However, one particularly surprising feature of Theorem 2 above is the fact that the limiting behavior described by (5) holds for all $L(n) \to \infty$, no matter how slowly. This includes in particular the cases where $L(n) = o(\log n)$ and the random graph $G(t)$ does *not* have a giant component. Hence, for choices of $L(n)$ such that $L(n) \to \infty$ but $L(n) = o(n)$, the quantity $\delta(\sigma_t)$ has a phase transition at time $n/2$, even though the random graph $G(t)$ does not get a giant component at this time.

1.3. *Relation to other work, and open problems. Long-range percolation.* A similar model has been studied by Penrose (1993). There the model considered is on the *infinite* square grid $\mathbb{Z}^d$, rather than the finite (one-dimensional) torus which we consider here. In the *infinite* case, $d = 1$ is trivial since percolation (occurrence of an infinite cluster) only occurs if $p = 1$ for obvious reasons. Penrose studied the case $d \geq 2$ and showed that if c is the expected degree of the origin, and L the maximum distance between the two ends of a bond, where the parameter $L \to \infty$ and c is fixed, then the percolation probability approaches $\theta(c)$ [where $\theta(c)$ is the same as in (1), i.e., the survival probability for a Galton–Watson process with Poisson(c) offspring distribution]. As is the case here, his arguments use a natural comparison with branching random walks.

It is interesting that, while the infinite case is essentially trivial when $d = 1$, the *finite-n* case is considerably more intricate than the infinite case, as witnessed by the different behaviors in (1) and (2) depending on how fast $L(n) \to \infty$. Regarding

the finite-n situation, it is an interesting open question to see whether there are giant components if $t = cn/2$ and $c > 1$ with $\log n \leq L(n) \leq (\log n)^2$. Another interesting problem concerns the size of the largest components when there is no giant component, in particular, if $L = o(\log n)$. Indeed, our proof makes it clear that when $L(n) \to \infty$, even if the largest component is not macroscopic, there is a positive proportion of vertices in components of *mesoscopic* size. We anticipate that as $c > 1$ is fixed and L increases, the size of the largest component, normalized by n, jumps from 0 to $\theta(c)$ as $L(n)$ passes through a critical threshold between $\log n$ and $(\log n)^2$. As pointed out by a referee, this is suggested by a work of Aizenman and Newman (1986) on long-range bond percolation on $\mathbb{Z}$ where the connection probability between vertices at distance $x > 0$ decays like $1/x^2$. Their main result (Proposition 1.1) shows that such discontinuities occur in this case.

Epidemic models. The question of giant components in random graph models can, as usual, be rephrased in terms of epidemic processes. More precisely, fix a vertex v and a number $p \in (0, 1)$. Consider an SIR epidemic model that begins with all vertices susceptible but vertex v infected. Once a vertex is infected, it transmits the infection to each of its neighbors in the base graph (V, E) at rate $\lambda > 0$ and dies or is removed at rate 1. Then the total size of the epidemic is equal to the size of the component containing v in the random graph with edge-probability $p = \lambda/(1 + \lambda)$. As pointed out by an anonymous referee, Bramson, Durrett and Swindle (1989) consider the related SIS model (or contact process) on $\mathbb{Z}^d$ where, as here, long-range connections are possible. Similar techniques are employed as in this article to calculate the critical rate of infection and the probability of percolation. Letting infections occur at rate $\lambda/\operatorname{Vol} B(L)$ where $B(L)$ is a ball or radius L in $\mathbb{Z}^d$, they show that the critical infection rate λ_c converges to 1 in all dimensions as $L \to \infty$. They also identify the rate of convergence, which turns out to depend on the dimension in an interesting way.

Higher-dimensional analogs of Theorem 1. Our proofs do not cover the higher-dimensional cases but it would not be very difficult to adapt them. In particular, the analogue of (1) would hold if $d \geq 2$ no matter how slowly $L(n) \to \infty$. In other words, only for the one-dimensional case is it important to have some quantitative estimates on $L(n)$. Intuitively this is because, in one dimension, one is forced to go through potentially bad regions whereas this problem does not arise in higher dimensions.

Regarding site percolation, we point out that recently Bollobás, Janson and Riordan (2009) have described an interesting behavior for a site percolation model on the torus in dimensions $d \geq 2$ where two vertices are joined if they agree in one coordinate and differ by at most L in the other. For $d = 2$ they show that the critical percolation probability, $p_c(L)$, satisfies $\lim_{L \to \infty} L p_c(L) = \log(3/2)$. This is surprising as the expected degree of a given vertex at the phase transition is then strictly greater than 1. There again, approximation by branching random walks plays an important role in the proof.

Slowdown transitions for random walks. In the mean-field case $L(n) = n/2$ of uniformly chosen random transpositions, the quantity $\delta(\sigma_t)$ may be interpreted as the graph-theoretic distance between the starting position of the random walk ($\sigma_0 = $ the identity element) and the current position of the walk. Theorem 2 in this case [which, as already mentioned, is Theorem 4 of Berestycki and Durrett (2006)], may thus be interpreted as a *slowdown* transition of the evolution of the random walk; at time $n/2$, the acceleration [second derivative of $\delta(\sigma_t)$] drops from 0 to $-\infty$. By contrast, Berestycki and Durrett (2008) studied the evolution of the graph-theoretic distance in the case of random adjacent transpositions. This essentially corresponds to the case $L = 1$, with the difference that the transposition $(1\ n)$ is not allowed. They found that no sudden transition occurs in the deceleration of the random walk. It would be extremely interesting to study the evolution of the graph-theoretic distance of the random walk when $L = L(n)$ is a given function that may or may not tend to infinity as $n \to \infty$. Unfortunately, this problem seems untractable at the moment as it is far from obvious how to compute (or estimate) the graph distance between two given permutations. [We note that even in the case $L = 1$ where the transposition $(1\ n)$ is allowed, this question is partly open; see Conjecture 3 in Berestycki and Durrett (2008).] Nevertheless it is tempting to take Theorems 2 and 3 as an indication that a slowdown transition for the random walk occurs if and only if $L(n) \to \infty$, with the phase transition always occurring at time $n/2$.

Organization of the paper. In Section 2 we show how Theorems 2 and 3 relate to a biological problem and in particular discuss the possible relevance of these results for a dataset coming from two Drosophila species. In Section 3 we state and prove results on the evolution of the clusters in a random graph which evolves in a more general way to $G(t)$. In Section 4 we give a proof of Theorem 1. Section 5 contains a proof of a result stronger than Theorem 2 using the more general random graph process defined in Section 3. Finally, in Section 6 we present the proof of Theorem 3.

2. Applications in comparative genomics.

2.1. *Statement of problem and history.* Part of the motivation for this paper comes from a biological background, more specifically, in answering a question about the evolution of the gene order of chromosomes. We begin with an example. In 2001 Ranz, Casals, and Ruiz located 79 genes on chromosome 2 of Drosophila repleta and on chromosome arm 3R of Drosophila melanogaster. While the genetic material is overall essentially identical, the order of the genes is quite different. If we number the genes according to their order in D. repleta then their order in D. melanogaster is given in Table 1.

Since the divergence of the two species, this chromosome region has been subjected to many reversals or chromosomal inversions, which are moves that reverse the order of whole gene segments. Because they involve many base pairs at a time

TABLE 1
Order of the genes in D. repleta compared to their order in D. melanogaster

36	37	17	40	16	15	14	63	10	9
55	28	13	51	22	79	39	70	66	5
6	7	35	64	33	32	60	61	18	65
62	12	1	11	23	20	4	52	68	29
48	3	21	53	8	43	72	58	57	56
19	49	34	59	30	77	31	67	44	2
27	38	50	26	25	76	69	41	24	75
71	78	73	47	54	45	74	42	46	

rather than the more common substitutions, insertions and deletions, these mutations are called large-scale. They are usually called inversions in the biology literature, but we stick with the word reversal as "inversions" is often used among combinatorists with a different meaning [see, e.g., Diaconis and Graham (1977)]. One question of interest in the field of computational biology is the following: *How many such reversals have occurred*?

Hannenhalli and Pevzner (1999) have devised a widely used algorithm which computes the *parsimony distance*, the minimal number of reversals that are needed to transform one chromosome into the other (this will be denoted here by d_∞). By definition, the number of reversals that did occur is at least d_∞. Berestycki and Durrett (2006) complemented this by rigorously analyzing the limiting behavior of the discrepancy between the true distance and the parsimony distance [described by the function $u(c)$ in Theorem 2], under the mean-field assumption that all reversals are equally likely.

However, that assumption does not seem to be entirely justified and it might be more accurate to restrict the length of the segment being reversed. According to Durrett (2003), "To seek a biological explanation of the nonuniformity we note that the gene-to-gene pairing of homologous chromosomes implies that if one chromosome of the pair contains an inversion that the other does not, a loop will form in the region in which the gene order is inverted.... If a recombination occurs in the inverted region then the recombined chromosomes will contain two copies of some regions and zero of others, which can have unpleasant consequences. A simple way to take this into account is... [to] restrict our attention to the L-reversal model." The reasoning here is that as the length of the segment reversed increases, the probability of recombination increases. Here, the L-reversal model is to allow only reversals that switch segments of, at most, length L and all such reversals have equal probability. A further argument can be seen in Durrett (2002) who argues that not all inversions occur at the same rate; when a large amount of DNA is absent from a chromosome, the offspring is typically not viable, so longer inversions will occur at a lower rate.

2.2. *Estimating the number of chromosomal inversions.* To estimate the number of chromosomal inversions (or reversals) in the long-range spatial model, one natural idea is to use the parsimony approach; that is, compute the d_L-distance (minimal number d_L of L-reversals needed to transform one genome into the other) and then prove a limit theorem for the evolution of $d_L(t)$ under random L-reversals. However, this appears completely out of reach at this stage; the crucial problem is that we do not know of any algorithm to compute the L-reversal distance. [Even in the case $L = 1$, if particles are lying on a circle, this is a delicate problem; see Conjecture 3 in Berestycki and Durrett (2008).] Thus, even if a limit theorem could be proved, we would not know how to apply it to two given genomes.

In order to tackle this difficulty, we propose here the following alternative approach. We keep looking at the d_∞-distance (minimal number of reversals needed to transform one chromosome into the other, no matter their length) but now we think of d_∞ only as an easily computed statistic on which we can make some inference, even though not all reversals were equally likely. More precisely, we are able to describe the evolution of the quantity $d_\infty(t)$ under the application of random L-reversals, and use that result to estimate t from the data $d_\infty(t)$.

We first state the result in this context, and illustrate our idea with a numerical example in Section 2.3. The distance d_∞ is defined in terms of an object known as the *breakpoint graph*. For definitions of these notions we refer the interested reader to Chapter 9 of Durrett (2002). For signed permutations σ, σ' we let $\hat{\delta}(\sigma, \sigma') = n + 1 - c$, where c is the number of components of the breakpoint graph. In general [see Durrett (2002), Theorem 9.1], $d_\infty(\sigma, \sigma') \geq \hat{\delta}(\sigma, \sigma')$. The quantity $\hat{\delta}(\sigma, \sigma')$ ignores obstacles known as "hurdles" and "fortresses of hurdles." All these are thought to be negligible in biologically relevant cases, so we will use $\hat{\delta}(\sigma, \sigma')$ as a proxy for $d_\infty(\sigma, \sigma')$. Let σ_t be the signed permutation obtained by composing Poisson(t) independent L-reversals. We slightly abuse notation and write $\hat{\delta}(\sigma_t)$ for $\hat{\delta}(\sigma_0, \sigma_t)$.

THEOREM 4. *Assume that $L(n) \to \infty$. Then*

$$\frac{1}{n}\hat{\delta}(\sigma_{cn/2}) \to u(c).$$

However, when $L(n)$ stays bounded, we get a behavior similar to Theorem 3.

THEOREM 5. *Assume $\lim_{n\to\infty} L(n)$ exists. Then we have convergence in probability; for all $c > 0$,*

$$\frac{1}{n}\hat{\delta}(\sigma_{cn/2}) \to w(c)$$

as $n \to \infty$, for some C^2 function $w(c)$ which satisfies $0 < w(c) < c/2$ for all $c > 0$.

The proofs for these two results are *verbatim* identical to those of Theorems 2 and 3. The choice of stating our results for transpositions is merely one of convenience, as transpositions are easier to describe and more familiar to many mathematicians.

2.3. *Numerical application to Drosiphila set.* We now illustrate on the dataset from Table 1 the possible relevance of Theorems 4 and 5. We first compute the parsimony distance in this case. Here there are $n = 79$ genes, and even though the orientation of each gene is not written, it is not difficult to find an assignment of orientations which minimizes the parsimony distance $d_\infty(\sigma)$. We find that the parsimony distance is $d_\infty(\sigma) = 54$.

First assume that all reversals are equally likely, or that L is large enough that the behavior described in Theorem 4 holds, and let us estimate the actual number of reversals that were performed. We are thus looking for t such that $d_\infty(\sigma_t) = 54$ when $n = 79$. Changing variables $t = cn/2$, we are looking for $c > 0$ such that $u(c) = 54/79 \approx 0.68$. Thus, inverting u we find $c \approx 1.6$ and hence, we may estimate the number of reversals to be around $t = 63$. Note that the discrepancy with parsimony ($d_\infty = 54$) is already significant.

This estimate keeps increasing as L decreases and the behavior of Theorem 3 starts kicking in. For instance, with $L = 4$ (so that $L/n \approx 5\%$), simulations give $c \approx 2.4$ or $t \approx 95$ reversals, or 175% of the initial parsimony estimate!

Ideally, we would want to use estimates in the biology literature on the typical range of reversals, in combination with the results of this paper, to produce a refined estimate. Kent et al. (2003) estimated the median length of a reversal in human/mouse genomes to be about 1 kb, corresponding very roughly speaking to L being a few units, say $1 \leq L \leq 4$. (However, they find a distribution for the reversal lengths which is bimodal and hence, quite different from the one we have chosen for simplicity in this paper.) Other estimates we found in several biology papers differed by several orders of magnitude, so that there does not appear to be a consensus on this question. Instead, we performed some reverse engineering, and compared our method with other existing methods. York, Durrett and Nielsen (2002) used a Bayesian approach in a model comparable to ours. The mode of the posterior distribution was at $t \approx 87$, with the parsimony estimate lying outside the 95% confidence interval [see Durrett (2002), Section 9.2.2, for further details]. This suggests that L is slightly more than 4, broadly speaking consistent with the estimate of Kent et al. (2003).

2.4. *Simulations for transpositions.* We complement the above example with plots (see Figure 1) to show how $\hat{\delta}(t) = \hat{\delta}(\sigma_t)$ evolves with t for finite n by straightforward MCMC, averaging over 1000 simulations in each case. The dotted line shows $u(c)$ and the solid line shows the average over the simulations. We observe that as L increases, $u(c)$ provides a better estimate to the parsimony.
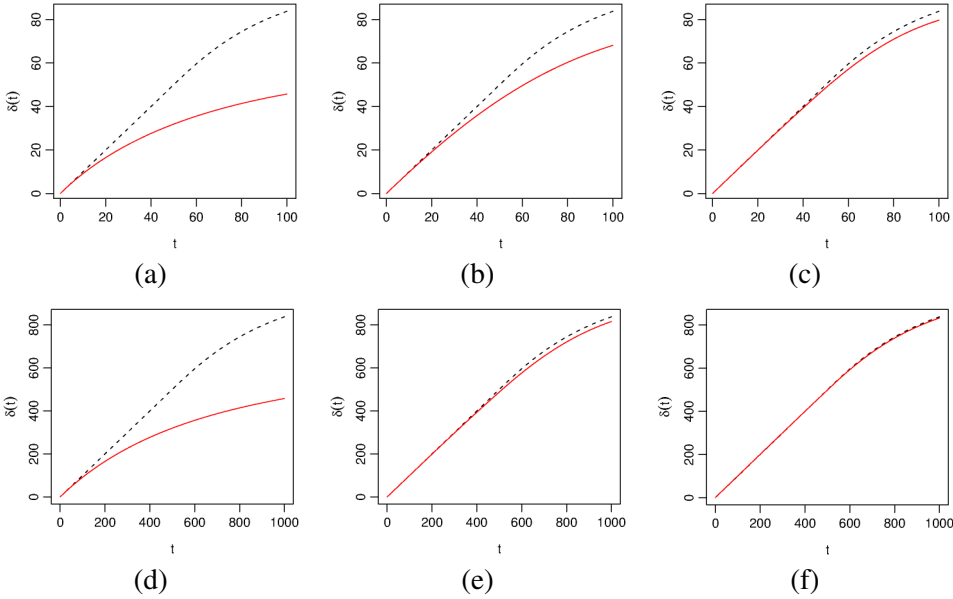
FIG. 1.   *Here* (a) $n = 100$, $L = 1$; (b) $n = 100$, $L = 5$; (c) $n = 100$, $L = 50$; (d) $n = 1000$, $L = 1$;
(e) $n = 1000$, $L = 50$; (f) $n = 1000$, $L = 500$.

3. Evolution of the components of the random graph.   We begin by proving
a few results relating to the components of a random graph which evolves in a more
general way than previously defined. For each $n \geq 2$, fix a probability distribution
$(p_\ell^n)_{1 \leq \ell \leq \lfloor n/2 \rfloor}$. We will omit the superscript n in all calculations below in order
to not overload the notation. For the rest of this section we redefine $(G(t), t \geq 0)$
to be the random graph process where at rate 1 we choose a random variable D
according to the distribution (p_ℓ), and open a uniformly chosen edge from those
of graph distance D. We define

$$(6) \qquad\qquad \varepsilon_n := \max_{1 \leq \ell \leq \lfloor n/2 \rfloor} p_\ell.$$

We begin by analyzing how the components in the random graph $G(t)$ evolve over
time.

LEMMA 1.   *Let $C(t)$ be the connected component of $G(t)$ containing some
fixed vertex $v \in V$, and let $t = cn/2$ for some $c > 0$. Assume that $\varepsilon_n \to 0$. We have
that $C(t) \preceq Z$ (where $\preceq$ stands for stochastic domination) and*

$$|C(t)| \to Z \qquad as\ n \to \infty$$

*in distribution. Here Z is the total progeny of a Galton–Watson branching process
in which each individual has a* Poisson(c) *number of offspring.*

REMARK 1.   The argument below is simpler to follow in the case where (p_ℓ) is the uniform distribution in $\{1, \ldots, L(n)\}$. We will need a more precise version of this lemma later on, in Lemma 5. It may thus be helpful for the reader to look at Lemma 5 first in order to understand the main idea of the proof of Lemma 1.

PROOF OF LEMMA 1.   We use the *breadth-first search* exploration of the component $C(t)$. That is, we expose the vertices that form $C(t)$ by looking iteratively at neighborhoods of increasing radius about v. In doing so, the vertices of $C(t)$ are naturally ordered according to levels $\ell = 1, 2, \ldots$ which represent the distance of any vertex from that level to the vertex v. To be more precise, if $A \subset V$ let $\mathcal{N}(A)$ denote the neighborhood of A, that is,

$$\mathcal{N}(A) = \{y \in V : y \sim x \text{ in } G(t) \text{ for some } x \in A\}.$$

Let $A^0 = \{v\}$ and then define inductively for $i \geq 0$,

$$A^{i+1} = \mathcal{N}(A^i) \setminus \bigcup_{j=0}^{i} A^j.$$

The statement of the lemma will follow from the observation that when $t = cn/2$, the sequence $(|A^0|, |A^1|, \ldots)$ converges in the sense of finite-dimensional distributions toward $(Z^0, Z^1, \ldots)$, the successive generation sizes of a Poisson(c) Galton–Watson process. Thus, fix an integer-valued sequence $(n_0, n_1, \ldots)$ with $n_0 = 1$. We wish to show that

$$\mathbb{P}(|A^0| = n_0, \ldots, |A^i| = n_i) \to \mathbb{P}(Z^0 = n_0, \ldots, Z^i = n_i)$$

as $n \to \infty$, which we do by induction on $i \geq 0$. The statement is trivial for $i = 0$. Now let $i \geq 0$. Given $\mathcal{A}_i = \{A^0 = n_0, \ldots, A^i = n_i\}$, we look at the neighbors in level $i + 1$ of each vertex $v_1, \ldots, v_{n_i}$ in level i, one at a time.

Let $\bar{G}(t)$ be the multigraph on V with identical connections as $G(t)$, but where each edge is counted with the multiplicity of the number of times the transposition (i, j) has occurred prior to time t. Equivalently, for each unordered pair of vertices $(i, j) \in V^2$ at distance $\|i - j\| = \ell \geq 1$, consider an independent Poisson process $N^{(i,j)}(t)$ of parameter $2p_\ell/n$. Then the multigraph $\bar{G}(t)$ contains $N^{(i,j)}(t)$ copies of the edge (i, j), while the graph $G(t)$ contains the edge (i, j) if and only if $N^{(i,j)}(t) \geq 1$.

Note that if $w \in V$, then the degree $\bar{d}_w$ of w in $\bar{G}(t)$ is

$$\bar{d}_w = \sum_{\ell \geq 1} \text{Poisson}(2tp_\ell/n) =_d \text{Poisson}(c).$$

Let $\mathcal{F}_i = \sigma(A^0, \ldots, A^i)$. Conditionally on $\mathcal{F}_i$, order the vertices from A^i in some arbitrary order, say $v_1, \ldots, v_{n_i}$. Observe that

$$A^{i+1} = \bigcup_{j=1}^{n_i} \left[\mathcal{N}(v_j) \setminus \left(\bigcup_{j=0}^{i} A^j \cup \bigcup_{k=1}^{j-1} \mathcal{N}(v_k) \right) \right].$$

It follows directly that, conditionally on $\mathcal{F}_i$,

$$|A^{i+1}| \preceq \sum_{j=1}^{n_i} P_j, \tag{7}$$

where P_j are independent Poisson(c) random variables which are further independent from $\mathcal{F}_i$. (The stochastic domination $|C_v| \preceq Z$ already follows from this observation.) For $1 \le j \le n_i$, let $\mathcal{F}_{i,j} = \mathcal{F}_i \vee \sigma(\mathcal{N}(v_1) \cup \cdots \cup \mathcal{N}(v_j))$. Observe that, conditionally given $\mathcal{F}_{i,j-1}$, then

$$N_j = \left| \mathcal{N}(v_j) \setminus \left(\bigcup_{j=0}^{i} A^j \cup \bigcup_{k=1}^{j-1} \mathcal{N}(v_k) \right) \right| \tag{8}$$

is stochastically dominated by P_j but also dominates a thinning of P_j which is a Poisson random variable with parameter $c(1 - M\varepsilon_n)$, where $M = n_0 + \cdots + n_i + |\bar{\mathcal{N}}(v_1)| + \cdots + |\bar{\mathcal{N}}(v_{j-1})|$, where $\bar{\mathcal{N}}(w)$ denotes the neighborhood of w in $\bar{G}(t)$ (hence, neighbors are counted with multiplicity). Furthermore, note that the random variables $(N_j, 1 \le j \le n_i)$ are conditionally independent given $\mathcal{F}_i$. Since $\mathbb{E}(M\varepsilon_n | \mathcal{F}_i) \to 0$ by the stochastic domination (7), it follows that

$$\mathbb{P}(|A^{i+1}| = n_{i+1} | \mathcal{F}_i)\mathbf{1}_{|A^i|=n_i} \to \mathbb{P}\left(\sum_{j=1}^{n_i} P_j = n_{i+1} \right).$$

This completes the induction step and finishes the proof of convergence in distribution. $\quad\square$

A useful consequence of this result is the following lemma.

LEMMA 2.   *Let $t = cn/2$, where $c > 0$. Then as $n \to \infty$, the number, K_t, of connected components of $G(t)$ satisfies*

$$\mathbb{E}(K_t) \sim n \sum_{k=1}^{\infty} \frac{k^{k-2}}{ck!} (ce^{-c})^k. \tag{9}$$

PROOF.   For $v \in V$, let C_v be the component containing vertex v. Then observe that the total number of components is given by $\sum_{v \in V} \frac{1}{|C_v|}$ and thus by exchangeability, the expected number of components is $n\mathbb{E}(1/|C_v|)$. Dividing by n and applying the bounded convergence theorem (since $1/|C_v| \le 1$) as well as Lemma 1, we obtain

$$\frac{1}{n}\mathbb{E}(K_t) \to \sum_{k=1}^{\infty} \frac{1}{k}\mathbb{P}(Z=k) = \sum_{k=1}^{\infty} \frac{k^{k-2}}{ck!}(ce^{-c})^k,$$

where the exact value $\mathbb{P}(Z=k)$ of the probability mass function of Z is the well-known Borel–Tanner distribution [see, e.g., Berestycki and Durrett (2006), Corollary 1]. $\quad\square$

We now prove that the number of components K_t is concentrated around its mean.

LEMMA 3. *Let $c > 0$ and let $t = cn/2$. Assume that $\varepsilon_n \to 0$. We have $K_t/\mathbb{E}(K_t) \to 1$ in probability as $n \to \infty$.*

PROOF. We write $K_t = Y_t + W_t$ where Y_t counts the components smaller than a threshold $T = (1/\varepsilon_n)^{1/4}$ and W_t those that are greater than this threshold. Note that $W_t \le n/T = o(n)$ and thus it suffices to show that Y_t is concentrated around its mean, that is, $\mathrm{Var}(Y_t) = o(n^2)$.

Note that we can always write

$$Y_t = \sum_{v \in V} \frac{1}{|C_v|} \mathbf{1}_{\{|C_v| \le T\}}$$

and thus

$$\mathrm{Var}(Y_t) = n\,\mathrm{Var}\big((1/|C_v|)\mathbf{1}_{\{|C_v| \le T\}}\big) + \sum_{v \ne w} \mathrm{Cov}\bigg(\frac{1}{|C_v|}\mathbf{1}_{\{|C_v| \le T\}}, \frac{1}{|C_w|}\mathbf{1}_{\{|C_w| \le T\}}\bigg).$$

Since $1/|C_v| \le 1$, the first term in the right-hand side is smaller than n. Define $S_v = \frac{1}{|C_v|}\mathbf{1}_{\{|C_v| \le T\}}$, $S_w = \frac{1}{|C_w|}\mathbf{1}_{\{|C_w| \le T\}}$. To know the value of S_v and S_w, it suffices to explore by breadth-first search a relatively small number of vertices in the components of v and w. While we do so, it is unlikely that the exploration of these components will ever intersect, hence, the random variables S_v and S_w are nearly independent.

To formalize this idea, let C_v^T (resp., C_w^T) denote the subset of C_v (resp., C_w) obtained by exploring at most T individuals using breadth-first search as above. Let $\tilde{S}_v$ be a copy of S_v, independent from C_w. Then conditionally on C_w^T, exploring C_v until at most T vertices have been exposed using breadth-first search, we may take $S_v = \tilde{S}_v$ except if C_v^T intersects with C_w^T, an event which we denote by $\mathcal{A}$. (To see this, imagine generating an independent copy $\tilde{C}_v^T$, using the same number of offsprings and positions for each individual in the breadth-first search exploration of C_v^T as in, but stop if at any point $\tilde{C}_v^T$ has an intersection with C_w^T.)

Thus, letting $m_v = \mathbb{E}(S_v) = \mathbb{E}(\tilde{S}_v)$, since $\tilde{S}_v$ is independent from C_w^T, and since $S_v = \tilde{S}_v$ on $\mathcal{A}^\complement$,

$$\mathbb{E}(S_v - m_v | C_w^T) = \mathbb{E}\big((S_v - \tilde{S}_v) | C_w^T\big) + \mathbb{E}\big((\tilde{S}_v - m_v) | C_w^T\big)$$
$$= \mathbb{E}\big((S_v - \tilde{S}_v)\mathbf{1}_{\mathcal{A}} | C_w^T\big) + \mathbb{E}\big((S_v - \tilde{S}_v)\mathbf{1}_{\mathcal{A}^\complement} | C_w^T\big)$$
$$= \mathbb{E}\big((S_v - \tilde{S}_v)\mathbf{1}_{\mathcal{A}} | C_w^T\big) \qquad \text{a.s.,}$$

and thus since $0 \le S_v \le 1$ and $0 \le \tilde{S}_v \le 1$,

$$\big|\mathbb{E}(S_v - m_v | C_w^T)\big| \le 2\mathbb{P}(\mathcal{A} | C_w^T) \qquad \text{a.s.,}$$

so that

$$\mathrm{Cov}(S_v, S_w) \le 4\mathbb{P}(\mathcal{A}).$$

Now observe that by Markov's inequality, $\mathbb{P}(\mathcal{A}) \le \mathbb{E}(e(C_v^T, C_w^T))$, where $e(A, B)$ denotes the number of edges between A and B. Since $|C_v^T| \le T$ and $|C_w^T| \le T$ by definition, we have $\mathbb{E}(e(C_v^T, C_w^T)) \le c\varepsilon_n T^2/2 = O(\varepsilon_n^{1/2}) \to 0$. The lemma follows. $\quad\square$

4. Proof of Theorem 1.

4.1. *Connection with branching random walk.* In this section we return to considering the random graph model $(G(t), t \ge 0)$ as given in the Introduction. The proof of (i) in Theorem 1 is easy and follows directly from the observation that for a given vertex v, $|C_v|$ is stochastically dominated by Z, the total progeny of a Poisson(c) Galton–Watson tree (see Lemma 1). When $c < 1$ it is easy to see that there exists $\lambda > 0$ and $C < \infty$ such that $\mathbb{P}(Z > k) \le Ce^{-\lambda k}$. Taking $k = b \log n$ with b sufficiently large, (i) now follows from a simple union bound.

We turn to the proof of (ii) in Theorem 1, which is the most challenging technically in this paper, and assume that $c > 1$. The key to the investigation of the properties of $G(t)$ with $t = cn/2$ is the following observation, which connects the geometry of a given component to the range of a certain branching random walk. We start by introducing notation and definitions. Let T be a Galton–Watson tree with a given offspring distribution and denote by T_i the ith level of the tree T. Let $(S(v), v \in T)$ denote a T-indexed random walk. That is, let $(X(e))_{e \in T}$ be a collection of i.i.d. random variables with a prescribed step distribution, and for all vertices $v \in T$, define $S(v) = S(o) + \sum_{e \prec v} X_v$, where the sum $e \prec v$ runs along all edges that are on the shortest path between the root o and v.

Let $t = cn/2$. Let $w \in V$, say $w = 0$, and let $C = C_w$ be the component containing w in $G(t)$. Consider the breadth-first exploration of C introduced in Lemma 1. Recall that $A^{i+1} = \mathcal{N}(A^i) \setminus \bigcup_{j=0}^{i} A^j$. Observe that it could be that two vertices $w, w' \in A^i$ each select a same neighbor z. We refer to this type of connection as a self-intersection. We view each A^i as a subset of $\mathbb{Z}$ by identifying V with

$$\{-\lfloor n/2 \rfloor + 1, \dots, -1, 0, 1, \dots, \lceil n/2 \rceil\}.$$

The following is a warm-up for the more complicated kind of couplings which will be needed later on.

LEMMA 4. *Let $c > 0$ and let $t = cn/2$. For each $k \ge 1$,*

$$\left(\sum_{v \in A^i} \delta_{v/L}, 1 \le i \le k\right) \to \left(\sum_{v \in T_i} \delta_{S_v}, 1 \le i \le k\right)$$

weakly in distribution as $n \to \infty$, where $(S_v)_{v \in T}$ denotes a branching random walk started from 0 with offspring distribution Poisson(c) *and step distribution uniform on $(-1, 1)$, and δ_x denotes the Dirac pointmass at x.*

PROOF. The proof of the lemma is an easy extension of Lemma 1, since in the case where there are no self-intersections, all the displacements of the children of vertices in any given generation form i.i.d. uniform random variables on $\mathcal{N}_L = \{-L, \ldots, -1, 1, \ldots, L\}$. Details are left to the reader. □

In practice, the finite-dimensional distribution convergence result of Lemma 4 will not be strong enough as we will typically need to explore more than a finite number of generations. The following lemma strengthens this to show that the breadth-first exploration of a cluster may be coupled *exactly* with a slightly modified branching random walk up until the first time the latter has a self-intersection. More precisely, let T be a Galton–Watson tree with offspring distribution Poisson(c), and let $S_v, v \in T$, be defined as above except that if $v \in T$ with offspring $v_1, \ldots, v_k$ [let e_i denote the edge (v, v_i)], we define the displacement variables $X(e_1), \ldots, X(e_k)$ to be sampled *with replacement* uniformly from $\mathcal{N}_L$. The sampling is still done independently for different vertices $v \in T$. We call this process branching random walk with replacement for future reference. We also introduce a version *with erasure*, where if v and w are such that $S_v = S_w$ (what we call a self-intersection) with v discovered before w in the breadth-first search, then the entire descendance of w is ignored or killed. We call this process an erased branching random walk and denote it by $(\tilde{S}_v, v \in T)$.

LEMMA 5. *Let $(S_v, v \in T)$ denote a branching random walk as above and $(\tilde{S}_v, v \in T)$ its corresponding erasure. Then there exists a coupling of $(\tilde{S}_v)_{v \in T}$ and $(A^i, i \geq 0)$ such that the sets A^i and $\{\tilde{S}_v, v \in T_i\}$ coincide exactly for each $i \geq 0$. In particular, let τ be the first self-intersection level; $\tau = \inf\{n \geq 1 : \exists v \neq w \in V(T_n), S_v = S_w\}$. Then we can couple A^i and $(S_v, v \in T_i)$ for each $i < \tau$.*

PROOF. For the most part this is a variation on Lemma 1, but there are some subtleties. Assume we are exploring the connections of a vertex $v \in \mathcal{A}^i$ for some $i \geq 0$. Let A be the $2L - 1$ potential neighbors of v, and let $B \subset A$ be the set of those within A which have already been exposed so far. For each of the $|A \setminus B|$ potential new neighbors of v to be added to $\mathcal{A}^{i+1}$, the edge joining it to v has appeared a Poisson$(t/(nL))$ number of times. Of course, if an edge appears several times, this amounts to connecting to the same vertex, and this is why we choose sampling *with* replacement. The action of sampling uniformly with replacement from A or from $A \setminus B$ can be chosen to be identical, until the first time that sampling from A uses an element from B. The rest of the details are left to the reader.
□

REMARK 2. Note that by the classical birthday problem, τ is unlikely to occur before at least of order $\sqrt{L}$ vertices have been added. Thus we can couple exactly the breadth-first search exploration of C_v and a branching random walk until of order $\sqrt{L}$ vertices have been discovered.

In fact, this will still not be strong enough and we will need to push this exploration until of order $o(L)$ vertices have been discovered. Of course, self-intersections can then not be ignored, but there are not enough of them that they cause a serious problem, so the breadth-first search exploration coincides with "most" of the branching random walk.

4.2. *Survival of killed branching random walk.* The basic idea for the proof of (ii) in Theorem 1 is a renormalization (sometimes also called "block") argument.

We show that if the component of a given vertex is larger than some fixed number, then this component is likely to reach distance KL, where $K > 0$ is a large number (which may even depend on L) to be suitably chosen. This may be iterated to show that two points selected at random from V will be connected with probability approximately $\theta(c)^2$, where $\theta(c)$ is the survival probability of T. For now, we will need a few basic estimates about killed branching random walks. In many ways, some of the results are more natural to state when we let $L \to \infty$ rather than $n \to \infty$. Since $L(n) \to \infty$, the two statements are identical.

Consider a branching random walk as above, started at $v \in V$, with step distribution uniform in $\{-L, \ldots, -1, 1, \ldots, L\}$ and some arbitrary offspring distribution with probability generating function $\phi(s)$. By killed branching random walk (KBRW) we refer to a branching random walk where, in addition, particles die if they escape a given interval containing the starting point.

LEMMA 6. *Let θ denote the survival probability of the branching random walk, that is, $\rho = 1 - \theta$ is the smallest root of $z = \phi(z)$. For each $\varepsilon > 0$ we can choose $K = K(\varepsilon, \phi)$ such that if all particles are killed upon escaping $[v - KL, v + KL]$, then for all L sufficiently large (depending solely on ε and ϕ) the survival probability θ^K of KBRW satisfies $\theta^K \geq \theta(1 - \varepsilon)$.*

PROOF. Let T denote the Galton–Watson tree describing the descendants of v. Conditionally on survival of T, the subset U of T for which all vertices in U have infinite progeny (i.e., the set of infinite rays) forms a Galton–Watson process with modified progeny; the generating function satisfies

$$(10) \qquad \tilde{\phi}(s) = \frac{1}{\theta}[\phi(\theta s + 1 - \theta) - 1 + \theta].$$

Define $\mathcal{N}_v := [v - L, v + L]$. Consider a subset W of U obtained as follows. Let $\xi = (u_0, u_1, \ldots, u_R)$ be a fixed ray in U where R is the first time the ray leaves $[v - KL, v + KL]$. Thus $(S_{u_0}, S_{u_1}, \ldots, S_{u_R})$ is a random walk with the underlying step distribution, killed upon exiting $[v - KL, v + KL]$. Then W restricted to ξ will consist of the subsequence of (u_{n_i}) such that $S_{u_{n_i}} \in \mathcal{N}_v$. More precisely, we take W to be the union of all such subsequences over all rays ξ in U. The vertices of W have a natural tree structure, and we claim that W dominates a branching process where the offspring progeny is

$$(11) \qquad \phi_K(s) = \tilde{\phi}(\varepsilon_K + (1 - \varepsilon_K)s),$$

where $\varepsilon_K = 1/(K+2)$. The reason for this is as follows. Suppose $u \in W$, so that $S_u \in \mathcal{N}_v$. Then u has (in U) a random number, say N, of offsprings, where the generating function is given by $\tilde{\phi}$ in (10). Since the trajectory of a random walk with jumps uniform in $\{-L, \ldots, -1, 1, \ldots, L\}$ forms a martingale and the jumps are bounded by L, a classical application of the optional stopping theorem shows that any particular fixed ray emanating from each offspring of u returns to $\mathcal{N}_v$ before hitting $v \pm KL$ with probability at least $1 - \varepsilon_K$. Formula (11) follows easily. Now, survival probability is therefore at least as large as the survival probability of the Galton–Watson process with offspring distribution given by ϕ_K. Let ρ_K be the extinction probability. Then $\rho_K = \phi_K(\rho_K)$ and ρ_K is the unique root of this equation in $(0, 1)$, and moreover, ρ_K is decreasing as a function of K. Since $\rho_K \geq 0$, call $\rho = \lim_{K \to \infty} \rho_K$. It is trivial to conclude by continuity of $\tilde{\phi}$ that $\rho = \tilde{\phi}(\rho)$ and that $\rho < 1$, from which it follows that ρ is the extinction probability of $\tilde{\phi}$ and is thus equal to 0. Thus we may choose K sufficiently large that $\rho_K < \varepsilon$. □

We now consider a certain subprocess of the killed branching random walk and show that this also survives, growing exponentially and leaving many offsprings very near the starting point v. Rather than stating a general result we will state only what we need. Fix a function $\omega(L)$ such that $\omega(L) \to \infty$ sufficiently slowly, say $\omega(L) = \log L$, and let $f_0(L)$ be any function such that $f_0(L) \leq L/\omega(L)$. Fix an integer $d \geq 1$, and explore no more than d offsprings for any individual, that is, declare dead any additional offspring. Fix $\lambda = \lambda_{K,d}$ and also declare a vertex v dead if the most recent common ancestor u of v such that $S_u \in \mathcal{N}_v$ is more than λ generations away. Refer to this process as $KBRW_1$. Note that $KBRW_1$ is a subprocess of $KBRW$ and thus of BRW. Note also that the erased $KBRW_1$ is a subprocess of the erased $KBRW$.

LEMMA 7. *Assume that $\phi''(1) < \infty$ so that the offspring distribution has finite second moments. For all $\varepsilon > 0$, there exists $K = K(\varepsilon, \phi)$, $d \geq 1$ and λ such that if all particles are also killed upon escaping $[v - KL, v + KL]$, then for all sufficiently large L (depending solely on ϕ and ε), with probability at least $(1 - \varepsilon)\theta$, the following hold*:

(i) *$KBRW_1$ gets at least $f_0(L)$ descendants in at most $c \log f_0(L)$ generations for some $c > 0$,*
(ii) *$f_0(L)/K$ of them are in $\mathcal{N}_v$.*

PROOF. Consider the *KBRW* of Lemma 6 and let W be as in the proof of that lemma. Consider $W \cap KBRW_1$ and note that this is a Galton–Watson process with offspring distribution which dominates one with a generating function given by (11), where now $1 - \varepsilon_K$ is the probability that a random walk with step distribution uniform on $\{-L, \ldots, -1, 1, \ldots, L\}$ returns to $\mathcal{N}_v$ before exiting

$[v - KL, v + KL]$, and that this takes less than λ steps. By choosing K sufficiently large, d sufficiently large and λ sufficiently large (in that order), ε_K is arbitrarily small and thus we find that $KBRW_1$ survives forever with probability at least $(1 - \varepsilon)\theta$, as in Lemma 6. Note also that $W \cap KBRW_1$, being a Galton–Watson tree and having finite second moments, grows exponentially fast by the Kesten–Stigum theorem. Thus fewer than $c \log f_0(L)$ levels are needed to grow $W \cap KBRW_1$ to size $f_0(L)$ for some $c > 0$, and so at this level we will certainly have at least $f_0(L)$ explored in $KBRW_1$.

Let $\mathcal{T}$ be the $KBRW_1$ stopped when the population size exceeds $f_0(L)$. Define the following marking procedure in $KBRW_1$. Mark any node $u \in KBRW_1$ if the position of the branching random walk S_u at this node is in the interval $\mathcal{N}_v$. Let $\mathcal{M} \subset \mathcal{T}$ be the set of marked nodes. Since by construction, every node $u \in \mathcal{T}$ has an ancestor at (genealogical) distance at most λ which is a marked node, and since the degree of any node in $\mathcal{T}$ is at most $d + 1$, it follows that

$$(12) \qquad |\mathcal{M}| \geq \eta|\mathcal{T}| \qquad \text{where } \eta = \frac{1}{1 + d + d^2 + \cdots + d^\lambda}.$$

[To see (12), just notice that for every new mark, one can add at most $1/\eta$ nodes in the tree without adding a new mark, and proceed by induction.] For (ii) to occur, it suffices that $|\mathcal{M}| \geq f_0(L)/K$. Since by construction $|\mathcal{T}| \geq f_0(L)$, choosing $\lambda = \lambda_{K,d} \geq \lfloor (\log K)/(\log d) \rfloor - 1$ shows that (ii) occurs as soon as (i) holds. The proof of the lemma is complete. $\quad\square$

We now strengthen this last result by showing that the erased random walk also has a large number of offsprings in $[v - KL, v + KL]$. Further, we suppose also that there is a set F of locations which, if an individual lands on, results in that individual being removed. We call these *forbidden locations*.

LEMMA 8. *Consider an erased branching random walk, started at $v \in V$, with step distribution uniform in $\{-L, \ldots, -1, 1, \ldots, L\}$ and some arbitrary offspring distribution with probability generating function $\phi(s)$ with $\phi''(1) < \infty$. Suppose also that there is a set F of forbidden locations, with $|F| \leq L/\omega(L)$ and $\omega(L) \to \infty$. Let θ denote the survival probability of the branching random walk. For all $\varepsilon > 0$ we can choose $K = K(\varepsilon, \phi)$ such that if all particles are also killed upon escaping $[v - KL, v + KL]$, then for all sufficiently large L (depending solely on ϕ and ε), with probability at least $(1 - \varepsilon)\theta$, the following hold*:

(i) *the erased $KBRW_1$ gets at least $f_0(L)$ descendants in at most $c \log f_0(L)$ generations for some $c > 0$,*

(ii) *$f_0(L)/K$ of them are in $\mathcal{N}_v$.*

PROOF. Let τ be the first time that the killed branching random walk has more than $2f_0(L)$ descendants. Let us show that the associated erased branching

random walk has at least $f_0(L)$ individuals at that point with high probability. To see this, we first observe that by (i) in Lemma 7 the number of generations, τ, is at most $c \log f_0(L)$ for some $c > 0$. Before time τ, for each new vertex added to the branching random walk, the probability that it is a self-intersection or hits an element of F is no more than $(2 f_0(L) + |F|)/(2L - 1)$. Thus the probability that a particular ray of no more than $c \log f_0(L)$ generations contains a self-intersection is, by Markov's inequality, at most

$$\frac{c(\log f_0(L))(2 f_0(L) + |F|)}{2L - 1} \to 0$$

as $L \to \infty$. Therefore, the number of vertices that are present in the $KBRW_1$ but not in the erased $KBRW_1$ is, by Markov's inequality again, at most $(1/2) \times (2 f_0(L)) = f_0(L)$ with high probability. We shall denote by $EKBRW_1$ the erased $KBRW_1$ which has a set F of forbidden locations.

By Lemma 7, we also know that $2 f_0(L)/K$ individuals of the $KBRW_1$ population are located in $\mathcal{N}_v$. Since we have just shown that the total number of individuals not in $EKBRW_1$ is $o(f_0(L))$ with high probability, we deduce that at least $f_0(L)/K$ individuals of $EKBRW_1$ are located in $\mathcal{N}_v$. The proof of the lemma is complete. $\quad\square$

4.3. *Breadth-first search explorations.* The next three lemmas give us some information on the breadth-first search exploration of a component C_v of a given vertex $v \in V$ in the random graph $G(t)$. For reasons that will soon become clear, we wish to assume that by the point we start exploring the component C_v, part of the graph has already been explored (a vertex has been explored once all its neighbors have been observed). The part that has already been explored (denoted F) represents forbidden vertices, in the sense that since we have already searched F, the breadth-first search of C_v can no longer connect to it.

We now specialize to the case where ϕ is the generating function of a Poisson(c) distribution, and in all that follows we let $\theta = \theta(c)$ be the survival probability of a Poisson(c) Galton–Watson tree. It turns out that we need to separately treat the case where L is very close to n, and this will be done later in Lemma 15.

LEMMA 9. *Fix $c > 1$, $\varepsilon > 0$, $v \in V$ and fix $K = K(\varepsilon, c)$ as in Lemma 8. We assume that a set F containing at most $L/\omega(L)$ vertices have already been discovered in $[v - KL, v + KL]$ (and v is not one of them). Then for all n large enough (depending only on ε and c), if $L < n/(2K)$, then with probability at least $\theta(1 - \varepsilon)$, a search procedure of C_v can uncover at least $f_0(L)/K$ vertices of C_v in $\mathcal{N}_v$ without exploring more than $f_0(L)$ vertices in total in $[v - KL, v + KL]$, and none outside.*

PROOF. Consider the breadth-first search exploration of C_v, with the following modifications. We stop exploring the descendants of any vertex outside of

$[v - KL, v + KL]$. We also completely stop the exploration when more than $f_0(L)$ vertices have been discovered. Also, we stop exploring the descendent of any vertex in F and we fix $d \geq 1$ and truncate the offspring progeny at d, so that if an individual has more than d offspring, only the first d encountered are explored. This keeps the degree of any node in the genealogical tree bounded by $d + 1$. We choose $d = d_K$ as in Lemma 7. We also stop exploring the descendants of an individual if the time elapsed since the last time an ancestor of this individual visited $\mathcal{N}_v$ exceeds $\lambda = \lambda_{K,d}$, where $\lambda_{K,d}$ is as in Lemma 7. We refer to this process as $KBFS_1$. More formally, we use the following algorithm:

Step 1. Set $\Omega_E = \varnothing$, $\Omega_A = \{v\}$. These correspond to the explored and active vertices, respectively.

Step 2. If $|\Omega_E| \geq f_0(L)$ we stop. Otherwise we proceed to Step 3.

Step 3. Set $\Omega_N = \varnothing$. For each $w \in \Omega_A$, add its neighbors (excluding the parent of w) to Ω_N until d have been added, or there are no more.

Step 4. Add the vertices in Ω_A to Ω_E.

Step 5. Set $\Omega_A = \Omega_N \setminus \{\Omega_E \cup F\}$. If $\Omega_A = \varnothing$, then we stop.

Step 6. Remove from Ω_A all vertices outside of $[v - KL, v + KL]$ and those that do not have an ancestor in $\mathcal{N}_v$ fewer than λ generations away.

Step 7. Go to Step 2.

This exploration can be exactly coupled with the $EKBRW_1$ considered up to the first time τ that the total population size exceeds $f_0(L)$, by taking in Lemma 8 the set F as it is defined here. Lemma 9 thus follows directly from Lemma 8. $\square$

To establish the existence of a connection between two vertices v and w, it will be useful to add another twist to the breadth-first search exploration of C_v and C_w, by *reserving* some of the vertices we discover along the way. That is, we decide not to reveal their neighbors until a later stage, if necessary. This allows us to keep a reserve of "fresh" vertices to explore at different locations and that we know are already part of C_v or C_w. To be more precise, let $\varepsilon > 0$. Let $1 < c' < c$ be such that $\theta(c') \geq \theta(c)(1 - \varepsilon/2)$. Let ν be small enough that $c(1 - \nu) > c'$. When exploring C_v through a method derived from breadth-first search, we choose which vertices to reserve as follows: for each new vertex that we explore, if it has any offsprings, we choose one uniformly at random, and reserve it with probability ν independently of anything else. (See below for a rigorous formulation.) Note, in particular, that the set of vertices that get reserved is dominated by a Poisson thinning of the original exploration procedure, with thinning probability ν. Let $K = K(\varepsilon/2, c')$ be as in Lemma 8. Note that with this choice of ν and K, the survival probability θ' of $EKBRW_1$ is at least

$$(13) \qquad\qquad \theta'(c) \geq \theta(c')(1 - \varepsilon/2) \geq \theta(c)(1 - \varepsilon)$$

for all L sufficiently large (depending solely on c and ε).

Thus, starting from a vertex v, a branching random walk killed when escaping $[v - KL, v + KL]$ with this reservation procedure survives forever with probability at least $\theta(c)(1 - \varepsilon)$. From this we deduce without too much trouble the following result.

LEMMA 10. *Fix $c > 1$, $\varepsilon > 0$, $v \in V$. Let $\nu = \nu(\varepsilon)$ as above, and assume that a set F containing no more than $L/\omega(L)$ vertices have been discovered. Then for all sufficiently large n (depending solely on ε and c), if $L < n/(2K)$, the following hold with probability at least $(1 - 2\varepsilon)\theta$:*

   (i) *A search procedure can uncover at least $f_1(L) = f_0(L)/K$ vertices in $\mathcal{N}_v$ without uncovering more than $2f_0(L)$ vertices in total in $[v - KL, v + KL]$, and none outside.*
   (ii) *At least $\delta f_0(L)$ vertices are reserved in $\mathcal{N}_v$, for $\delta = (1 - e^{-c})\nu(\varepsilon)/(4K)$.*

PROOF. We apply the above reservation method to $KBFS_1$ (see the proof of Lemma 9). Formally, we introduce a set Ω_R of reserved vertices (initially $\Omega_R = \varnothing$). We use the same algorithm as for the modified breadth-first search but now Step 7 becomes:

Step $7'$. Partition Ω_A into classes of vertices with the same parent in the exploration. Choose uniformly from each class a representative and with probability ν this representative is added to Ω_R and removed from Ω_A. Go to Step 2.

We call this new search procedure $KBFS_2$. Let τ be the time we have discovered $f_0(L)$ nonreserved vertices. At this time the total number of explored vertices is less than $2f_0(L)$ and thus, similar to the proof of Lemma 9, we can couple the exploration with an erased $KBRW_1$ where the offspring distribution has a slightly modified offspring distribution (a randomly chosen offspring is removed with probability ν). We call this an erased $KBRW_2$. Reasoning as in Lemma 9, and using (13), we see that (i) holds with probability at least $\theta(1 - \varepsilon)$, provided that n is large enough and $L < n/(2K)$. For each new vertex exposed by $KBFS_2$ in $\mathcal{N}_v$, it has a reserved offspring in $\mathcal{N}_v$ with probability at least $(1 - e^{-c})\nu/2$, as if $u \in \mathcal{N}_v$ and X are uniformly distributed on $\{-L, \ldots, -1, 1, \ldots, L\}$, then $u + X \in \mathcal{N}_v$ with probability at least $1/2$. Thus (ii) follows from (i) and from Chebyshev's inequality. □

With this lemma we are now able to show that a vertex v connects to $v \pm KL$ with probability essentially θ, and that many vertices in the same component may be found without revealing too much inside $[v - KL, v + KL]$.

LEMMA 11. *Fix $c > 1$, $\varepsilon > 0$, and let K be as in Lemma 9. Let $0 < \zeta < \zeta' < 1/2$ and let $v \in V$. Assume that a set F of no more than $L/\omega(L)$ vertices have already been explored in $[v - KL, v + KL]$ and v is not one of them. Let $\mathcal{B}_{K,v}$*

denote the event that v is connected to at least L^ζ unexplored vertices in the range $[v + KL, v + (K + 1)L]$ which may be discovered by searching no more than $L^{\zeta'}$ vertices. Then for all sufficiently large n (depending solely on ε, c, ζ and ζ'), if $L < n/(2(K_0(\varepsilon) + 1))$, then

$$\mathbb{P}(\mathcal{B}_{K,v}) \geq \theta(c)(1 - \varepsilon).$$

PROOF. Consider the $KBFS_2$ exploration of C_v, stopped when a total of $f_0(L) = L^{\zeta'}/2$ vertices of C_v have been exposed (additional to those exposed initially). By Lemma 10, with probability at least $\theta(c)(1 - \varepsilon)$ if n is large enough and $L < n/(2K)$, this search reveals at least $k = \delta f_0(L)$ reserved vertices within $\mathcal{N}_v$, and no more than $L^{\zeta'}$ vertices in the range $[v - KL, v + KL]$ have been explored (let $\mathcal{A}_v$ denote this event). On $\mathcal{A}_v$, label $v_1, \ldots, v_k$ the first k such vertices to have been discovered in $\mathcal{N}_v$. After this stage, we then continue the $KBFS_2$ exploration started from $\{v_1, \ldots, v_k\}$ only, until a total of $L^{\zeta'}/2$ further vertices are exposed. Note that the exploration can be coupled with a system of k erased $KBRW_2$, started from $v_1, \ldots, v_k$. The total number of vertices searched by the end of the second stage will be no more than $f_0(L) + L^{\zeta'}/2 \leq L^{1/2}$. Thus, as in Lemma 8, the probability each particular vertex gives rise to a self-intersection is no more than $(|F| + \sqrt{L})/(2L - 1) \to 0$ as $n \to \infty$.

Moreover, using domination by a branching process (Lemma 1), it is easy to see that the number of generations for the $L^{\zeta'}/2$ vertices to be discovered by the branching random walk is at least $b \log L$ for some $b > 0$, with probability tending to 1 as $n \to \infty$. Now, for every $1 \leq i \leq k$, the probability that the erased branching random walk (erased $KBRW_2$) started from v_i has a descendant that hits $[v + KL, v + (K + 1)L]$ in fewer than $b \log L$ steps is at least $\theta(c)(1 - o(1)) \geq \theta(c)/2$ for n large enough (depending solely on ε and c). Thus we deduce that, on the event $\mathcal{A}_v$, the number of particles that hit $[v + KL, v + (K + 1)L]$ stochastically dominates a binomial random variable $\text{Bin}(\delta f_0(L), \theta(c)/2)$. By applying Chebyshev's inequality, this is with high probability greater than $\delta f_0(L)\theta(c)/4 \geq L^\zeta$ for all n sufficiently large (depending on c, ε, ζ, ζ'). When this occurs, $\mathcal{B}_{K,v}$ holds, so the result follows. □

Let $c > 1$, $\varepsilon > 0$ and fix $K = K_0(c, \varepsilon)$ as in Lemma 11. We now prove that if the connected component of a given vertex is not finite then it must spread more or less uniformly over V. As desired, this is achieved by keeping a density of explored sites small, lower than $1/\omega(L)$. Let $v \in V$ and split the vertex set V into $r + 1 = \lceil n/(KL) \rceil$ disjoint strips $(I_0, \ldots, I_r)$ of size KL, except for the last one which may be of size smaller than KL. Let J_i denote the initial segment of I_i of length L. Since $L(n) \geq (\log n)^{2+\xi}$ for some positive ξ by assumption, we may find $\zeta < 1/2$ such that $L(n) > (\frac{4}{\theta} \log n)^{1/\zeta}$. Let $\zeta' = (\zeta + 1/2)/2$, hence $\zeta < \zeta' < 1/2$.

LEMMA 12. *With the above notation, assume that no more than $L^{2\zeta'}$ vertices have already been exposed in each strip $I_0, \ldots, I_r$. Let $\mathcal{C}_{K,v}$ denote the event that*

v is connected to at least $k = L^\zeta$ vertices in strips $I_3, \ldots, I_{r-1}$, which may be discovered without exposing more than an additional $L^{2\zeta'}$ vertices in each strip, and that in each J_i, $3 \leq i \leq r - 1$, at least $k/2$ vertices connected to v are unexplored at the end of the search procedure. Then

$$\mathbb{P}(\mathcal{C}_{K,v}) \geq \theta(c)(1 - \varepsilon)$$

for all n large enough (depending solely on ε and c), and provided $L < n/(2K)$.

PROOF.   This is basically proved by iterating Lemma 11. We can assume that v is in strip I_1. In the first step we explore C_v using $KBFS_2$ killing at the boundary of $I_0 \cup I_1 \cup I_2$. The arguments of Lemma 11 still carry through to obtain that $\mathcal{B}^0_{K,v}$ holds with probability at least $\theta(c)(1 - \varepsilon)$ where $\mathcal{B}^0_{K,v}$ is the event that v is connected to at least k unexplored vertices in J_3, and fewer than $L^{\zeta'}$ vertices are explored in finding them.

Then define inductively for $1 \leq i \leq r - 4$, $\mathcal{B}^i_{K,v} = \mathcal{B}^{i-1}_{K,v} \cap \mathcal{C}^i_{K,v}$, where $\mathcal{C}^i_{K,v}$ is defined as follows. On $\mathcal{B}^{i-1}_{K,v}$, let $v_1, \ldots, v_k$ be a list of k vertices in the range J_{i+2} that are the first to be discovered in this search procedure in this range. Then $\mathcal{C}^i_{K,v}$ is the event that we can find at least k connections between $v_1, \ldots, v_k$ and J_{i+3} without exploring more than an additional $L^{2\zeta'}$ new vertices in $I_{i+2} \cup I_{i+3}$.

This is where it starts to pay off to allow for the exploration to unfold in a partially revealed environment in Lemma 11. Indeed, let $i \geq 1$ and condition on $\mathcal{B}^{i-1}_{K,v}$. We reserve (i.e., do not explore further) $v_{k/2+1}, \ldots, v_k$. We explore successively the components $C_{v_1}, \ldots, C_{v_{k/2}}$, each time performing $KBFS_2$ of Lemma 11. Since we never reveal more than $L^{\zeta'}$ vertices at each of those $k/2$ steps, and since we did not reveal more than $(k/2)L^{\zeta'} = L^{\zeta+\zeta'}/2 < L^{2\zeta'}/2$ other vertices in I_{i+2} previously (since $\mathcal{C}^{i-1}_v$ holds), we see that the search may be coupled with high probability (depending solely on c and ε) to $(k/2)$ erased $KBRW_2$ started at $v_1, \ldots, v_{k/2}$. Thus the total number of connections between $v_1, \ldots, v_{k/2}$, to J_{i+3} is dominated from below by $k \operatorname{Bin}(k/2, \theta(c)/2)$. Indeed, for each of $(k/2)$ trials there is a probability $\theta(c)(1 - \varepsilon) \geq \theta(c)/2$ of success (by Lemma 11), in which case k connections are added. Thus, using standard Chernoff bounds on binomial random variables,

$$\begin{aligned}
\mathbb{P}((\mathcal{B}^i_{K,v})^{\complement} | \mathcal{B}^{i-1}_{K,v}) &= \mathbb{P}((\mathcal{C}^i_{K,v})^{\complement} | \mathcal{B}^{i-1}_{K,v}) \\
&\leq \mathbb{P}(k \cdot \operatorname{Binomial}(k/2, \theta(c)/2) < k) \\
&\leq \exp(-\theta(c)k/4).
\end{aligned}$$

It follows by easy induction that for all n large enough, letting $\mathcal{C}'_{K,v} = \bigcap_{i=0}^r \mathcal{B}^i_{K,v}$,

$$\mathbb{P}(\mathcal{C}'_{K,v}) \geq (1 - e^{-\theta k/4})^r \mathbb{P}(\mathcal{B}^0_{K,v}).$$

Since $L > ((4/\theta) \log n)^{1/\zeta}$, $r = \lceil n/(2LK) \rceil$ and $k = L^\zeta$, it follows that $(1 - e^{-\theta k/4})^r \sim \exp(-r e^{-\theta k/4}) \geq \exp(-r/n) \to 1$. Vertices can only be discovered

during two consecutive steps of the proof, and hence, the total number of vertices discovered in each strip is no more than $L^{2\zeta'}$. Thus $\mathcal{C}_{K,v} \supset \mathcal{C}'_{K,v}$. The proof of the lemma is complete. $\quad\square$

LEMMA 13.   *Let* $\mathcal{D}_v = \{|C_v| > \log L\}$. *Then for any fixed* $v \in V$,

$$\mathbb{P}(\mathcal{D}_v) \to \theta(c)$$

and for v, w *fixed in* V,

$$\mathbb{P}(\mathcal{D}_v \cap \mathcal{D}_w) \to \theta(c)^2.$$

PROOF.   This is a direct consequence of Lemma 5 and the remark following it.
$\quad\square$

LEMMA 14.   *Fix* $c > 1$, $\varepsilon > 0$. *Let* v, w *be chosen uniformly at random in* V, *and let* $\mathcal{E} = \{C_v = C_w\}$ *be the event that they are connected. If* $L < n/(5K)$, *and* n *is large enough* (*depending solely on* ε *and* c) *then*

(14) $$\mathbb{P}(\mathcal{E}^{\complement}|\mathcal{D}_v \cap \mathcal{D}_w) \leq \varepsilon.$$

PROOF.   We fix $K(c, \varepsilon)$ as in Lemma 12. We apply this lemma a first time by exploring C_v as specified in this lemma with a set of forbidden vertices (vertices previously explored) being empty. We then let F be the set of all vertices explored during that procedure.

We apply one more time Lemma 12 by exploring C_w using a set of forbidden vertices given by F (which must necessarily satisfy the assumptions of Lemma 12, since the search of C_v did not reveal more than $L^{2\zeta'}$ vertices in each strip). Note that conditionally given $\mathcal{D}_v \cap \mathcal{D}_w$, both $\mathcal{C}_{K,v}$ and $\mathcal{C}_{K,w}$ must hold with high probability (depending solely on c). Let us show that $\mathcal{E}$ must then hold with high probability.

Since $\mathcal{C}_{K,v}$ and $\mathcal{C}_{K,w}$ hold, we know that each interval J_i, $3 \leq i \leq r - 1$, contains at least $L^{\zeta}/2$ unexplored vertices from both C_v and C_w. We now apply Lemma 10 repeatedly, starting from each of these unexplored vertices. Since $L < n/5K$, we have that $r \geq 4$. While fewer than $\delta f_0(L)$ vertices have been reserved in J_i, we know that fewer than $f_0(L)$ vertices have in total been explored and thus Lemma 10 can still be applied. We deduce that (conditionally given $\mathcal{D}_v \cap \mathcal{D}_w$) in each J_i, $3 \leq i \leq r - 1$, with probability greater than $1 - o(1)$ depending solely on ε and c, there are $\delta L/\omega(L)$ reserved vertices from C_v and $\delta L/\omega(L)$ reserved vertices from C_w, with δ as in Lemma 10. Thus at least one J_i (say J_1) contains $\delta L/\omega(L)$ unexplored vertices from both C_v and C_w. The probability to not observe a connection between these $(\delta L/\omega(L))^2$ pairs of vertices inside J_1 is at most

(by revealing only the status of the edges connecting each such pair)

$$\left(1 - \frac{c}{2L}\right)^{(\delta L/\omega(L))^2} \leq \exp\left(-\frac{c}{2L}\frac{\delta L^2}{\omega(L)^2}\right)$$

$$= \exp\left(-\frac{c\delta L}{2\omega(L)^2}\right) \to 0$$

for all n sufficiently large. Thus $C_v = C_w$ with high probability (depending solely on ε and c) given $\mathcal{D}_v \cap \mathcal{D}_w$, and hence, $\mathcal{E}$ holds with high probability (depending solely on ε and c) given $\mathcal{D}_v \cap \mathcal{D}_w$. $\square$

We deal with the case $L \geq n/(5K)$ separately.

LEMMA 15. *Fix $c > 1$, $\varepsilon > 0$. Let v, w be chosen uniformly at random in V, and let $\mathcal{E} = \{C_v = C_w\}$ be the event that they are connected. If $L \geq n/(5K)$, and n is large enough (depending solely on ε and c) then*

(15) $$\mathbb{P}(\mathcal{E}^{\complement}; \mathcal{D}_v \cap \mathcal{D}_w) \leq \varepsilon.$$

PROOF.   We let $\bar{\mathcal{E}}$ denote the event that we can explore C_v until at most $L^{0.7}$ vertices have been exposed finding at least $L^{0.6}$ reserved vertices and also can explore C_w until at most $L^{0.45}$ vertices have been exposed finding at least $k = L^{0.44}$ reserved vertices. By a simple modification of Lemma 10, the probability of this event given $\mathcal{D}_v \cap \mathcal{D}_w$ is at least $1 - o(1)$.

Let us show that, given the above event $\bar{\mathcal{E}}$, C_v and C_w intersect with high probability. We partition $\{1, \ldots, n\}$ into $s = 5K + 1$ disjoint intervals of size less than or equal to L. On the above event, by the pigeonhole principle there must be at least one region of size at most L, denoted I, with more than $L^{0.6}/s$ reserved vertices from C_v. We denote by $w_1, \ldots, w_k$ the k reserved vertices from C_w. For each $1 \leq i \leq k$, we continue to explore C_{w_i} by breadth-first search for $6(s+1)^2$ generations, or until a descendent is observed in interval I. Since s depends only on ε and c, with probability at least $1 - o(1)$ depending on ε and c, no self-intersections occur throughout this evolution. We claim that the probability the evolution of C_{w_i} results in us finding a descendent in I is at least $\theta(c)/3$ for n large enough depending solely on ε and c. Indeed this occurs if we can find a ray emanating from w_i where the corresponding random walk goes around the circle in less than $6(s+1)^2$ levels. We let $(X_j)_{j \geq 1}$ denote the location of a random walk on $\mathbb{Z}$ which starts at 0 and where the jump distribution is uniform on $\{-L, \ldots, -1, 1, \ldots, -L\}$. It is clear that $(X_j)_{j \geq 0}$ is a martingale, as is

$$\left(X_j^2 - jL(L+1)(2L+1)/(6L)\right)_{j \geq 0}.$$

Letting T denote the time the walk goes above sL, or below $-sL$, we see that by optional stopping $\mathbb{E}(T) < 3(s+1)^2$. Thus, by Markov's inequality, $\mathbb{P}(T >$

$6(s + 1)^2 < 1/2$. Hence, a random walk on V with jumps uniformly distributed on $\{-L, \ldots, -1, 1, \ldots, -L\}$ goes around the circle in less than $6(s + 1)^2$ steps with probability at least $1/2$. It follows that the desired ray exists with probability at least $(1 - o(1))\theta(c)/2$, depending solely on ε and c. Thus, given $\bar{\mathcal{E}}$, we can find $L^{0.43}$ reserved vertices from C_w and $L^{0.6}$ reserved vertices from C_v in the interval I, with probability $1 - o(1)$ depending solely on ε and c. Looking one level further, the number of connections between C_v and C_w in this region is $\text{Bin}(L^{1.03}, c/(2L))$ which is larger than 1 with probability $1 - o(1)$ depending solely on ε and c. Equation (15) now follows. $\quad\square$

We are now ready to finish the proof of (ii) in Theorem 1.

PROOF OF THEOREM 1(ii). Let v, w be chosen uniformly on V. Let $\mathcal{E} = \{C_v = C_w\}$ be the event that they are connected.

Consider $W = \{v \in V : \mathcal{D}_v \text{ holds}\}$. We already know that $\mathbb{E}(|W|/n) \to \theta$ and $\mathbb{E}(|W|^2/n^2) \to \theta^2$, so that

$$\frac{|W|}{n} \to \theta$$

in probability. Furthermore, observe that if v, w are uniformly chosen in W, then $C_v = C_w$ with high probability depending solely on c by Lemmas 14 and 15. Also, if $v \in W$, then clearly $C_v \subset W$. Hence, W consists of a union of clusters. Let X_n denote the size of a cluster from W chosen according to size-biased picking, that is, $X_n =_d |C_v|/|W|$, where v is chosen uniformly at random in W. It is a well-known consequence of exchangeability (and easy to see) that

$$\mathbb{P}(C_v = C_w | v, w \in W) = \mathbb{E}(X_n).$$

By (14), if $L < n/(4LK)$ [resp., by (15) if $L \geq n/(4LK)$], we have that

$$\mathbb{P}(C_v = C_w | v, w \in W) \to 1,$$

hence, $\mathbb{E}(X_n) \to 1$. Since $X_n \leq 1$, it follows that $X_n \to 1$ in probability. This implies that, for all $\varepsilon > 0$, with high probability depending solely on ε and c, W contains a component of size at least $\frac{|W|}{n}(1 - \varepsilon) \geq \theta(c)(1 - 2\varepsilon)$. This proves the existence of a giant component of mass relative to V equal to θ in the limit $n \to \infty$.

Let us show that all other components are small. Note that by the above, we already know that the second largest component size, L_W^2, is such that $L_W^2/|W| \to 0$ in probability. Hence, $L_W^2/n \to 0$ as well. Let $L_{W^{\complement}}^1$ be the largest component size in $W^{\complement}$. By definition, $L_{W^{\complement}}^1$ is smaller in size than $\log L$. Since $L_n^2 \leq \max(L_{W^{\complement}}^1, L_W^2)$, we conclude that

$$\frac{L_n^2}{n} \to 0$$

in probability, as desired. The proof of (ii) in Theorem 1 is complete. □

We now conclude with the proof of (iii) in Theorem 1.

PROOF OF THEOREM 1(iii).   Since $L = o(\log n)$, we have $L < \frac{4a}{6c} \log n$, with $a > 0$ as in the statement of (iii). We begin by dividing $1, \ldots, n$ into $n^{1-a} \log n$ disjoint intervals of size $n^a / \log n$, labeled $A_1, \ldots, A_{n^{1-a} \log n}$. In each interval we show that we can find an interval of size L, none of whose vertices have been involved in a transposition by time t with high probability. We show in fact, that all the $n^{1-a} \log n$ intervals contain such a sub-interval with high probability. Thus the largest component must be of size smaller than $2n^a / \log n$, and hence, in particular, there will be no giant components.

For a given interval of size L, the number of potential edges connected to vertices in this interval is $2L^2 - \binom{L}{2} = (3L^2 + L)/2$. Each of these edges is present with probability $c/(2L)$. We call the interval empty if none of the edges are present. The probability a given interval of size L is empty is

$$\left(1 - c/(2L)\right)^{(3L^2+L)/2} \sim \exp\left(-\frac{c}{4}(3L + 1)\right).$$

We divide each A_i into $\lfloor n^a/(L \log n) \rfloor$ intervals of size L, denoted

$$A_i^1, A_i^2, \ldots, A_i^{\lfloor n^a/(L \log n) \rfloor}.$$

Let $S_i^{2k} = \{A_i^{2k} \text{ is empty}\}$. We consider the set of events

$$\{S_i^{2k}, 1 \le k \le \tfrac{1}{2} \lfloor n^a/(L \log n) \rfloor\},$$

which are independent since each interval is at distance at least L from any other. For each i, we let

$$B_i = \sum_{k=1}^{\lfloor n^a/(L \log n) \rfloor/2} \mathbf{1}_{\{S_i^{2k}\}}.$$

We have

$$\mathbb{P}(B_i > 0) \sim 1 - \exp\left[-\frac{1}{2}\left\lfloor \frac{n^a}{L \log n} \right\rfloor e^{-c(3L+1)/4}\right]$$

and so

$$\mathbb{P}(B_i > 0 \text{ for all } 1 \le i \le n^{1-a} \log n)$$

$$\sim \exp\left[-n^{1-a} \log n \exp\left(-\frac{1}{2}\left\lfloor \frac{n^a}{L \log n} \right\rfloor e^{-c(3L+1)/4}\right)\right]$$

$$\to 1 \qquad \text{as } n \to \infty$$

since $L < \frac{4a}{6c} \log n$. The proof of Theorem 1 is complete. □

5. Proof of Theorem 2. We will prove a stronger result than Theorem 2 by allowing the distribution of edge-lengths to be more general. Recall the definitions of (p_ℓ) and ε_n given at the beginning of Section 3. Let $(\tau_i)_{i \geq 1}$ be a sequence of i.i.d. transpositions with $\tau_1 = (i \ j)$ where i, j are chosen uniformly from $\{u, v \in V : \|u - v\| = L\}$. Then we construct the permutation $\sigma_t = \tau_1 \circ \cdots \circ \tau_{N_t}$, where $(N_t, t \geq 0)$ is an independent Poisson process. In words, at rate 1, we transpose two markers at random with distance D, where D is chosen according to the distribution (p_ℓ). We recover the process $(\sigma_t \geq 0)$ when (p_ℓ) is the uniform distribution on $\{1, \ldots, L(n)\}$.

THEOREM 6. *Assume $\varepsilon_n \to 0$ as $n \to \infty$. Then we have the following convergence in probability as $n \to \infty$: for all $c > 0$,*

$$\frac{1}{n}\delta(\sigma_{cn/2}) \to u(c).$$

There is a natural coupling between the process $(\sigma_t, t \geq 0)$ and the random graph $(G(t), t \geq 0)$ defined in Section 3. The coupling is an adaptation of the coupling with the Erdős–Renyi random graph in Berestycki and Durrett (2006). Consider the following procedure. Initially, G_0 consists of isolated vertices. Suppose that at time t, a transposition $\tau = (i, j)$ is performed. If $G(t)$ already contains the edge (i, j) we do nothing, else we add it to the graph.

The relationship between σ_t and $G(t)$ is not one-to-one; however, the following deterministic observation holds as in Berestycki and Durrett (2006). For every $t \geq 0$, every cycle of σ_t is a subset of a certain connected component of $G(t)$. That is, the partition of V obtained from considering the cycle decomposition of σ_t is a refinement of the partition obtained from considering the connected components of $G(t)$. This is easily proved by induction on the number N_t of transpositions up to time t, after observing that the cycle decomposition of σ_t undergoes a coagulation-fragmentation process. Indeed, every transposition (i, j) that involves two particles from the same cycle yields a fragmentation of that cycle, while if the two particles are in distinct cycles they merge.

This coupling is the basis of our proof. Armed with Lemmas 2 and 3, in order to prove Theorem 6 we need to show that K_t and $|\sigma_t|$ differ by $o(n)$ (Lemma 16), where we recall that $|\sigma|$ is the number of cycles of the permutation σ.

LEMMA 16. *Assume $\varepsilon_n \to 0$. Let $t = cn/2$, where $c > 0$. As $n \to \infty$,*

$$\frac{|\sigma_t| - K_t}{n} \to 0$$

in probability.

PROOF. This argument is somewhat analogous to the proof of Lemma 6 in Berestycki (2011). First we note that by the properties of the coupling between

σ_t and $G(t)$, it is with probability 1 the case that $K_t \leq |\sigma_t|$. To prove a bound in the converse direction, we need to distinguish between small and large cycles or components. We say that a cycle of σ_t or a component of $G(t)$ is *small* if it has a size less than $1/\sqrt{\varepsilon_n}$ and *large* if it has size at least $1/\sqrt{\varepsilon_n}$.

Note that the number of large cycles and the number of large components is at most $n\sqrt{\varepsilon_n} = o(n)$. It thus suffices to control the difference between the number of small cycles and the number of small components. However, note that at any time, the probability of generating a small cycle by fragmentation is at most $4(1/\sqrt{\varepsilon_n})\varepsilon_n$. To see where this comes from, suppose the current permutation is $\sigma_t = \sigma$, and the first position for the transposition (i, j) to be performed has been chosen. Thus j will be one of the $n - 1$ other vertices chosen according to the distribution (p_ℓ). Then to produce a cycle of size exactly k, j must be equal to $\sigma^{k-1}(i)$ or $\sigma^{-k+1}(i)$. (Depending on the exact size of the cycle containing i, there may be two other points allowed.) Thus, conditioning on the point i, the probability of creating a fragment of size smaller than $1/\sqrt{\varepsilon_n}$ is at most $4(1/\sqrt{\varepsilon_n})\varepsilon_n$, as claimed. It follows that since each excess small cycle must have been generated by such a fragmentation at some time $s \leq t$, and since transpositions occur at rate 1,

$$\mathbb{E}[|\sigma_t| - K_t] \leq n\sqrt{\varepsilon_n} + 4t\sqrt{\varepsilon_n}.$$

Thus by Markov's inequality, taking $t = cn/2$, for all $\delta > 0$,

$$\mathbb{P}\left(\frac{|\sigma_t| - K_t}{n} > \delta\right) \leq \frac{\mathbb{E}(|\sigma_t| - K_t)}{\delta n}$$

$$\leq \sqrt{\varepsilon_n}\frac{1 + 2c}{\delta} \to 0$$

as $n \to \infty$. The proof is complete. $\square$

PROOF OF THEOREM 6. The proof of Theorem 6 now follows directly from Lemmas 2, 3 and 16. Indeed, $\delta(\sigma_t) = n - |\sigma_t|$. By Lemma 16, $n^{-1}(|\sigma_t| - K_t)$ tends to 0 in probability. We have concentration of K_t around its mean by Lemma 3, and the mean is obtained in Lemma 2. Putting these pieces together we obtain Theorem 6. $\square$

6. Proof of Theorem 3. We consider the case where L is bounded (say by some constant C) and show that if $t = cn/2$ with $c > 0$, then $\delta(t)/n$ is bounded away from $c/2$, where we write $\delta(t) = \delta(\sigma_t)$.

LEMMA 17. *Assume L is bounded. Fix $c > 0$ and let $t = cn/2$. Then there exists $\eta = \eta_c > 0$ such that $\delta(t) \leq (1 - \eta)cn/2$ with high probability.*

In the statement above and in what follows, the expression *with high probability* means with probability tending to 1 as $n \to \infty$.

PROOF OF LEMMA 17.   Since each transposition decreases the number of cycles by 1 if there is a coagulation and increases it by 1 if there is a fragmentation, we have

$$(16) \qquad\qquad\qquad \delta(t) = N_t - 2F_t,$$

where F_t is the total number of fragmentations by time t. It suffices to show that $F_t \geq \eta n$ when $t = cn/2$, for some $\eta > 0$. Let $(i, j) \in \mathcal{R}_L$. Consider the event $\mathcal{A}_{ij}$ that the transposition (i, j) occurred twice by time t, and that no other transposition involved either i or j by time t. There are $4L - 2$ possible transpositions involving i or j but not both, with each occurring at rate $1/(nL)$. Thus the number of such transpositions that occur by this time is $\text{Poi}(t(4L - 2)/nL)$ which has a positive probability, q_c, of being 0. Further, the number of times transposition (i, j) occurs by time t is $\text{Poi}(t/nL)$ and thus we have a positive probability, p_c, of it occurring exactly twice. Thus $\mathbb{P}(\mathcal{A}_{i,j}) = q_c p_c > 0$ for each $(i, j) \in \mathcal{R}_L$.

Moreover, the events $(\mathcal{A}_{2iL,2iL+1})_{0 \leq i \leq \lfloor n/2L \rfloor - 1}$ are independent and each occurs with probability $q_c p_c$. Note that the number F_t of fragmentations satisfies

$$F_t \geq \sum_{i=0}^{\lfloor n/2L \rfloor - 1} \mathbf{1}_{\mathcal{A}_{2iL,2iL+1}}.$$

It thus follows from Chebyshev's inequality that $\mathbb{P}(F_t > n q_c p_c/(4C)) \to 1$, where C is an upper-bound on L. Hence, $F_t \geq \eta_c n$ with $\eta_c = q_c p_c/(4C)$. Plugging back in (16) completes the proof. □

We now turn toward the proof of Theorem 3. Assume without loss of generality that $L(n) = L$ is constant. As $N_t =_d \text{Poisson}(t)$, we obtain directly from Chebyshev's inequality that $\frac{1}{n} N_{cn/2} \to c/2$ in probability.

It thus suffices to show that $\frac{1}{n} F_{cn/2}$ also has a limit as $n \to \infty$. Let $(\mathcal{F}_t)_{t \geq 0}$ be the filtration associated with the entire history of the process; that is, $\mathcal{F}_t = \sigma(\tau_i, i \leq N_t)$. For $s \geq 0$, let $g_n(s)$ denote the $\mathcal{F}_s$-measurable random variable giving the instantaneous rate of fragmentation given σ_s. Let

$$A_t = \int_0^t g_n(s) \, ds$$

and observe that if $M_t = F_t - A_t$, then $(M_t, t \geq 0)$ is a martingale with respect to the filtration $(\mathcal{F}_t)_{t \geq 0}$, for each $n \geq 1$.

We prove convergence of $n^{-1} F_t$ (with $t = cn/2$) in two steps:

  (i) $n^{-1} A_t$ converges,
  (ii) $n^{-1} M_t \to 0$ in probability, which will follow from Doob's inequality.

Note first that by a change of variable,

$$(17) \qquad\qquad \frac{1}{n} A_{cn/2} = \frac{1}{2} \int_0^c g_n(sn/2) \, ds.$$

LEMMA 18. *There exists a nonrandom function $g(s)$ such that $\mathbb{E}(\frac{1}{n}A_{cn/2}) \rightarrow \frac{1}{2}\int_0^c g(s)\,ds$.*

PROOF. Since $g_n(s) \leq 1$ almost surely, it suffices to show (by Fubini's theorem and Lebesgue's dominated convergence theorem) that $\mathbb{E}(g_n(sn/2)) \rightarrow g(s)$ for all fixed $s > 0$. Let $\hat{C}_s$ be the cycle of σ_s containing the origin. By exchangeability, note that

$$\mathbb{E}(g_n(sn/2)) = \mathbb{P}(v \in \hat{C}_{sn/2}),$$

where v is chosen uniformly among the $2L - 1$ neighbors of 0. Fix such a neighbor v. The idea for the proof of this lemma is that the cycle structure of v can be coupled with the cycle structure of the origin in a random transposition process on the infinite line $\mathbb{Z}$, rather than on the torus. More precisely, let G_∞ be the graph where the vertex set is $V_\infty = \mathbb{Z}$ and the edge set is $E_\infty = \{(i, j) \in \mathbb{Z} \times \mathbb{Z}, |i - j| \leq L\}$. Consider the process $(\sigma_t^\infty, t \geq 0)$, with values in the permutation of V_∞, obtained by transposing each edge $(i, j) \in E_\infty$ at rate $1/(2L)$. It is not obvious that this process is well defined as there are an infinite number of edges. However, the process may be constructed using a standard *graphical construction* [see, e.g., Liggett (1985)]. Briefly speaking, for every (nonoriented) edge $e \in E_\infty$, consider an independent Poisson process which rings at rate $1/(2L)$. Then the value $\sigma_t^\infty(w)$ is defined for every $t \geq 0$ and $w \in V_\infty$ by following the trajectory between times 0 and t of a particle which is initially on w and moves to a neighbor j of its current position i each time the edge $e = (i, j)$ rings. It is easy to see (and will be shown below) that almost surely there are empty patches (where no edge has rung) surrounding the origin. Thus the trajectory cannot accumulate an infinite number of jumps in a compact interval, and hence, is well defined. Moreover, the cycle $\hat{C}_s^\infty$ of the origin in σ_s^∞ contains only finitely many points almost surely for $s \geq 0$, since it must be contained in between two empty patches.

Let $c > 0$. We claim that there is an event $\mathcal{G} = \mathcal{G}_n$ such that $\mathbb{P}(\mathcal{G}) \rightarrow 1$ as $n \rightarrow \infty$ and such that on $\mathcal{G}$, $\hat{C}_{sn/2}$ and $\hat{C}_s^\infty$ are identical. (Here we use the obvious identification of $V = \mathbb{Z}/n\mathbb{Z}$ as a subset of $\mathbb{Z}$, as $V = \{-\lfloor n/2 \rfloor + 1, \ldots, -1, 0, 1, \ldots, \lfloor n/2 \rfloor\}$.) We choose $g(s) = \mathbb{P}(v \in \hat{C}_s^\infty)$.

The event $\mathcal{G}$ we choose is

$$\mathcal{G}_n = \{\hat{C}_u \subset [-\log n, \log n] \text{ for all } u \leq sn/2\}.$$

The coupling between $\hat{C}_{sn/2}$ and $\hat{C}_s^\infty$ is obvious on $\mathcal{G}_n$ since we can use the same graphical construction for both σ_t and σ_t^∞. It remains to show that $\mathbb{P}(\mathcal{G}) \rightarrow 1$. To do this it suffices that there is a strip of size at least L in $[-\log n, 0]$ and in $[0, \log n]$ where each vertex in the strip has never been involved in a transposition by time $sn/2$ (we say that such a vertex has degree 0), what we called earlier an empty patch. A given interval of size L contains exactly $L(2L - 1) - \binom{L}{2} = (3L^2 - L)/2$ distinct edges, hence, the probability that it is an empty patch is

$$\exp\left(-\frac{s}{2L}\frac{3L^2 - L}{2}\right) =: p(s) > 0.$$

If some patches of size L share no edge in common, then the events that they are empty are mutually independent. Since we can find at least $\alpha \log n$ distinct patches that do not share any edge in $[0, \log n]$, for some $\alpha > 0$ depending only on L, the probability that there is no empty patch in $[0, \log n]$ is at most $(1 - p(s))^{\alpha \log n} \to 0$. Hence, $\mathbb{P}(\mathcal{G}_n) \to 1$ and Lemma 18 is proved. $\quad\square$

LEMMA 19. $\operatorname{Var}(\frac{1}{n} A_{cn/2}) \to 0$ as $n \to \infty$.

PROOF. Using (17) and Cauchy–Schwarz's inequality,

$$\operatorname{Var}\left(\frac{1}{n} A_{cn/2}\right) \le \frac{c}{4} \int_0^c \operatorname{Var}(g_n(sn/2)) \, ds.$$

Since $g_n(s) \le 1$, it suffices to show that $\operatorname{Var}(g_n(sn/2)) \to 0$ for all fixed $s > 0$. Now, note that

$$g_n(sn/2) = \frac{1}{n} \sum_{v \in V} f_v,$$

where

$$f_v = \frac{1}{2L - 1} \sum_{\|w - v\| \le L} \mathbf{1}_{\{w \in \hat{C}_v(sn/2)\}},$$

and where $\hat{C}_v(s)$ denotes the cycle containing v in σ_s. Let

$$\mathcal{A}_v = \{\hat{C}_v(r) \subset [v - \log n, v + \log n] \text{ for all } r \le sn/2\},$$

where the addition and substraction is done modulo n. If $\|v - v'\| > 2 \log n$, then on $\mathcal{A}_v \cap \mathcal{A}_{v'}$ the random variables f_v and $f_{v'}$ may be taken to be independent. Reasoning as in Lemma 3 shows that $\operatorname{Var}(\frac{1}{n} \sum_v f_v) \to 0$, since by Lemma 18 we know that $\mathbb{P}(\mathcal{A}_v \cap \mathcal{A}_{v'}) \to 1$. $\quad\square$

Our final step is to show that $M_{cn/2}/n$ converges in probability to 0.

LEMMA 20. *For all $\varepsilon > 0$,*

$$\mathbb{P}\left(\sup_{s \le cn/2} |M_s| > \varepsilon n\right) \to 0.$$

PROOF. By Markov's inequality,

$$\mathbb{P}\left(\sup_{s \le t} |M_s/n| > \varepsilon\right) = \mathbb{P}\left(\sup_{s \le t} |M_s/n|^2 > \varepsilon^2\right)$$

$$\le \frac{\mathbb{E}(\sup_{s \le t} |M_s/n|^2)}{\varepsilon^2}$$

$$\le \frac{4\mathbb{E}(M_t^2/n^2)}{\varepsilon^2}$$

by Doob's inequality. Now note that since M_t is a martingale whose jumps are only of size 1,

$$M_t^2 - \int_0^t g_n(s)\,ds$$

is again an $(\mathcal{F}_t)_{t \geq 0}$-martingale. [To see this, observe that $F_{A_t^{-1}}$ is $\mathrm{Poi}(t)$ and hence, $M_{A_t^{-1}}^2 - t$ is a martingale.] Thus $\mathbb{E}(M_t^2) \leq t$ for all $t \geq 0$ and when $t = cn/2$,

$$\mathbb{P}\Big(\sup_{s \leq cn/2} |M_s| > \varepsilon n \Big) \leq \frac{2c}{n\varepsilon^2} \to 0$$

as claimed. $\quad\square$

PROOF OF THEOREM 3.   It follows from Lemmas 18 and 19 that $\frac{1}{n} A_{cn/2} \to \frac{1}{2} \int_0^c g(s)\,ds$ in probability, where $g(s) = \mathbb{P}(v \in \hat{C}_s^\infty)$ has been defined in Lemma 18. By Lemma 20, we deduce that

$$\frac{1}{n} F_{cn/2} \to \frac{1}{2} \int_0^c g(s)\,ds$$

in probability. Since $\delta(t) = N_t - 2F_t$ for all $t \geq 0$, it follows that

$$\frac{1}{n} \delta(cn/2) \to v(c) = \frac{c}{2} - \int_0^c g(s)\,ds.$$

By Lemma 17, we must have $v(c) < c/2$ for all $c > 0$. It thus suffices to show that g is continuously differentiable on $[0, \infty)$. Assume that the process (σ_t^∞) is in some state such that the (finite) cycle C containing 0 also contains v. Let $f_1(C)$ denote the instantaneous rate at which v becomes part of a different cycle; note that this rate depends indeed only on C and not on the rest of σ_t^∞, and satisfies $f_1(C) \leq |C|^2/(2L)$. Likewise, assume that the cycle containing v, C', is distinct from C. Let $f_2(C, C')$ be the instantaneous rate at which these cycles merge. Then $f_2(C, C') \leq |C| \times |C'|/(2L)$.

Note that $|\hat{C}_c^\infty| \geq k$ implies that there are $\lfloor k/L \rfloor$ consecutive intervals of size L around 0 all containing at least one edge in the associated percolation process. By considering every other interval, this implies that we can find $\lfloor k/(2L) \rfloor$ disjoint intervals of size L, all of which contain at least one edge. Such events are independent, and hence, if $p_\infty(c) > 0$ is the probability that at time c an interval of size L is an empty patch, we find (summing over at most k possible locations for the leftmost point of this sequence of consecutive intervals),

$$\mathbb{P}(|\hat{C}_c^\infty| \geq k) \leq k\big(1 - p_\infty(c)\big)^{\lfloor k/2L \rfloor},$$

so that $|C_c^\infty|$ has exponential tails. It follows directly that $\mathbb{E}(|\hat{C}_c^\infty|^2) < \infty$, and if $C_c^\infty(v)$ denotes the cycle containing v at time c, $\mathbb{E}(|\hat{C}_c^\infty||\hat{C}_c^\infty(v)|) < \infty$ by Cauchy–Schwarz's inequality. A routine argument thus shows that

$$g'(c) = \mathbb{E}\big[\mathbf{1}_{\{v \notin \hat{C}_c^\infty\}} f_2(\hat{C}_c^\infty, \hat{C}_c^\infty(v))\big] - \mathbb{E}\big[\mathbf{1}_{\{v \in \hat{C}_c^\infty\}} f_1(\hat{C}_c^\infty)\big].$$

By the same arguments, we see that $g'(c)$ is continuous, which in turn shows that v is continuously twice differentiable. The proof of Theorem 3 is complete. $\square$

REFERENCES

AIZENMAN, M. and NEWMAN, C. M. (1986). Discontinuity of the percolation density in one-dimensional $1/|x - y|^2$ percolation models. *Comm. Math. Phys.* **107** 611–647. MR0868738

BERESTYCKI, N. (2011). Emergence of giant cycles and slowdown transition in random transpositions and k-cycles. *Electron. J. Probab.* **16** 152–173. MR2754801

BERESTYCKI, N. and DURRETT, R. (2006). A phase transition in the random transposition random walk. *Probab. Theory Related Fields* **136** 203–233. MR2240787

BERESTYCKI, N. and DURRETT, R. (2008). Limiting behavior for the distance of a random walk. *Electron. J. Probab.* **13** 374–395. MR2386737

BOLLOBÁS, B. (1985). *Random Graphs*. Academic Press, London. MR0809996

BOLLOBÁS, B., JANSON, S. and RIORDAN, O. (2009). Line-of-sight percolation. *Combin. Probab. Comput.* **18** 83–106. MR2497375

BRAMSON, M., DURRETT, R. and SWINDLE, G. (1989). Statistical mechanics of crabgrass. *Ann. Probab.* **17** 444–481. MR0985373

DIACONIS, P. and GRAHAM, R. L. (1977). Spearman's footrule as a measure of disarray. *J. Roy. Statist. Soc. Ser. B* **39** 262–268. MR0652736

DURRETT, R. (2002). *Probability Models for DNA Sequence Evolution*. Springer, New York. MR1903526

DURRETT, R. (2003). Shuffling chromosomes. *J. Theoret. Probab.* **16** 725–750. MR2009200

DURRETT, R. (2010). *Random Graph Dynamics*. Cambridge Univ. Press, Cambridge. MR2656427

HANNENHALLI, S. and PEVZNER, P. A. (1999). Transforming cabbage into turnip: Polynomial algorithm for sorting signed permutations by reversals. *J. ACM* **46** 1–27. MR1692509

KENT, W. J., BAERTSCH, R., HINRICHS, A., MILLER, W. and HAUSSLER, D. (2003). Evolutions cauldron: Duplication, deletion, and rearrangement in the mouse and human genomes. *Proc. Natl. Acad. Sci. USA* **100** 11484–11489.

LIGGETT, T. M. (1985). *Interacting Particle Systems. Grundlehren der Mathematischen Wissenschaften* [*Fundamental Principles of Mathematical Sciences*] **276**. Springer, New York. MR0776231

PENROSE, M. D. (1993). On the spread-out limit for bond and continuum percolation. *Ann. Appl. Probab.* **3** 253–276. MR1202526

YORK, T. L., DURRETT, R. and NIELSEN, R. (2002). Bayesian estimation of inversions in the history of two chromosomes. *J. Comput. Biol.* **9** 808–818.

STATISTICAL LABORATORY
UNIVERSITY OF CAMBRIDGE
WILBERFORCE ROAD
CAMBRIDGE, CB3 0WB
UNITED KINGDOM
E-MAIL: N.Berestycki@statslab.cam.ac.uk
            R.Pymar@statslab.cam.ac.uk