

Nathanaël Berestycki · Rick Durrett

## A phase transition in the random transposition random walk

Received: 5 March 2004 / Revised version: 13 August 2005 /  
Published online: ■■ 2005 – © Springer-Verlag 2005

**Abstract.** Our work is motivated by Bourque and Pevzner's (2002) simulation study of the effectiveness of the parsimony method in studying genome rearrangement, and leads to a surprising result about the random transposition walk on the group of permutations on  $n$  elements. Consider this walk in continuous time starting at the identity and let  $D_t$  be the minimum number of transpositions needed to go back to the identity from the location at time  $t$ .  $D_t$  undergoes a phase transition: the distance  $D_{cn/2} \sim u(c)n$ , where  $u$  is an explicit function satisfying  $u(c) = c/2$  for  $c \leq 1$  and  $u(c) < c/2$  for  $c > 1$ . In addition, we describe the fluctuations of  $D_{cn/2}$  about its mean in each of the three regimes (subcritical, critical and supercritical). The techniques used involve viewing the cycles in the random permutation as a coagulation-fragmentation process and relating the behavior to the Erdős-Renyi random graph model.

### 1. General motivation

The relationship between the orders of genes in two species can be described by a signed permutation. For example the relationship between the human and mouse  $X$  chromosomes may be encoded as (see Pevzner and Tesler (2003))

$$1 \quad -7 \quad 6 \quad -10 \quad 9 \quad -8 \quad 2 \quad -11 \quad -3 \quad 5 \quad 4$$


In words the two  $X$  chromosomes can be partitioned into 11 segments. The first segment of the mouse  $X$  chromosome is the same as that of humans, the second segment of mouse is the 7th human segment with its orientation reversed, etc. The parsimony approach to estimation of evolutionary changes of the  $X$  chromosome between human and mouse is to ask: what is the minimum number of reversals (i.e., moves that reverse the order of a segment and therefore change its *sign*) needed to transform the arrangement above back into  $1, \dots, 11$ ? In other words, what is the (reversal) distance between the human and mouse  $X$  chromosomes?

Hannehalli and Pevzner (1995) developed a polynomial algorithm for answering this question. The first step in preparing to use the Hannehalli-Pevzner algorithm

N. Berestycki: Ecole Normale Supérieure, Département de Mathématiques et Applications, 45, rue d'Ulm F-75005 Paris, France

N. Berestycki, R. Durrett: Department of Mathematics, Malott Hall, Cornell University, Ithaca, NY 14853, U.S.A.

*Key words or phrases:* Random transposition – Random graphs – Phase transition – Coagulation-fragmentation – Genome rearrangement – Parsimony method

	<b>4 4 0 0 4 7 9 B</b>	Dispatch: 8/10/2005	Journal: PTRF
	Jour. No	Ms. No.	Total pages: 31
		Disk Received <input checked="" type="checkbox"/>	Not Used <input type="checkbox"/>
		Disk Used <input checked="" type="checkbox"/>	Corrupted <input type="checkbox"/>
			Mismatch <input type="checkbox"/>

is to double the markers. When segment  $i$  is doubled we replace it by two consecutive numbers  $2i - 1$  and  $2i$ , e.g., 6 becomes 11 and 12. A reversed segment  $-i$  is replaced by  $2i$  and  $2i - 1$ , for example,  $-7$  is replaced by 14 and 13. The doubled markers use up the integers 1 to 22. To these numbers we add a 0 at the front and a 23 at the end. Using commas to separate the ends of the markers we can write the two genomes as follows:

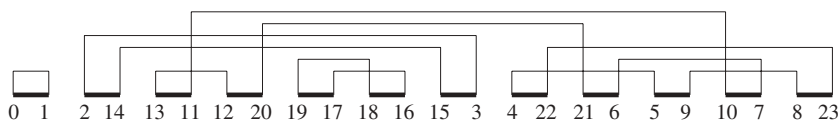
mouse 0, 1 2, 14 13, 11 12, 20 19, 17 18, 16 15, 3 4, 22 21, 6 5, 9 10, 7 8, 23  
 human 0, 1 2, 3 4, 5 6, 7 8, 9 10, 11 12, 13 14, 15 16, 17 18, 19 20, 21 22, 23

The next step is to construct the breakpoint graph (see Figure 1) that results when the commas are replaced by edges that connect vertices with the corresponding numbers. In the picture we have written the vertices in their order in the mouse genome. Commas in the mouse order become thick lines (black edges), while those in the human genome are thin lines (gray edges).

Each vertex has one black and one gray edge, so the connected components of the graph are easy to find: start with a vertex and follow the connections in either direction until you come back to where you start. In this example there are five components:

0 - 1 - 0      2 - 14 - 15 - 3 - 2      4 - 22 - 23 - 8 - 9 - 5 - 4  
 19 - 17 - 16 - 18 - 19      13 - 11 - 10 - 7 - 6 - 21 - 20 - 12 - 13

To compute a lower bound for the distance, we take the number of commas seen when we write out one genome. In this example that is 12. In general, it is 1 plus the number of markers. We then subtract the number of components in the breakpoint graph. In this example that is 5, so the result is 7. This is a lower bound on the distance, since any reversal can at most reduce this quantity by 1, and it is 0 when the two genomes are the same. We can verify that 7 is the minimum distance by constructing a sequence of 7 moves that transforms the mouse  $X$  chromosome into the human order. There are thousands of solutions, so we leave this as an exercise for the reader. Here are some hints: (i) To do this it suffices, at each step, to choose a reversal that increases the number of cycles by 1. (ii) This never occurs if the two chosen black edges are in different cycles. (iii) If the two black edges are in the same cycle and are  $(a, b)$  and  $(c, d)$  as we read from left to right, this will occur unless in the cycle minus these two edges  $a$  is connected to  $d$  and  $b$  to  $c$ , in which case the number of cycles will not change. For example, in the graph in Figure 1 a reversal that breaks black edges 19-17 and 18-16 will increase the number of cycles but the one that breaks 2-14 and 15-3 will not.



**Fig. 1.** Breakpoint graph for human-mouse  $X$  chromosome comparison

In general, the distance between genomes can be larger than the lower bound from the breakpoint graph. There can be obstructions called *hurdles* that can prevent us from decreasing the distance, and hurdles can be intertwined in a *fortress of hurdles* that takes an extra move to break. See Hannehalli and Pevzner (1995). In symbols, if  $\pi$  is the signed permutation that represents the relative order and orientation of segments in the two genomes, then

$$d(\pi) = n + 1 - c(\pi) + h(\pi) + f(\pi)$$

where  $d(\pi)$  is the distance from the identity,  $n$  is the number of markers,  $c(\pi)$  is the number of components in the breakpoint graph,  $h(\pi)$  is the number of hurdles, and  $f(\pi)$  is the indicator of the event  $\pi$  is a fortress of hurdles. See Section 5.2 of Durrett (2002) or Chapter 10 of Pevzner (2000) for more details.

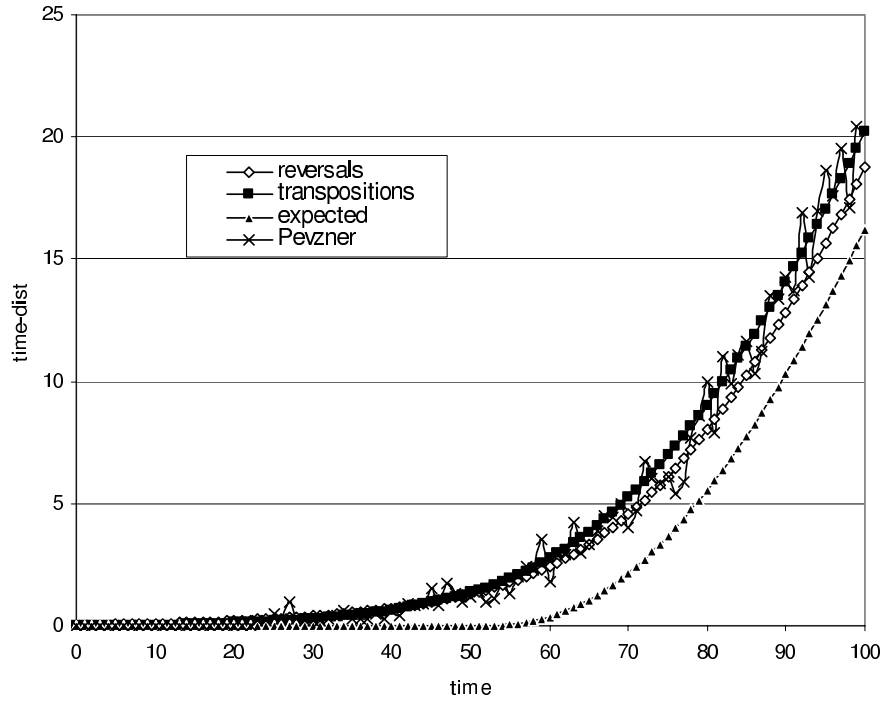
Although  $d_0(\pi) = n + 1 - c(\pi)$  is only a lower bound on the distance, it is the right answer in most biological examples. Bafna and Pevzner (1995) consider 11 comparisons of mitochondrial and chloroplast genomes and found that this lower bound gave the right answer in all cases. This pattern has continued in more recent work, see York, Durrett, and Nielsen (2002), and Durrett, Nielsen, and York (2003). The simulations in Figure 2 will give more evidence that  $d_0(\pi)$  and  $d(\pi)$  are close in many cases.

To motivate our main question, we will introduce a second data set. Ranz, Casals, and Ruiz (2001) located 79 genes on chromosome 2 of *D. repleta* and on chromosome arm 3R of *D. melanogaster*. If we number the genes according to their order in *D. repleta* then their order in *D. melanogaster* is given in Table 1. This time we do not know the orientation of the segments, but that is not a serious problem. Using simulated annealing, one can easily find an assignment of signs that minimizes the distance, which in this case is 54. Given the large number of rearrangements relative to the number of markers, we should ask: when is the parsimony estimate reliable?

Bourque and Pevzner (2002) approached this question by taking 100 markers in order, performing  $k$  randomly chosen reversals to get a permutation  $\pi_k$ , computing the minimum number of reversals needed to return to the identity,  $d(\pi_k)$ , and then plotting the average value of  $d(\pi_k) - k \leq 0$  for 100 simulations. They concluded, based on their simulations, that the parsimony distance for  $n$  markers was a good estimate as long as the number of reversals performed was at most  $0.4n$ . In Figure 2 we have given  $-1$  times their data. We have also repeated their experiment for

**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	



**Fig. 2.** Simulations with  $n = 100$  markers. Average values of  $k - D_k$  for 10,000 simulations of the random transposition chain, 10,000 simulations of  $k - d_0(\pi_k)$  and Pevzner's 100 simulations of  $k - d(\pi_k)$  for the reversal chain. The smooth curve of small triangles gives the limiting behavior as  $n \rightarrow \infty$  of  $(cn/2 - D_{cn/2})$  from Theorem 3

the approximate distance  $d_0(\pi) = n + 1 - c(\pi)$  and plotted the average value of  $k - d_0(\pi_k) \geq 0$  for 10,000 replications. Our curve is less random, but close to data of Bourque and Pevzner (2002). The smooth curve gives the result of Theorem 3 for the limiting behavior of  $(tn - d_0(\pi_{tn}))/n$  (as a function of  $t$ ).

The biological question concerns the random reversal walk. However, it is also interesting to consider the analogous problem for random transpositions. In that case the distance from the identity can be easily computed: it is the number of markers  $n$  minus the number of cycles in the permutation. For an example, consider the following permutation of 14 objects written in its cyclic decomposition:

$$(1\ 7\ 4)\ (2)\ (3\ 12)\ (5\ 13\ 9\ 11\ 6)\ (8\ 10\ 14)$$

which indicates that  $1 \rightarrow 7, 7 \rightarrow 4, 4 \rightarrow 1, 2 \rightarrow 2, 3 \rightarrow 12, 12 \rightarrow 3$ , etc. There are 5 cycles so the distance from the identity is 9. If we perform a transposition that includes markers from two different cycles (e.g., 7 and 9) the two cycles merge into 1, while if we pick two in the same cycle (e.g., 13 and 11) it splits into two.

The situation is similar but slightly more complicated for reversals. There a reversal that involves edges in two different components merges them into 1, but a reversal that involves two edges of the same cycle may or may not increase the

number of cycles. One can attempt to couple the components of the breakpoint graph for random reversals on  $n - 1$  markers and the cycles of random transposition of  $n$  markers as follows: number the edges between markers in the reversal chain (including the ends 0 and  $n$ ); when markers  $i$  and  $j$  are transposed, do the inversion of edges numbered  $i$  and  $j$ . The result of the coupled simulation is given in Figure 2. As expected time minus distance is smaller for reversals but the qualitative behavior is similar. Thus, we will begin by considering the biologically less relevant case of random transpositions, and ask a question that in terms of the rate 1 continuous time random walk on the symmetric group is: how far from the identity are we at time  $cn/2$ ? We will see later that parts of the answer can be extended to the reversal random walk.

## 2. The coagulation-fragmentation process and the random graph process

Let  $(\sigma_t, t \geq 0)$  be the continuous-time random walk on the group of permutations, starting at the identity, in which, at times of a rate one Poisson process, we perform a transposition of two elements chosen uniformly at random, with replacement, from  $\{1, \dots, n\}$ . Choosing with replacement causes the chain to do nothing with probability  $1/n$ , but makes some of the calculations a little nicer. If we think of the permutation  $\sigma$  as being represented by numbered balls sitting on numbered locations with ball  $\sigma(k)$  sitting at  $k$ , then transposition of  $i$  and  $j$ ,  $\rho_{i,j}$ , can be implemented in two ways. We can exchange the balls at  $i$  and  $j$  or the balls numbered  $i$  and  $j$ . Algebraically these correspond to  $\rho_{i,j}\sigma$  and  $\sigma\rho_{i,j}$ . Since  $(\sigma\rho_{i,j})^{-1} = \rho_{i,j}\sigma^{-1}$  and the partition of  $\{1, \dots, n\}$  induced by the cycle decompositions of  $\sigma$  and  $\sigma^{-1}$  are equal, the results are the same for either random walk.

Define the distance to the identity  $D_t$  to be the minimum number of transpositions one needs to perform on  $\sigma_t$  to go back to the identity element. A different way of looking at  $D_t$  is the following.  $(\sigma_t, t \geq 0)$  can be viewed as a random walk on a graph  $G$ , where  $G$  is the Cayley graph of the symmetric group for the set of generators given by the set of all transpositions. Using this language, we see that  $D_t$  is nothing but the graph distance from  $\sigma_t$  to the origin, the identity element.

It is clear that if  $N_t$  is the number of transpositions distinct from the identity performed up to time  $t$  (a Poisson random variable with mean  $t(1 - 1/n)$ ), then  $D_t \leq N_t$ . As mentioned earlier  $D_t$  is given by  $D_t = n - |\sigma_t|$ , where  $|\sigma_t|$  is the number of cycles in the cycle decomposition of  $\sigma_t$ . This formula allows us to turn any question about  $D_t$  into a question about  $|\sigma_t|$ . The key to studying  $|\sigma_t|$  is that the cycles evolve according to the dynamics of a coagulation-fragmentation process. When a transposition  $\rho_{i,j}$  occurs, if  $i$  and  $j$  belong to two different cycles then the cycles merge. On the contrary, if they belong to the same cycle, this cycle is split into two cycles. From the definition it can be seen that the ranked sizes of the cycles form a coagulation-fragmentation process (see Aldous (1999) and Pitman (2002, 2003)) in which components of size  $x$  and  $y$  merge at rate  $K_n(x, y) = 2xy/n^2$  and components of size  $x$  split at rate  $F_n(x) = x(x - 1)/n^2$  and are broken at a uniformly chosen random point. Diaconis, Mayer-Wolf, Zeitouni, and Zerner (2003) have recently considered the corresponding Markov chain on partitions of the unit

interval and shown that the Poisson-Dirichlet distribution is the unique invariant measure.

To study the evolution of the cycles in the random permutation, we construct a random graph process  $G_t^*$ . Start with the initial graph on vertices  $\{1, \dots, n\}$  with no edge between the vertices. When a transposition of  $i$  and  $j$  occurs in the random walk, we draw an edge between the vertices  $i$  and  $j$ , even if one is already present. Elementary properties of the Poisson process imply that if we collapse multiple edges in  $G_t^*$  into one, then the resulting graph  $G_t$  is a realization of the Erdős-Renyi random graph  $G(n, p)$ , in which edges are independently present with probability  $p = 1 - \exp(-2t/n^2)$ . The probability of picking an edge twice is  $\leq (2t/n^2)^2 = O(1/n^2)$  when  $t = cn/2$ , so the expected number of multiple edges is  $O(1)$ . Multiple edges are a nuisance but not a real problem. We have to be careful in Theorem 1 where the random variable of interest is also  $O(1)$ , but in the other cases the quantities of interest  $\rightarrow \infty$ , so multiple edges can be ignored.

It is easy to see that in order for two integers to be in the same cycle in the permutation it is necessary that they are in the same component of the random graph. To estimate the difference between cycles and components, let  $F_t$  denote the event that a fragmentation occurs at time  $t$ . It is clear that

$$D_t = N_t - 2 \sum_{s \leq t} \mathbf{1}_{\{F_s\}} \quad (1)$$

A fragmentation occurs in the random permutation when a transposition occurs between two integers in the same cycle, so tree components in the random graph  $G_t^*$  correspond to unfragmented cycles in the random walk. (To be precise, a tree is a connected component with no closed circuits and hence no multiple edges.) Unicyclic components in  $G_t^*$  (connected components with an equal number of vertices and edges) correspond to cycles in the permutation that have experienced exactly one fragmentation, but we need to know the order in which the edges were added to determine the resulting cycles. For more complex components, the relationship between the random graph and the permutation is less clear. Fortunately, these can be ignored in the proofs of our results. Coming back to the problem of multiple edges, the reader should check that in the proof of Theorem 1, in particular in Lemma 4, it is indeed the number of fragmentations that is being counted, and not just the number of cycles in  $G_t$ .

### 3. Limit Theorems

We will now describe our results and sketch their proofs. Rigorous proofs of the results stated in this section can be found in Sections 4, 5 and 6.

#### 3.1. The subcritical regime

**Theorem 1.** *Let  $0 < c < 1$ . The number of fragmentations*

$$Z_c := \sum_{s \leq cn/2} \mathbf{1}_{\{F_s\}} \Rightarrow \text{Poisson}(\kappa(c)) \quad (2)$$

where  $\kappa(c) = (-\log(1-c) - c)/2$ . In fact, the convergence holds for the process  $\{Z_c : 0 \leq c < 1\}$  with the limit being a Poisson process with compensator  $\kappa(c)$ .

*Remark.* This result should be regarded as a fluctuation result for  $D_{cn/2}$  about its mean  $cn/2$  (a corollary of the Theorem). However we have chosen to formulate it in terms of fragmentations rather than the distance, since by (1)  $D_{cn/2} - cn/2 \approx N_{cn/2} - cn/2 = O(n^{1/2})$ . That is, in continuous time and in the subcritical regime the fluctuations are due to those of the Poisson process. However for the embedded discrete time chain, if  $k = \lfloor cn/2 \rfloor$ , then

$$(k - D_k)/2 \Rightarrow \text{Poisson}(\kappa(c)) \text{ as } n \rightarrow \infty \quad (3)$$

(We divide by 2 since a fragmentation reduces the distance by 1 instead of increasing it by 1). To deduce (3) from (2) we note that time  $k$  in the discrete walk corresponds to time  $N^{-1}(k) \approx cn/2$  in the continuous time walk.

**Sketch of the proof.** The process  $\{Z_c, 0 \leq c < 1\}$  is a càdlàg counting process. Therefore by arguments from Jacod and Shiryaev (1987), it is enough to show that its compensator  $\kappa^n$  converges to the deterministic limit  $\kappa(c)$ . If  $f_k(t)$  is the fraction of vertices that belong to cycles of size  $k$ , the rate at which fragmentations occur is just  $\sum_k f_k(t)(k-1)/n$ . Hence  $\kappa^n$  is just the integral with respect to time of this rate. We first show that the variance converges to 0 and then, by Chebycheff's inequality, it only remains to show  $E\kappa^n(c) \rightarrow \kappa(c)$ . But by exchangeability  $E[f_k(t)] = P[|\mathcal{C}_1| = k]$  where  $|\mathcal{C}_1|$  is the size of the component that contains 1 at time  $t$ . It is not hard to see that this quantity at time  $bn/2$  converges in distribution to the total progeny  $\tau$  of a Galton-Watson branching process with offspring distribution  $\text{Poisson}(b)$ , or  $PGW(b)$ . Summing the geometric series, we see that  $E\tau = 1/(1-b)$ . Integrating with respect to  $b$  we get the desired expected value,  $\kappa(c)$ .  $\square$

To prepare for later developments, it is useful to take a second combinatorial approach to this result. We begin with Cayley's result that there are  $k^{k-2}$  trees with  $k$  labeled vertices. At time  $cn/2$  each edge is present with probability  $1 - \exp(-c/n) \sim c/n$  so the expected number of trees of size  $k$  present is

$$\sim \binom{n}{k} k^{k-2} \left(\frac{c}{n}\right)^{k-1} \left(1 - \frac{c}{n}\right)^{k(n-k) + \binom{k}{2} - k + 1} \quad (4)$$

since each of the  $k-1$  edges needs to be present and there can be no edges connecting the  $k$  point set to its complement or any other edges connecting the  $k$  points. (To justify replacing  $\exp(-c/n)$  by  $1 - c/n$  note that the difference is  $O(1/n^2)$ .) For fixed  $k$  (4) is asymptotic to

$$n \frac{k^{k-2}}{k!} c^{k-1} \left(1 - \frac{c}{n}\right)^{kn}$$

The quantity in parentheses at the end converges to  $e^{-ck}$  so we have an asymptotic formula for the number of tree components at time  $cn/2$ . As a side result we get the following known result:

**Corollary 1.** *The probability distribution of the total progeny  $T$  of a Poisson( $c$ ) branching process with  $c < 1$  is given by  $P(T = k) = \frac{1}{c} \frac{k^{k-1}}{k!} (ce^{-c})^k$*

See Section 4.1 of Pitman (1999) for another proof of this result. It was first discovered by Borel (1942) and the distribution of  $T$  is called the Borel distribution. It is a particular case of the so-called Borel-Tanner distribution, see Devroye (1992) and Pitman (1998) for further references. In this context it appeared in the problem of the total number of units served in the first busy period of a queue with Poisson arrivals and constant service times. See also Tanner (1961). Of course, this becomes a branching process if we think of the customers that arrive during a person's service time as their children.

### 3.2. The critical regime

It is well known in the theory of random graphs that the correct time-scale to describe the critical regime is  $(n/2)(1 + \lambda n^{-1/3})$ ,  $\lambda \in (-\infty, \infty)$ . See Aldous (1997) for an interesting account that relates the growth of large clusters in the critical random graph to the multiplicative coalescent. At times  $(n/2)(1 - n^{-r})$  with  $r < 1/3$ , we are still in the subcritical regime, so the arguments in the proof of Theorem 1, when done more carefully, are still valid. More precisely, we can show that if  $c_n(r) = 1 - n^{-r/3}$  for  $0 \leq r \leq 1$ , then the expected number of fragmentations up to time  $c_n(r)n/2$  is again given by  $\kappa(c_n(r)) \sim (r/6) \log n$ . Hence define:

$$W_n(r) = \left( \frac{6}{\log n} \right)^{1/2} \left( \sum_{s \leq c_n(r)n/2} \mathbf{1}_{\{F_s\}} - \frac{r}{6} \log n \right) \quad (5)$$

**Theorem 2.** *As  $n \rightarrow \infty$ ,  $W_n(\cdot)$  converge weakly, with respect to the Skorokhod topology on the space of càdlàg functions on  $[0, 1]$ , to  $\{W(r), 0 \leq r \leq 1\}$ , a standard Brownian Motion on  $[0, 1]$ . Furthermore,*

$$\left( \frac{6}{\log n} \right)^{1/2} \left( \sum_{s \leq n/2} \mathbf{1}_{\{F_s\}} - \frac{1}{6} \log n \right) \Rightarrow W(1), \quad (6)$$

**Sketch of the proof.** Intuitively, the first result is an immediate consequence of the Poisson limit in Theorem 1 and the normal approximation to the Poisson. To prove it, we show that  $W_n(r)$  is a martingale, whose jumps are asymptotically zero, and whose quadratic variation process is  $r$  thanks to our time-change  $c_n(r) = 1 - n^{-r/3}$ . Therefore it converges to Brownian Motion.

At times  $(1 - n^{-1/3})n/2 \leq t \leq n/2$  we are in the critical range of the random graph. Results of Luczak, Pittel, and Wierman (1994) and computations with (4) imply that the number of fragmentations in this interval is bounded in expectation and hence can be ignored.  $\square$

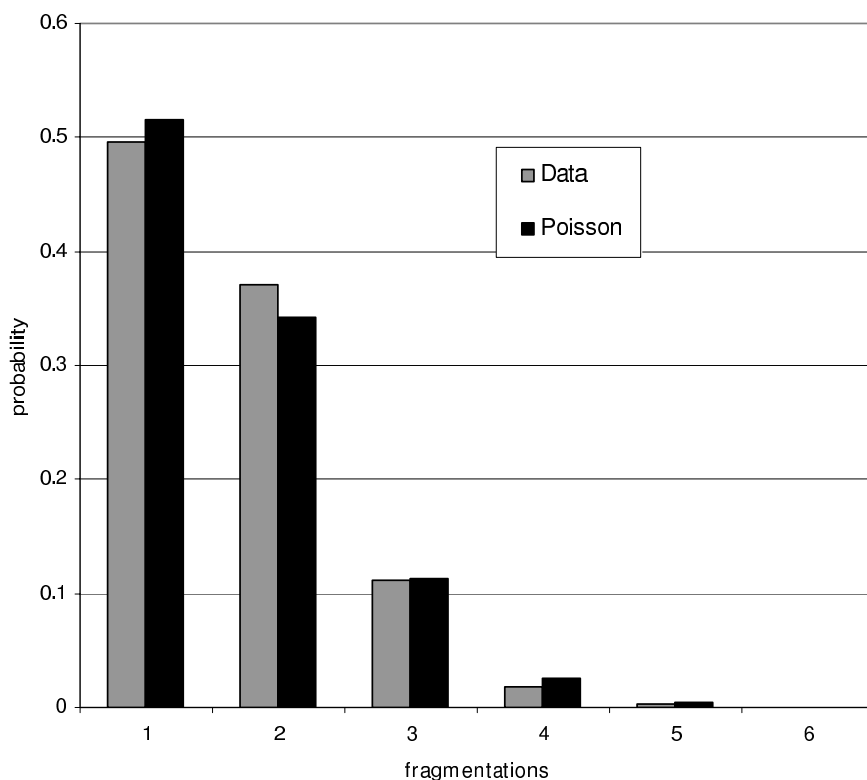
*Remark.* While Theorem 2 is a nice theoretical result, it does not have much to say about any biological example. If we think of the human genome and set  $n = 3$

billion nucleotides, Theorem 2 says that after  $n/2 = 1.5$  billion transpositions there have been an average of  $(\log n)/6 = 3.63$  fragmentations, with a standard deviation of 1.91. These numbers are small so even for  $n = 3$  billion, we can't expect a very good approximation to the normal distribution. In the example that we simulated  $n = 100$  and  $(\log n)/6 = 0.767$  versus an observed average number of fragmentations = 0.662 (which translates into a value of 1.224 in Figure 2). While our estimation of the mean is not very accurate, Figure 3 shows that the distribution of the number of fragmentations is almost Poisson.

### 3.3. The supercritical regime

This is the most interesting case, and also the hardest one. We start by establishing a law of large numbers. For all  $c > 0$  define

$$\beta_k(c) = \frac{1}{c} \frac{k^{k-1}}{k!} (ce^{-c})^k$$



**Fig. 3.** Comparison of the distribution of the number of fragmentations in 10,000 simulations of the random transposition chain with the Poisson distribution with the same mean

so that for  $c < 1$  it coincides with the Borel distribution of Corollary 1. When  $c > 1$ ,

$$\lim_{n \rightarrow \infty} P(|\mathcal{C}_1| = k) = \beta_k(c)$$

still holds but the  $\beta_k(c)$ 's no longer sum up to 1 because there is a probability  $\beta_\infty(c) = 1 - \sum_{k \geq 1} \beta_k(c) > 0$  that  $\mathcal{C}_1$  is the giant component.

Let us denote by  $\Upsilon(c)$  a random variable that takes the value  $1/k$  with probability  $\beta_k(c)$  when  $1 \leq k < \infty$  and the value 0 with probability  $\beta_\infty(c)$ . The motivation for this definition is that in the limit as  $n \rightarrow \infty$   $1/|\mathcal{C}_m|$  has the same distribution as  $\Upsilon(c)$  and  $\sum_{m=1}^n 1/|\mathcal{C}_m|$  gives the number of components in the random graph.

**Theorem 3.** *Let  $c > 0$  be a fixed positive number. Then the number of cycles in the random permutation at time  $cn/2$ ,  $|\sigma_{cn/2}| = g(c)n + \omega(\sqrt{n})$ , where*

$$g(c) := E\Upsilon(c) = \sum_{k=1}^{\infty} \frac{1}{c} \frac{k^{k-2}}{k!} (ce^{-c})^k \quad (7)$$

and the error term  $\omega(\sqrt{n})/a_n\sqrt{n} \rightarrow 0$  in probability if  $a_n \rightarrow \infty$ .

Note that the theorem is valid for all regimes and implies that the distance is given by  $D_{cn/2} = u(c)n + \omega(\sqrt{n})$  where  $u(c) = 1 - g(c)$ . Although it is not obvious from the formula,  $u(c) = c/2$  for  $c < 1$  and  $u(c) < c/2$  when  $c > 1$ . Using Stirling's formula,  $k! \sim k^k e^{-k} \sqrt{2\pi k}$ , it is easy to check that  $g'$  exists for all  $c$  and is continuous, but  $g''(1)$  does not exist. In words, there is phase transition in the behavior of the distance of the random walk to the identity at time  $n/2$  from linear to sublinear.

*Proof.* In the supercritical regime the dynamics of the large components is quite complicated, but there can never be more than  $\sqrt{n}$  components of size  $\sqrt{n}$  or larger. On the other hand, the expected number of all fragmentations that produce (regardless of the initial size of the cycle) clusters of size smaller than  $\sqrt{n}$  by time  $cn/2$  is at most  $O(n^{1/2})$ . This follows from the following important remark, which will be used on several other occasions implicitly: suppose  $C = (x_1, \dots, x_k)$  is a cycle of the permutation  $\sigma$ . If we transpose  $x_i$  and  $x_j$  with  $x_i \neq x_j$  then  $C$  breaks into  $(x_1, \dots, x_{i-1}, x_j, x_{j+1}, \dots, x_k)$  on the one hand and  $(x_i, \dots, x_{j-1})$  on the other hand. So, to generate a fragment of size  $s$  we must transpose two elements  $(x, y)$  in the same cycle and such that  $x$  and  $y$  are separated by exactly  $s - 1$  other integers in this cycle. In particular, fragmentations that produce pieces smaller than  $s$  occur with a rate smaller than  $2s/n$ . When  $s = n^{1/2}$  this gives a rate smaller than  $O(n^{-1/2})$ . Thus by time  $cn/2$  there have been no more than  $O(n^{-1/2}) \cdot cn/2 = O(n^{1/2})$  such fragmentations. From this and Chebyshev's inequality we see that up to a term  $\omega(n^{1/2})$ ,  $|\sigma_{cn/2}|$  is the number of components of the random graph, and the result follows Theorem 12 in Chapter V of Bollobás (1985).  $\square$

**Theorem 4.** *Let  $c > 1$ . As  $n \rightarrow \infty$ ,*

$$\frac{D_{cn/2} - u(c)n}{n^{1/2}} \Rightarrow \mathcal{N}(0, \sigma^2) \quad (8)$$

where  $\sigma = \rho[1 + \rho(c/2 - 1)]$ , and  $\rho = 1 - \theta(c)$  is the extinction probability of a supercritical PGW( $c$ ).

*Remark.* Note that the constant  $\sigma$  is different from the one given in Berestycki and Durrett (2003). We were correct in claiming that the central limit theorem in Theorem 4 is the same as the one for the number of components of the random graph, but we naively thought that the terms in  $\sum_{k=1}^n 1/|C_k|$  were sufficiently independent so that  $\sigma^2 = \text{var}(\Upsilon(c))$ .

**Sketch of Proof.** By Pittel's (1990) central limit theorem for the number of components of a random graph, it suffices to prove that the number of extra components due to fragmentation at time  $cn/2$  is  $o(\sqrt{n})$  (see his Corollary 1 and note that  $T/c = \rho$ ). Our first step is to increase the cutoff for large cycles to  $n^a$  where  $a > 1/2$ , so that the number of large cycles is at most  $n^{1-a} = o(n^{1/2})$ . The number of fragmentations that produce "small" cycles is now  $n^{-(1-a)} \cdot cn/2 = O(n^a)$  and cannot be ignored, so we need to use the fact that fragmented cycles are reabsorbed by the large components. If the fraction of mass in large cycles ("upstairs") at time  $tn$  is  $\lambda_t$  then new fragments of size  $k$  are produced at rate  $\leq 2\lambda_t$  and each fragment of size  $k$  is reabsorbed at rate  $2k\lambda_t$ . After time change this is bounded by an  $M/M/\infty$  queue in which the expected number of customers in equilibrium is  $1/k$ . Using this, we can show that with high probability the number of small fragments at any time is at most  $(\log n)^2$ . Of course, the coagulation fragmentation process is not exactly the queuing system. Customers can split into two, coalesce with other customers, gain weight (and increase their fragmentation rate) by eating small components, etc. However,  $(\log n)^2$  is much smaller than  $n^{1/2}$  so crude but robust estimates and patience eventually lead to a proof.  $\square$

### 3.4. Results for Reversals

Theorems 3 and 4 extend easily to the approximate distance for reversal chain. Recall that the main difference lies in the fact that, a reversal involving edges from different components in the breakpoint graph always yields a coagulation, but one involving two edges in the same component may or may not cause a fragmentation. The proofs of Theorems 3 and 4 for transpositions are based on showing that fragmentations can be ignored, so this difference is unimportant and these results extend to reversals. As Figure 2 shows, this is not true for the more precise results in Theorems 1 and 2. For example, the underlying data shows that up to  $c = 1$ , an average of 23% of the reversals have caused no change in the distance. Since inversions that affect an edge are much more frequent than those that involve it, it seems reasonable to guess that in the limit as  $n \rightarrow \infty$  the relative orientations of the black edges in a component of the breakpoint graph are independent. This would imply that the Poisson process of fragmentations in the reversal case is a 1/2-thinning of the one for transpositions, and Theorem 2 would hold with 6 replaced by 12.

### 3.5. Emergence of a giant cycle?

Since cycles in the random permutation are smaller than components of the random graph, it follows that if  $c < 1$  then the largest cycle at time  $cn/2$  has fewer than  $\alpha(c)^{-1} \log n$  vertices, where  $\alpha(c) = (c - 1 - \log c)$ . (See Theorem 10 in Chapter V of Bollobás (1985) or Lemma 3 below.)

For  $c > 1$ , the largest component of the random graph is, as is well known, “giant,” meaning that it is of order  $n$ . In fact it is asymptotic to  $\theta(c)n$  where  $\theta(c)$  is the survival probability of a supercritical Poisson Galton-Watson with mean  $c$ . It is a natural question to ask whether the largest cycle of the random permutation is also giant in the supercritical regime.

*Conjecture.* Let  $L_1(t)$  be the size of the largest cycle at time  $t$ . If  $c > 1$  then

$$\frac{L_1(cn/2)}{\theta(c)n} \Rightarrow V$$

where  $V$  is a random variable with  $0 < V \leq 1$  a.s.

This problem is quite different from our original one. However our techniques enable us to prove a partial result in this direction as a corollary of the proof of Theorem 4.

**Theorem 5.** *For any  $c > 1$ , at time  $cn/2$  there are at least  $\theta(c)n - o(n)$  vertices located on large cycles (i.e., of size greater than or equal to  $n^a$ , for any  $a < 2/3$ ).*

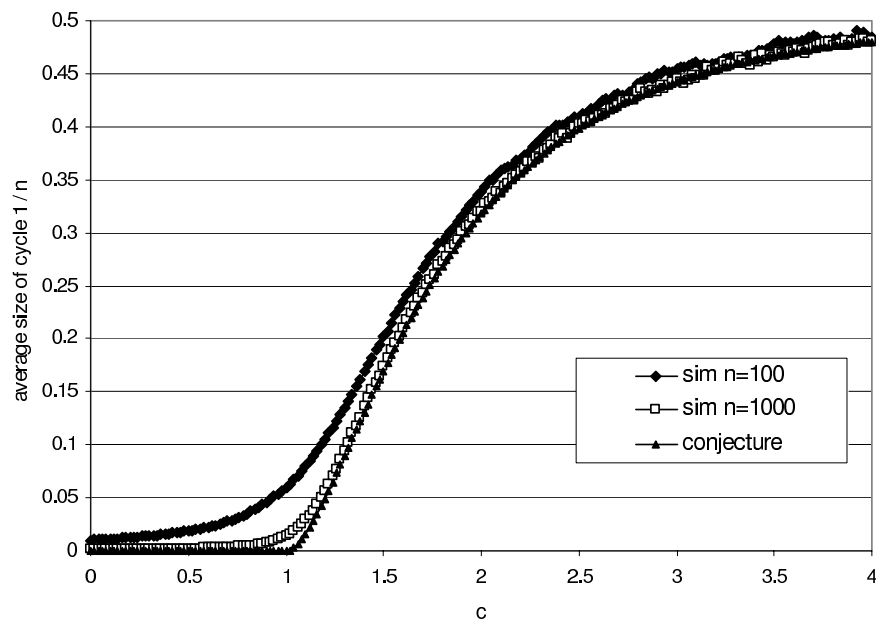
David Aldous (private communication) conjectures that the relative sizes of the pieces of the giant cycle are in equilibrium at all times in the supercritical regime, i.e., have the Poisson-Dirichlet  $PD(0, 1)$  distribution, which gives the limiting behavior of the ordered sizes of cycles in a uniform random permutation. According to this conjecture,  $V$  would be distributed as the first coordinate of a  $PD(0, 1)$  random variable. One way to approach this conjecture would be to generalize Aldous (1997) to show that the large cycles in the critical regime converge to a coagulation-fragmentation process and to study the growth of clusters in that process.

Alternatively, one could look at the size of the cycle containing 1,  $K_1(t)$ , and try to show that

$$\frac{K_1(cn/2)}{\theta(c)n} \Rightarrow U$$

where  $U$  has a point mass of size  $1 - \theta(c)$  at 0 and is otherwise uniform on  $(0, \theta(c))$ . Figure 4 shows the average growth of  $K_1(cn/2)/n$  in 10,000 simulations of  $n = 100$ ,  $n = 1000$ , and compares the results to  $EU = \theta(c)^2/2$ . Although this considers only one aspect of the distribution of large cycles, it agrees well with Aldous’ conjecture.

Figure 5 shows a histogram of the result of 100,000 simulations of  $K_1(100)$  when  $n = 100$ . As the graph shows, the spike in the frequency of clusters of size

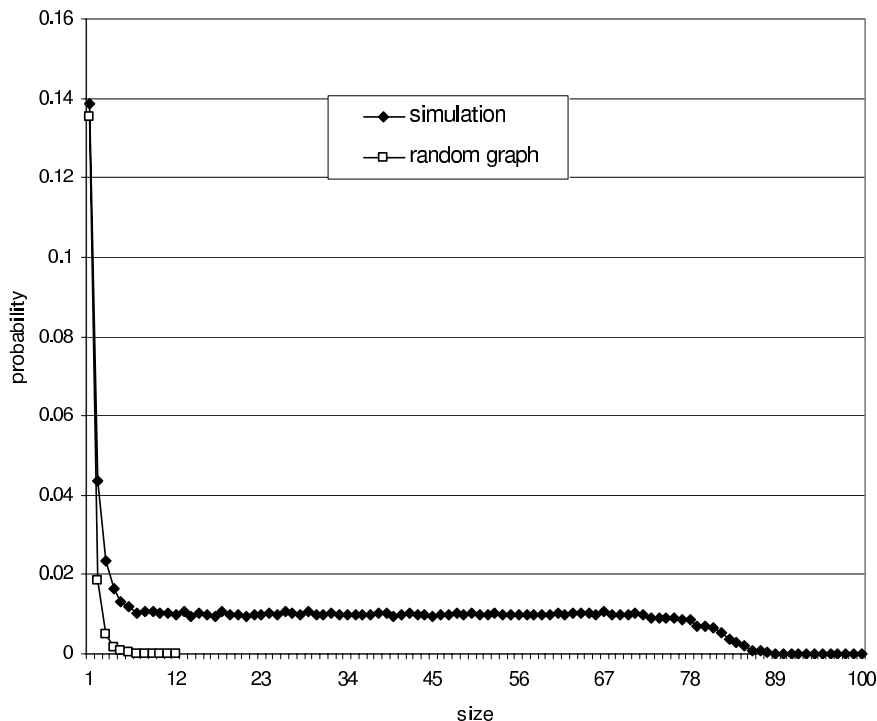


**Fig. 4.** Growth of the average fraction of vertices in the cycle containing 1 at time  $cn/2$  in 10,000 simulations of the random transposition chain with  $n = 100$  and  $n = 1000$  compared to  $\theta(c)^2/2$  ( $\theta(c)$  being the percolation probability of the corresponding random graph with  $p = c/n$ )

4 or smaller is what one would predict from the random graph cluster size distribution. The remainder of the distribution is roughly uniform except for rounding at the upper end. The latter is to be expected if Aldous' conjecture is correct, since the size of the giant component satisfies the central limit theorem.

As we were finishing this paper, we learned that Oded Schramm (2004) has proved David Aldous' conjecture.

*Remark.* The problem of the emergence of a giant cycle is closely related to Angel's (2003) work on the existence of infinite orbits for the *random stirring process*, which is the random transposition random walk on an infinite graph such as  $\mathbb{Z}^d$  or a tree, rather than the complete graph on  $\{1, \dots, n\}$  considered in this work. To explain the connection, suppose that we construct our process using a Poisson process with rate  $2/n^2$  for each  $i \neq j$ , and at these times draw an edge between  $i$  and  $j$  to indicate that  $i$  and  $j$  are to be transposed. To compute the cycles in the permutation at time  $cn/2$ , we repeat the first  $[0, cn/2]$  units of time periodically and then observe the sites that a walker starting at  $i$  visits at times  $kcn/2$ , for  $k = 1, 2, \dots$ . Angel (2003) calls this construction the *cyclic time random walk*. Its relevance to his work is that the cyclic time random walk is transient if, and only if, the cycles are infinite.



**Fig. 5.** Histogram of the size of the cycle containing 1 in 100,000 simulations of the random transposition chain with  $n = 100$  at time 100 ( $c = 2$ ). The open squares give the distribution of the size of finite clusters in the corresponding random graph

#### 4. The subcritical regime

Let us introduce some notations for the different probability laws involved. For each  $n$ , we have the coagulation-fragmentation process, and the Erdős-Renyi random graph model. To emphasize when computations are being done for the random graph we will use  $Q_p$ , for the random graph with Bernoulli percolation parameter  $p$ , and  $Q$  for the law of the evolving random graph that at time  $s$  has  $p_s = 1 - \exp(-2s/n^2)$ . When  $s = cn/2$  this probability is  $p(c, n) = 1 - \exp(-c/n) \leq c/n$ . To simplify notation we will use  $QX$  to denote the expected value of  $X$  with respect to the probability  $Q$ .

##### 4.1. Preliminary results : comparison with a branching process

Our first result provides a useful upper bound.

**Lemma 1.** *The cluster size  $|\mathcal{C}_1|$  in  $Q_{c/n}$  is dominated by  $Z$ , the total progeny of a branching process in which each individual has a Binomial( $n - 1, c/n$ ) number of children, i.e., we can construct these random variables on the same probability space so that  $|\mathcal{C}_1| \leq Z$  a.s. It follows from this that if  $c < 1$  then  $Q_{c/n}|\mathcal{C}_1| \leq EZ = 1/(1 - c)$ .*

*Proof.* This result and the one below (Lemma 2) are well-known. See for instance Ball (1983), Section 4, where earlier references are also given. However, as the proofs are elementary and help build the intuition about the problem, we prefer to include them directly here.

Intuitively, Lemma 1 holds since a vertex in generation  $k$  may have children among all of the  $n$  vertices of the graph except those of the first  $k$  generations. To begin to prove this formally, let  $\xi_{i,j}$ ,  $1 \leq i, j \leq n$  be independent random variables, taking values 1 with probability  $c/n$  and 0 with probability  $1 - c/n$ . To start the random graph let  $Y_0 = \{1\}$  and let  $Y_1 = \{j \notin Y_0 : \xi_{1,j} = 1\}$ . To start the branching process let  $Z_0 = 1$ ,  $Z_1 = |Y_1|$ , and let  $\phi_1 : Y_1 \rightarrow \{1, 2, \dots, Z_1\}$  be 1-1 and onto.

If the first  $k$  stages of the construction have been done and we have  $Y_k \neq \emptyset$  and a  $\phi_k : Y_k \rightarrow \{1, \dots, Z_k\}$  that is 1-1 (but not onto in general), then let

$$Y_{k+1} = \cup_{i \in Y_k} \{j \notin \cup_{\ell=0}^k Y_\ell : \xi_{i,j} = 1\}$$

We let individual  $\phi_k(i)$  in the  $k$ th generation of the branching process have  $|\{j \neq i : \xi_{i,j} = 1\}|$  children. The individuals in the branching process that are not in  $\phi_k(Y_k)$  have a number of children given by independent binomials. It should be clear from the construction that can again define  $\phi_{k+1} : Y_{k+1} \rightarrow \{1, \dots, Z_{k+1}\}$  to be 1-1, and the comparison follows by induction. The inequality follows by computing  $EZ$  (for instance by summing a geometric series).  $\square$

The next result shows that the bound in Lemma 1 is exact in the limit. Let  $\{Z_k\}_{k=0}^\infty$  be a Poisson Galton-Watson process with offspring mean  $c$  and let  $Z = \sum_{k=0}^\infty Z_k$  be its total progeny.

**Lemma 2.** *Let  $\mathcal{C}_1$  be the cluster that contains vertex 1. If  $0 \leq c < 1$  then as  $n \rightarrow \infty$*

$$Q_{p(c,n)}(|\mathcal{C}_1| = k) \rightarrow P(Z = k)$$

*Proof.* The number of children of vertex 1,  $Z_n^1 = |Y_1|$  has distribution Binomial( $n-1$ ,  $p(c, n)$ ), which converges to a Poisson( $c$ ) limit. Let  $k \geq 1$  and let  $(n_1, \dots, n_{k+1}) \in \mathbb{N}^{k+1}$ . If we let  $Z_j^n = |Y_j|$  then

$$Q_{p(c,n)}(Z_{k+1}^n = n_{k+1} | Z_1^n = n_1, \dots, Z_k^n = n_k) = P\left(\sum_{i=1}^{n_k} B_i^n = n_{k+1}\right)$$

where  $B_i^n$  are i.i.d. Binomial( $n - s$ ,  $p(c, n)$ ) random variables, and  $s = \sum_{i=0}^k n_k$  with  $n_0 = 1$ . From this it follows easily that the convergence of finite-dimensional distributions of  $\{Z_j^n\}_{j \geq 1}$  to those of  $PGW(c)$ . Markov's inequality and the domination result in Lemma 1 imply that

$$Q_{p(c,n)}\left(\sum_{k=K}^\infty Z_k^n > 0\right) \leq Q_{p(c,n)}\left(\sum_{k=K}^\infty Z_k^n\right) \leq c^K / (1 - c)$$

and the desired conclusion follows.  $\square$

Our next ingredient is

**Lemma 3.**  $Q_{c/n}(|\mathcal{C}_1| \geq y) \leq c^{-1} \exp(-(c-1-\ln c)y)$ .

*Proof.* In view of Lemma 1, it suffices to prove the result for  $Z$ , rather than  $|\mathcal{C}_1|$ . To do this, let

$$\begin{aligned} \phi_n(\theta) &= e^{-\theta} \sum_{m=0}^{n-1} \binom{n-1}{m} \left(\frac{c}{n}\right)^m \left(1 - \frac{c}{n}\right)^{n-1-m} e^{\theta m} \\ &= e^{-\theta} \left(1 - \frac{c}{n} + \frac{c}{n} e^{\theta}\right)^{n-1} \end{aligned}$$

be the moment generating function of the distribution of the number offspring minus 1. Let  $S_m$  be a random walk that takes steps with this distribution and  $S_0 = 1$ , so that  $S_m$  explores the Galton-Watson tree. Then  $\tau = \inf\{m : S_m = 0\}$  has the same distribution as  $Z$ . Let  $R_m = \exp(\theta S_m) / \phi_n(\theta)^m$ .  $R_m$  is a nonnegative martingale. Stopping at time  $\tau$  we have  $e^{\theta} \geq E(\phi_n(\theta)^{-\tau})$ . If  $\phi_n(\theta) < 1$  it follows that

$$P(\tau \geq y) \phi_n(\theta)^{-y} \leq E[\phi_n(\theta)^{-\tau}] \leq e^{\theta}$$

Using  $\phi_n(\theta) \leq e^{-\theta} \exp(c(e^{\theta} - 1))$  now we have

$$P(\tau \geq y) \leq e^{\theta} \left(e^{-\theta} \exp(c(e^{\theta} - 1))\right)^y$$

To optimize the bound we want to minimize  $c(e^{\theta} - 1) - \theta$ . Differentiating this means that we want  $ce^{\theta} - 1 = 0$  or  $\theta = -\log(c)$ . Plugging this and recalling that  $\tau$  and  $Z$  have the same distribution we have

$$P(Z \geq y) \leq \frac{1}{c} \exp(-(c-1-\ln c)y)$$

It follows that

$$Q_{c/n}(|\mathcal{C}_1| \geq y) \leq \frac{1}{c} \exp(-(c-1-\ln c)y)$$

which completes the proof of Lemma 3.  $\square$

Now recall that for  $c < 1$ ,  $Z_c = \sum_{s \leq cn/2} \mathbf{1}_{\{F_s\}}$  is the number of fragmentations up to time  $cn/2$ . The rate at which fragmentations occur is  $\sum_k f_k(t)(k-1)/n$  where  $f_k(t)$  is the fraction of vertices that belong to cycles of size  $k$ . Therefore the expected number of fragmentations is

$$EZ_c = \int_0^{cn/2} \sum_{k=1}^n E f_k(t) \frac{k-1}{n} dt$$

*Remark.* Note that this formula takes into account the case of multiple edges in the random graph, i.e. the possibility that a given transposition may be chosen twice as an increment of the random walk before time  $cn/2$ .

**Lemma 4.** *Let  $f_k(s)$  be the empirical fraction of vertices in cycles of size  $k$  at time  $s$ . If  $0 \leq c < 1$  then  $E f_k(cn/2) \rightarrow P(Z = k)$  and  $E Z_c \rightarrow \kappa(c)$ , where  $\kappa(c)$  was defined in Theorem 1.*

*Proof.* The cycle sizes at time  $s$  in the coagulation-fragmentation process are dominated by the cluster sizes in the random graph model with  $p_s = 1 - \exp(-2s/n^2) \leq 2s/n^2$ . Note that by exchangeability the expected fraction of vertices in clusters of the random graph of size  $k$  satisfies  $E(f_k) = P(|\mathcal{C}_1| = k)$ . (This also holds under  $Q$ , and we will use it below). By Lemma 3,

$$\begin{aligned} P(|\mathcal{C}_1| \text{ fragments before time } cn/2) &\leq E(\#\text{such fragmentations}) \\ &\leq \int_0^{cn/2} E\left(\frac{|\mathcal{C}_1|^2}{n^2}\right) dt \\ &\leq \int_0^{cn/2} Q_{2t/n^2}\left(\frac{|\mathcal{C}_1|^2}{n^2}\right) dt \rightarrow 0 \end{aligned}$$

Therefore  $P(|\mathcal{C}_1| = k)$  has the same asymptotics as  $Q(|\mathcal{C}_1| = k)$  and we can conclude for the first convergence by Lemma 2.

For the second convergence, first note that

$$E Z_c \leq \int_0^{cn/2} \sum_{k=1}^n Q f_k(s) \frac{k-1}{n} ds \leq \int_0^{cn/2} Q_{2s/n^2} \left( \frac{|\mathcal{C}_1| - 1}{n} \right) ds$$

Using Lemma 1  $Q_{2s/n^2} |\mathcal{C}_1| \leq 1/(1 - (2s/n))$ . Changing variables  $un/2 = s$  we have

$$E Z_c \leq -\frac{1}{2}(\log(1 - c) + c) = \kappa(c) \quad (9)$$

For the lower bound we use Fatou's lemma (twice) and the first convergence (recall also that if  $Z(s)$  is the total progeny of a  $PGW(s)$  then  $E Z(s) = 1/(1 - s)$ ):

$$\begin{aligned} \liminf_{n \rightarrow \infty} E Z_c &= \liminf_{n \rightarrow \infty} \int_0^{cn/2} \sum_{k=1}^n E f_k(t) \frac{k-1}{n} dt \\ &\geq \frac{1}{2} \int_0^c \liminf_{n \rightarrow \infty} \sum_{k=1}^n k E(f_k(sn/2)) - 1 ds \\ &\geq \frac{1}{2} \int_0^c \sum_{k=1}^{\infty} k P(Z(s) = k) - 1 ds = \frac{1}{2} \int_0^c \frac{1}{1-s} - 1 ds \\ &\geq \kappa(c) \end{aligned}$$

□

The final preparatory step is:

**Lemma 5.** *If  $c < 1$  the expected number of fragmentations that occur to cycles that have already been fragmented is  $\leq K_c(\log n)^2/n$ , and  $K_c = 9c\kappa(c)\alpha(c)^{-2}$ . (Recall  $\alpha(c) = (c - 1 - \log c)$ ).*

*Proof.* The expected number of such fragmentations is at most:

$$\begin{aligned} &\leq E \int_0^{cn/2} \frac{\#\text{vertices in fragments}}{n} \frac{L_1(bn/2)}{n} dt \\ &\quad \frac{n}{2} \int_0^c EZ_b \left( \frac{L_1(bn/2)}{n} \right)^2 db \end{aligned}$$

where  $L_1(t)$  is the size of the largest component at time  $t$ . In the event that  $L_1(cn/2) \leq 3\alpha(c)^{-1} \log n$ , the above is at most

$$(n/2)(3\alpha(c)^{-1} \log n/n)^2 \int_0^c \kappa(b) db \leq \frac{1}{2} K_c \frac{(\log n)^2}{n}$$

On the other hand by Lemma 3 the complement of this event has probability at most  $n^{-2}$ , and there can never be more than  $cn/2$  such fragmentations, so Lemma 5 is proved.  $\square$

#### 4.2. Proof of Theorem 1

We are now ready to prove Theorem 1. Let  $\bar{Z}_c^n = \sum_{s \leq cn/2} \mathbf{1}_{\{\bar{F}_s\}}$ ,  $0 \leq c < 1$  be the counting process of fragmentations that occur to cycles which (a) have not been fragmented previously and (b) have size  $\leq n^{0.7}$ . The second condition is irrelevant in this section, but imposing it now will help in the next one. Unfragmented cycles correspond to trees in the random graph (as explained in the introduction, by tree we mean a connected components with no closed circuit, including multiple edges), so the compensator of  $\bar{Z}_c^n$  is

$$\bar{\kappa}^n(c) = \int_0^{cn/2} \bar{\psi}_s^n ds \tag{10}$$

where  $\bar{\psi}_s^n = \sum_{k=1}^{n^{0.7}} \bar{f}_k(s)(k-1)/n$  and  $\bar{f}_k(s)$  is the fraction of vertices that belong to tree components of size  $k$ . As noted in the sketch of the proof, it is enough to show that for each fixed  $c$ ,  $\bar{\kappa}^n(c)$  converges in probability to  $\kappa(c)$ , or, by Lemma 5, that  $\bar{\kappa}^n(c)$  converges to  $\kappa(c)$  in probability. Lemmas 3 and 4 imply that  $E[\int_0^{cn/2} \bar{\psi}_s^n ds] \rightarrow \kappa(c)$ . It remains to show that  $\text{var} \int_0^{cn/2} \bar{\psi}_s^n ds \rightarrow 0$ . Our first step will be to prove :

$$\text{var}(\bar{\psi}_s^n) \leq \frac{K}{n^3} \mathcal{Q}_{p(c,n)}[|\mathcal{C}_1|^3] \tag{11}$$

for all time  $s \leq cn/2$ , where  $K$  is a constant that depends only on  $c$ .

To see this, first observe that in terms of cluster sizes

$$\bar{\psi}_s^n = \frac{1}{n^2} \sum_{i=1}^n (|\mathcal{C}_i| - 1) I_i$$

where  $I_i$  is the indicator of the event that  $\mathcal{C}_i$  is a tree. Let  $d_i = (|\mathcal{C}_i| - 1) I_i$ .

$$\text{var} \frac{1}{n^2} (d_1 + \dots + d_n) = \frac{1}{n^4} (n \text{var}(d_1) + n(n-1) \text{cov}(d_1, d_2)) \quad (12)$$

Monotonicity and Lemma 3 imply,

$$\text{var}(d_1) \leq \mathcal{Q}_{p(c,n)}[|\mathcal{C}_1|^2] \leq K \quad (13)$$

It remains to bound  $\text{cov}(d_1, d_2)$ . If we let

$$\pi_{n,i} = i^{i-2} (p_s)^{i-1} (1-p_s)^{i(n-i) + \binom{i}{2} - (i-1)}$$

where here  $p_s = \exp(-2s/n^2) 2s/n^2$  (this is the probability that a given edge appears exactly once, since we don't want any multiple edges), then by the reasoning for (4) we have

$$\begin{aligned} & \mathcal{Q}_{p_s}[\mathcal{C}_1 \cap \mathcal{C}_2 = \emptyset, |\mathcal{C}_1| = j, |\mathcal{C}_2| = k, \mathcal{C}_1 \text{ and } \mathcal{C}_2 \text{ are trees}] \\ &= \binom{n-2}{j-1} \pi_{n,j} \binom{n-j-1}{k-1} \pi_{n-j,k} \\ & \mathcal{Q}_{p_s}[\mathcal{C}_1 = \mathcal{C}_2, |\mathcal{C}_1| = k, \mathcal{C}_1 \text{ is a tree}] = \binom{n-2}{k-2} \pi_{n,k} \end{aligned}$$

From this it follows that  $\text{cov}(d_1, d_2)$

$$\begin{aligned} &= \sum_{j,k} \left[ \binom{n-2}{j-1} \binom{n-j-1}{k-1} (1-p_s)^{-j} \right. \\ & \quad \left. - \binom{n-1}{j-1} \binom{n-1}{k-1} \right] (j-1)(k-1) \pi_{n,j} \pi_{n,k} \\ & \quad + \sum_k \binom{n-2}{k-2} (k-1)^2 \pi_{n,k} \end{aligned}$$

For the first term in the right-hand side,

$$\begin{aligned} & \left[ \binom{n-2}{j-1} \binom{n-j-1}{k-1} (1-p_s)^{-j} - \binom{n-1}{j-1} \binom{n-1}{k-1} \right] \\ & \leq \frac{(n-2)! e^{2c}}{(j-1)!(k-1)!(n-j-k)!} - \frac{(n-1)!}{(j-1)!(n-j)!} \frac{(n-1)!}{(k-1)!(n-k)!} \\ & \leq 0 \end{aligned}$$

where we have used for the first inequality that  $-\log(1-x) \leq 2x$  if  $|x|$  is small enough, and for the second one,  $(n-2)! e^{2c} \leq (n-1)!$  for large  $n$  and

$$(n-j)!/(n-j-k)! \leq (n-1)!/(n-1-k)!.$$

For the second term,

$$\sum_k \binom{n-2}{k-2} (k-1)^2 \pi_{n,k} \leq \frac{1}{n-1} \sum_k k^3 \binom{n-1}{k-1} \pi_{n,k} \leq \frac{1}{n-1} \mathcal{Q}_{p(c,n)}[|\mathcal{C}_1|^3]$$

Combining this with (12) and (13) gives (11).

Hence by the Cauchy-Schwarz inequality we get:

$$\begin{aligned} \text{var} \left( \int_0^t \bar{\psi}_s^n ds \right) &= \mathcal{Q} \left[ \left( \int_0^t (\bar{\psi}_s^n - \mathcal{Q}[\bar{\psi}_s^n]) ds \right)^2 \right] \leq t \int_0^t \text{var}(\bar{\psi}_s^n) ds \\ &\leq \frac{cn}{2} \int_0^{cn/2} \frac{K}{n^3} ds = \frac{c^2 K}{4n} \rightarrow 0 \end{aligned} \quad (14)$$

where we have used both (11) and Lemma 3. This completes the proof of Theorem 1.

## 5. The critical regime

The first step in the proof of Theorem 2 is to argue that fragmentations of previously fragmented cycles can be ignored. The number of such fragmentations is smaller than the total number of cycles in multicyclic components (i.e., components with at least 2 cycles) in the random graph. Theorem 1 and Corollary 3 in Luczak, Pittel, and Wierman (1994) imply that the total number of cycles in multicyclic components in the critical regime is bounded in probability.<sup>1</sup> In particular, divided by  $(\log n)^{1/2}$  it converges to 0 in probability. As a result, by the converging together lemma (see e.g., Durrett (1996), Chap. 2, Ex. 2.10), it suffices to prove the central limit theorem for the number of fragmentations on tree components.

As in the previous section, we will in addition restrict our attention to fragmentations of tree components of size at most  $n^{0.7}$ , and continue to use the notation introduced there. (Indeed, classical results from the theory of random graphs, or Aldous (1997), show that asymptotically almost surely all clusters are smaller than  $n^{0.7}$ ).

Let  $\bar{W}_n(r) := (6/\log n)^{1/2}(\bar{Z}^n(r) - \bar{\kappa}^n(r))$ . By standard methodology in the theory of stochastic processes (see Jacod and Shiryaev (1987) or Revuz and Yor (1999) for instance), to prove convergence of  $\bar{W}_n(\cdot)$  to Brownian Motion, the two things we need to check are: (i)  $E[\sup_{0 \leq r \leq 1} |\bar{W}_n(r) - \bar{W}_n(r^-)|] \rightarrow 0$  and (ii) The quadratic variation of  $\bar{W}_n$ , i.e. the increasing process associated with  $\bar{W}_n(\cdot)^2$ , must converge to  $r$  at time  $r$ . (i) is obvious because  $\bar{Z}^n$  is a counting process, and (ii) turns into  $E(6\bar{\kappa}^n(r)/\log n) \rightarrow r$  and  $\text{var}(6\bar{\kappa}^n(r)/\log n) \rightarrow 0$ . These two steps are dealt with respectively in Lemmas 7 and 8.

But first, we need a technical lemma that will be useful on several occasions (e.g., for computing precise asymptotics of the number of trees of a given size).

<sup>1</sup> This result could also be derived from the Folk Theorem 1 in Aldous (1997) which gives the limit for the joint distribution of the component sizes and the number of cycles they contain. See the discussion page 850 of his paper.

**Lemma 6.** *If  $k \rightarrow \infty$  and  $k = o(n^{3/4})$  then*

$$\begin{aligned} \gamma_{n,k}(c) &\equiv \binom{n}{k} k^{k-2} \left(\frac{c}{n}\right)^{k-1} \left(1 - \frac{c}{n}\right)^{kn-k^2/2-3k/2+1} \\ &\sim \frac{nk^{-5/2}}{c\sqrt{2\pi}} \exp\left(-\alpha(c)k + (c-1)\frac{k^2}{2n} - \frac{k^3}{3n^2}\right) \equiv \lambda_{n,k}(c) \end{aligned}$$

where  $\alpha(c) = c - 1 - \log(c)$ . There is a constant  $K$  so that if  $1 \leq k \leq n^{0.7}$  and  $c \leq 1$  then  $\gamma_{n,k}(c) \leq K\lambda_{n,k}(c)$ .

*Proof.* Stirling's formula implies  $k! \sim k^k e^{-k} \sqrt{2\pi k}$ . Using this we have that

$$\gamma_{n,k}(c) \sim \frac{nk^{-5/2}}{c\sqrt{2\pi}} \left[ \prod_{j=1}^{k-1} \left(1 - \frac{j}{n}\right) \right] e^k c^k \left(1 - \frac{c}{n}\right)^{kn-k^2/2-3k/2+1}$$

Using the expansion  $\log(1-x) = -x - x^2/2 - x^3/3 - \dots$  we see that if  $k = o(n)$  then

$$\left(1 - \frac{c}{n}\right)^{kn-k^2/2-3k/2+1} \sim \exp(-ck + k^2/2n)$$

while if  $k = o(n^{3/4})$  we have

$$\begin{aligned} \prod_{j=1}^{k-1} \left(1 - \frac{j}{n}\right) &= \exp\left(-\frac{1}{n} \sum_{j=1}^{k-1} j - \frac{1}{n^2} \sum_{j=1}^{k-1} j^2 + O\left(\frac{j^4}{n^3}\right)\right) \\ &\sim \exp\left(-\frac{k(k-1)}{2n} - \frac{k(k-1)(2k-1)}{6n^2}\right) \sim \exp\left(-\frac{k^2}{2n} - \frac{k^3}{3n^2}\right) \end{aligned}$$

Combining the last three formulas gives the asymptotic formula. To prove the bound we note that Stirling's formula implies  $k! \geq \delta k^k e^{-k} \sqrt{2\pi k}$  for some  $\delta > 0$ . Using the bounds  $\log(1-x) \leq -x$  and  $\log(1-x) \leq -x - x^2/2$  in the last two calculations gives the upper bound.  $\square$

**Lemma 7.**

$$E \left[ \frac{6}{\log n} \int_0^{c_n(r)n/2} \bar{\psi}_s^n ds \right] \rightarrow r$$

*Proof.* The upper bound follows from (9) which holds for all  $c < 1$ . In the other direction, we will use Fatou's lemma so it is enough to know the asymptotic behavior of the integrand. Changing variables  $s = c_n(v)n/2$  where  $c_n(v) = 1 - n^{-v/3}$  and noting  $c'_n(v) = (1/3)(\log n)n^{-v/3}$  gives

$$\begin{aligned} E \left[ \frac{6}{\log n} \bar{Z}^n(r) \right] &= n \int_0^r Q[\bar{\psi}_{c_n(v)n/2}^n] n^{-v/3} dv \\ &= \int_0^r \sum_{k=1}^{n^{0.7}} \frac{k-1}{n} Q_{p(c_n(v),n)}[kT_k] n^{-v/3} dv \end{aligned} \quad (15)$$

where  $T_k$  is the number of tree components of size  $k$ , and  $p(c, n) = 1 - \exp(-c/n)$ .

We can take the limit of the last expression by using formula (4), combined with Lemma 6. Indeed formula (4) shows that  $ET_k = \gamma_{n,k}(c)$ , and  $k \leq n^{0.7} = o(n^{3/4})$ , so that the use of Lemma 6 is justified. Hence

$$E \left[ \frac{6}{\log n} \bar{Z}^n(r) \right] = \int_0^r \sum_{k=1}^{n^{0.7}} \frac{k(k-1)}{n} \gamma_{n,k}(c_n(v)) n^{-v/3} dv$$

Setting  $c = 1 - b$  with  $b = n^{-v/3} \rightarrow 0$  and using Taylor's theorem

$$-(c-1-\log(c))k - b \frac{k^2}{n} = -\frac{b^2}{2}k - b \frac{k^2}{2n} + o(b^2k) \quad (16)$$

The first term becomes significantly negative when  $k \approx 1/b^2 = n^{2v/3}$ , the second when  $k \approx \sqrt{n/b} = n^{(1+v)/3}$ . Let  $2v/3 < w < (1+v)/3$ . The terms with  $k \geq n^w$  tend to 0 exponentially fast in  $n$  and there are less than  $n^{0.7}$  of them so their sum goes to 0. In the range  $k \leq n^w$  the second term can be ignored, so Lemma 6 implies that if  $v < 1$

$$\sum_{k=1}^{n^{0.7}} \frac{k^2}{n} Q_{p(c_n(v),n)}[T_k] \sim \frac{1}{\sqrt{2\pi}} \sum_{k=1}^{n^w} k^{-1/2} \exp(-n^{-2v/3}k/2) \quad (17)$$

Here we have used the asymptotic formula of Lemma 6 for all  $k$ . However, the next computation will show that the sum grows like  $n^{v/3}$  so the contributions from small  $k$  can be ignored.

To recognize (17) as a Riemann sum with spacing  $n^{-2v/3}$ , we rewrite it as

$$n^{v/3} \sum_{k=1}^{n^{0.7}} n^{-2v/3} (n^{-2v/3}k)^{-1/2} \exp(-n^{-2v/3}k/2)$$

Since  $x^{-1/2}e^{-x}$  is decreasing it is straightforward to estimate the difference between the sum and the limiting integral and we conclude that

$$n^{-v/3} \sum_{k=1}^{n^{0.7}} \frac{k(k-1)}{n} Q_{p(c_n(v),n)}[T_k] \rightarrow \int_0^\infty \frac{x^{-1/2}}{\sqrt{2\pi}} e^{-x/2} dx$$

Changing variables  $x = y^2$ ,  $dx = 2y dy$  the integral becomes  $(2\pi)^{-1/2} \int_0^\infty 2e^{-y^2/2} dy = 1$ . Therefore, by Fatou's lemma:

$$\liminf_{n \rightarrow \infty} E \left[ \frac{6}{\log n} \bar{Z}^n(r) \right] \geq \int_0^r n^{-v/3} \cdot n^{v/3} dv = r$$

□

We turn now to the analysis of the variance.

**Lemma 8.**  $\text{var} \left( \frac{6}{\log n} \bar{\kappa}_n(r) \right) \rightarrow 0$

*Proof.* Changing variables as in (15) and using Cauchy-Schwarz inequality as in (14),

$$\begin{aligned} \text{var} \left( \frac{6}{\log n} \int_0^{c_n(r)n/2} \bar{\psi}_s^n ds \right) &= \text{var} \left( n \int_0^r \bar{\psi}_{c_n(v)n/2}^n n^{-2v/3} dv \right) \\ &\leq n^2 \int_0^r n^{-2v/3} \text{var}(\bar{\psi}_{c_n(v)n/2}^n) dv \\ &\leq \frac{2}{n} \int_0^r n^{-2v/3} Q_{p(c_n(v),n)}[|C_1|^3 I_1] dv \end{aligned}$$

Reasoning as in (17) but using the bound in Lemma 6

$$\sum_{k=1}^n k^3 Q_{p(c_n(v),n)}[kT_k] \leq K \sum_{k=1}^n k^{3/2} \exp(-n^{-2v/3}k/2)$$

To check the right-hand side note that the power of  $k$  has increased by 2, from the previous calculation. If we view the last sum as a Riemann sum with spacing  $n^{-2v/3}$ , we can rewrite it as

$$n^{5v/3} \sum_{k=1}^n n^{-2v/3} (n^{-2v/3}k)^{3/2} \exp(-n^{-2v/3}k/2)$$

Now  $x^{3/2}e^{-x/2}$  has derivative  $((3/2)x^{1/2} - (1/2)x^{3/2})e^{-x/2}$  so it is increasing on  $[0, 3]$  and then decreasing on  $[3, \infty)$ . Thus if we discard the term with the largest  $k$  so that  $n^{-2v/3}k \leq 3$  we have a lower bound on the integral.

$$n^{-2v/3} \sum_{k=1}^n k^3 Q_{p(c_n(v),n)}[kT_k] \leq n^v \frac{1}{\sqrt{2\pi}} \int_0^\infty x^{3/2} e^{-x/2} dx + n^{v/3} 3^{3/2} e^{-3/2}$$

Using this it follows that

$$\text{var} \left( \frac{6}{\log n} \bar{\kappa}^n(r) \right) \leq \frac{K}{n} \int_0^r n^u du$$

Writing  $n^u = \exp(-u \log n)$  and integrating we have that the right-hand side is  $\leq K/(\log n) \rightarrow 0$ . This concludes the proof of the first result in Theorem 2.  $\square$

The final step is to estimate the number of fragmentations that occur to tree components of size  $\leq n^{0.7}$  at times between  $(1 - n^{-1/3})n/2$  and  $n/2$ :

$$\int_{(1-n^{-1/3})n/2}^{n/2} Q \bar{\psi}_s^n ds$$

For each  $s$  in the interval the integrand is smaller than  $\sum_{k=1}^{n^{0.7}} \frac{k^2}{n} Q_{1/n} T_k$ . Using Lemma 6, the last quantity is smaller than

$$\frac{K}{n} \sum_{k=1}^{\infty} k^{-1/2} \exp(-k^3/3n^2)$$

which we can rewrite as

$$\frac{K}{n} n^{1/3} \sum_{k=1}^{\infty} n^{-2/3} (kn^{-2/3})^{-1/2} \exp(-(kn^{-2/3})^3/2)$$

The above sum is a Riemann sum so it converges to  $\int_0^{\infty} x^{-1/2} e^{-x^3/2} dx$ . Therefore,  $Q\bar{\psi}_s^n \leq Kn^{-2/3}$ . Since the duration of the critical regime is  $n^{2/3}/2$ , the expected number of fragmentations is bounded and the proof of Theorem 2 is complete.

## 6. The supercritical regime

By Pittel's (1990) central limit theorem for the number of components of a supercritical random graph, it is enough to show that, with probability going to 1 as  $n \rightarrow \infty$ , at time  $cn/2$  there are fewer than  $o(n^{1/2})$  extra components due to fragmentation. (This was already indicated in the sketch of the proof of Theorem 4).

Let  $a = 0.55$ . (In fact the results stated in this section would also be valid for any  $1/2 < a < 2/3$  but making this choice makes some proofs slightly easier). We call cycles of size  $k \geq n^a$  large. These can be ignored since there cannot be more than  $n^{1-a} = o(n^{1/2})$  such components. We define the amount of mass "upstairs" by

$$N_t^{\uparrow} = \sum_{k > n^a} k X_k(t)$$

where  $X_k(t)$  is the number of cycles of size  $k$  at time  $n/2 + t$ . (It is convenient in this section to shift the time so that  $t = 0$  corresponds to critical time  $n/2$ .) If all of the mass was upstairs, then the expected number of cycles of size less than  $n^a$  produced by fragmentation would be  $2n^{a-1}(cn/2) = O(n^a)$ . It is overly pessimistic to think that all of the mass will be upstairs, but by analogy with the random graph, we expect (and will eventually prove in Theorem 5) that at times  $c > 1$  a positive fraction of the total mass  $n$  will be there, so this estimate of the number of fragmentations is too large to ignore.

To improve this crude estimate, we take advantage of the fact that fragmented pieces are reabsorbed upstairs. Let  $X_k^{\downarrow}(t)$  be the number of cycles of size  $k$  produced by fragmentation of cycles upstairs.  $X_k^{\downarrow}(t)$  can only increase when a transposition is performed, and only if it is made of one of the  $N_t^{\uparrow}$  vertices upstairs and of one of the 2 points located  $k$  steps away when writing the corresponding cycle of the current permutation. This gives a rate at most  $2N_t^{\uparrow}/n^2$ . As for the death rate, one way to get rid of a component of size  $k$  is by picking one of the  $k$  vertices of one of the  $X_k^{\downarrow}(t)$  components and one of the  $N_t^{\uparrow}$  vertices upstairs. This happens with rate  $2kX_k^{\downarrow}(t)N_t^{\uparrow}/n^2$ . For the moment we are ignoring the fact that cycles may experience coalescence or fragmentation while downstairs. We will deal with these complexities once we have an understanding of the basic birth and death process of fragments of large clusters.

### 6.1. The cluster queuing system

It is fortunate that the unknown quantity  $N_t^\uparrow \leq n$  appears in both rates, so that as long as  $N_t^\uparrow > 0$  we can remove it by time change. Once this is done, we have a system of stochastic processes  $\xi_t^k$ , for  $1 \leq k \leq n^a$  that we call a *cluster queuing system*: let  $\xi_t^k$  be independent birth-and-death chains with birth rate 1 and death rate  $k\xi_t^k$ , that begin with  $\xi_0^k = 0$ .

**Lemma 9.** *With probability  $\rightarrow 1$  as  $n \rightarrow \infty$  we have*

$$\sum_{k=1}^{n^a} \xi_t^k \leq (\log n)^2 \quad \text{and} \quad \sum_{k=1}^{n^a} k\xi_t^k \leq n^a (\log n)^2$$

for all  $t \leq c$  ( $c > 0$ ).

*Remark.* Although this system of stochastic processes can be defined without any reference to our random walk problem, it is useful to bear in mind that the state of this cluster queuing system at time  $t$  describes the number of fragments of large cycles at time

$$\frac{n}{2} + \int_0^t \frac{n^2}{2N_s^\uparrow} ds \geq \frac{n}{2}(1+t)$$

since  $N_s^\uparrow \leq n$ . Thus the control obtained in the above lemma for all  $t \leq c$ , will provide useful information for the random walk between times  $n/2$  and  $(1+c)n/2$  for any  $c > 0$ . On our original time-scale, this corresponds exactly to the supercritical regime, i.e. up to time  $cn/2$  for any  $c > 1$ .

*Proof.* The second result is a trivial consequence of the first. The key idea to handle the processes  $\xi_t^k$  is to consider strips  $2^j \leq k < 2^{j+1}$ . Because there are no simultaneous jumps, we can prove that the queues  $\xi_t^k$  at each level  $k$  are independent processes (see e.g. Revuz and Yor (1999), chap. XII, prop. (1.7), for a proof of this fact in the case of Poisson processes). Therefore, for each  $1 \leq j \leq \log_2 n^a$ , the number of cycles with sizes in  $[2^j, 2^{j+1})$ ,  $\zeta_t^j$ , is dominated by a birth and death chain with birth and death rates respectively  $2^j$  and  $2^j \zeta_t^j$ . To analyze these processes, we consider the successive excursions away from 0. Their embedded discrete time processes  $Y_s$  jump from  $m$  to  $m-1$  with probability  $m/(m+1)$  and from  $m$  to  $m+1$  with probability  $1/(m+1)$ . Let us try to find a function  $\phi$  such that  $\phi(0) = 0$ ,  $\phi(1) = 1$  and  $\phi(Y_s)$  is a martingale. The latter implies

$$\frac{1}{m+1}[\phi(m+1) - \phi(m)] = \frac{m}{m+1}[\phi(m) - \phi(m-1)]$$

so  $\phi(x) = \sum_{k=1}^x (k-1)!$ . Since  $\phi(1) = 1$  and  $\phi(0) = 0$ , it follows by optional sampling that the maximum level reached during an excursion of  $\zeta^j$ ,  $M$ , satisfies

$$P(M > x) = 1/\phi(x+1) \leq 1/x! \quad (18)$$

To bound the number of excursions for the process in the  $j^{\text{th}}$  strip before time  $c$ ,  $N_j(c)$ , we note that jumps from 0 to 1 occur at rate  $2^j$  so ignoring the amount of time it takes to return to 0 from 1, the number of excursions by time  $c$  is bounded by a Poisson random variable with mean  $2^j c \leq cn^a$ . Markov's inequality implies that  $P(N_j(c) > n^2) \leq cn^{a-2}$  so

$$P\left(\max_{1 \leq j \leq a \log_2 n} N_j(c) > n^2\right) \rightarrow 0 \quad (19)$$

To estimate the probability that the maximum of  $n^2$  excursions is  $> \log n$  we recall (18) and that Stirling's formula implies  $k! \geq \delta_0 k^k e^{-k} / \sqrt{2\pi k}$  for some  $\delta_0 > 0$ , so

$$(\log n)! \geq \delta_1 (\log n)^{\log n} n^{-1} (\log n)^{-1/2} = \delta_1 n^{\log \log n - 1} (\log n)^{-1/2}$$

The right-hand side goes to  $\infty$  faster than  $n^2 \log_2 n$  so using (19) we have

$$P\left(\max_{1 \leq j \leq a \log_2 n} \max_{0 \leq t \leq c} \zeta_t^j > \log n\right) \rightarrow 0$$

When the last event does not occur we have

$$\sum_{k=1}^{n^a} \xi_t^k \leq a (\log_2 n) \log n = \frac{a}{\log 2} (\log n)^2$$

Since  $a < 2/3 < \log 2 \approx 0.69$ , this gives the desired result.  $\square$

## 6.2. Completion of the proof of Theorem 4

The cluster queuing system is the first approximation to the analysis of the dynamics of the supercritical regime. However, it ignores customer fragmentation and a number of "bad events" that we need to consider in order to give a rigorous proof of Theorem 4. Though *a priori* one might expect it to be difficult to take account of corrections of second order, third order, . . . , and have nightmares about adding up infinitely many terms, we were pleasantly surprised to see that the proof could be completed with a few simple estimates.

The first technical problem to confront is to show that the total amount of mass upstairs stays positive at any given time so we can apply our time change. This is done in Section 6.3.

The more difficult problem is to control the difference between the CQS and the real system of clusters. To do this, we need a notational scheme to verify that we have indeed taken care of all of the relevant events. We call clusters of size larger than  $n^a$  *large*, those in the CQS (i.e., those that were generated by a fragmentation of some large cycle), *medium*, and non-giant clusters in the random graph *small*. Writing *frag* and *coag* as shorthand for fragmentation and coagulation, we have three *frag* and six *coag* events to handle:

*coag(small, small)* is a natural part of the random graph so these events are not errors. The fragmentation of small clusters involves  $o(n^{1/2})$  clusters and hence does not significantly alter this process (see *frag(small)* and Lemma 12).

$coag(small, large)$  eliminates a small component, but in the random graph these correspond to the small cluster being absorbed into the giant component, so this is not an error.

$frag(small)$  is easy to take care of due to the duality principle which asserts that finite clusters in the random graph at time  $c > 1$  have the same distribution as clusters at time  $c\rho < 1$  where  $\rho$  is the probability of no percolation. This allows use to use our subcritical estimates for fragmentation of small supercritical clusters. More details are given in Lemma 12.

$coag(large, large)$  We do not care about these events since we do not need to keep track of the number of cycles upstairs.

$frag(large)$  These are the arrivals in the cluster queuing system.

$coag(medium, large)$  are (almost) the departures in the cluster queuing system. The problem is that the next three events can cause clusters to gain weight or split into two.

$coag(medium, medium)$  are helpful events since they reduce the number of customers in the CQS. This does make the fragmentation rate for the new cluster larger than the sum of the two previous clusters but Lemma 11 will take care of this. More importantly, it makes the departure rate of the new cluster larger. This, applied to  $coag(medium, medium)$  and  $coag(medium, small)$ , shows that the number of medium clusters is stochastically bounded by the CQS of Section 6.1, and is the content of Lemma 10.

$coag(medium, small)$  eliminates a small component, but in the random graph these correspond to the small cluster being absorbed into the giant component. Again, this also makes the fragmentation rate larger for the cluster that gained weight but Lemma 11 will take care of this.

$frag(medium)$  is taken care of by Lemma 11.

To complete the proof it remains to prove the three promised lemmas.

**Lemma 10.** *The number of medium clusters is dominated by that of the CQS. Therefore there are never more than  $(\log n)^2$  medium clusters, and never more than  $n^a (\log n)^2$  vertices in medium clusters.*

*Proof.* As was just mentioned, the only differences between the CQS and the medium clusters are generated by events of type  $coag(medium, medium)$  and  $coag(small, medium)$ . However both those events do not increase the number of medium clusters, and both those events make the death rate of the clusters concerned higher. Hence we can construct the CQS and the medium clusters process on the same probability space, in such a way that the *total* number of medium clusters is smaller than that of the CQS.  $\square$

**Lemma 11.** *The expected number of fragmentations of medium clusters is at most  $O(n^{2a-1} (\log n)^2)$ .*

*Proof.* There are never more than  $(\log n)^2$  medium clusters. Since there are at most  $n^a$  vertices per medium clusters the total number of vertices is at most  $n^a (\log n)^2$ . The rate at which those fragmentations happen is thus bounded by

$$\left( \frac{n^a (\log n)^2}{n} \right) \frac{n^a}{n}$$

so that the expected number of such fragmentations is indeed  $O(n^{2a-1}(\log n)^2)$ .  $\square$

**Lemma 12.** *The number of fragmentations of small components is  $o(n^{1/2})$ .*

*Proof.* By a now familiar estimate, the expected number of fragmentations that produce clusters of size smaller than  $n^p$  at times between  $n$  and  $n+t$  is at most  $2n^{p-1}t$ . So we can ignore fragmentations that (a) produce clusters of size smaller than  $n^{0.45}$  before time  $cn/2$  and (b) produce clusters of size smaller than  $n^{0.55}$  at times between  $n$  and  $n+n^{0.9}$ .

If  $c > 1$  the distribution of nongiant components in the random graph is given by progeny of a Poisson Galton Watson process with mean  $c$  on the event of its extinction. If we let  $\rho$  denote its extinction probability, then the offspring distribution conditional on extinction is given by

$$\frac{1}{\rho} e^{-c} \frac{(c\rho)^k}{k!} = e^{-c\rho} \frac{(c\rho)^k}{k!}$$

since  $\rho = e^{-c(1-\rho)}$ . In short,  $PGW(c)$  conditioned on extinction is  $PGW(c\rho)$ . The last observation implies that results for finite supercritical clusters can be derived from those for subcritical clusters. In particular, by Lemma 3, the largest nongiant components seen after time  $n+n^{0.9}$ , are smaller than  $n^{0.2}$ . Since fragmentations of such clusters necessarily produce pieces smaller than  $n^{0.2}$  these fragmentations can be ignored by (a).  $\square$

### 6.3. The initial mass upstairs

The last step in the proof of Theorem 4 is to ensure that upstairs never becomes empty in this process. In other words we must prove that  $N_t^\uparrow > 0$  for all  $t > 0$  with high probability, so that we can indeed time-change the queues by  $(N_t^\uparrow)^{-1}$ , and use rigorously all the analysis carried out on (CQS) in Section 6.1. This will be done by showing that initially there are already more vertices upstairs than will ever (with high probability) be taken away by fragmentation in the cluster queuing system.

**Lemma 13.** *Initially, upstairs contains at least  $N_0^\uparrow \geq Kn^{1-a/2}$  vertices. In particular  $N_0^\uparrow > n^a (\log n)^2$  and it never becomes empty during the supercritical regime.*

*Proof.* Lemma 6 implies that when  $c = 1$  the expected number of trees of size  $k$

$$ET_k \sim \frac{nk^{-5/2}}{\sqrt{2\pi}} \exp(-k^3/3n^2)$$

If we let  $|\mathcal{C}_{\geq a}| = \sum_{k=n^a}^{\infty} T_k$  then it follows that

$$E|\mathcal{C}_{\geq a}| \sim \frac{n}{\sqrt{2\pi}} \sum_{k=n^a}^{\infty} k^{-5/2} \sim \frac{2}{3\sqrt{2\pi}} n^{1-3a/2}$$

Bollobás (1985) has calculated (see page 107) that the expected number of ordered pairs of trees of sizes  $j$  and  $k$ ,

$$E(T_j, T_k) \leq ET_j ET_k$$

When  $j \neq k$  this implies  $\text{cov}(T_j, T_k) \leq 0$  and for  $j = k$  that  $ET_k(T_k - 1) \leq (ET_k)^2$  or  $\text{var}(T_k) \leq ET_k$ . Summing we have

$$\text{var}(|\mathcal{C}_{\geq a}|) \leq E|\mathcal{C}_{\geq a}|$$

and it follows from Chebyshev's inequality that  $|\mathcal{C}_{\geq a}|/E|\mathcal{C}_{\geq a}| \rightarrow 1$  in probability. These trees have not experienced fragmentation so their size is always at least  $n^a$  and the total mass in large components is at least  $Kn^{1-a/2}$ . When  $a < 2/3$  and  $n$  is large, this is much larger than the  $n^a(\log n)^2$  upper bound on the missing mass due to fragmentations.

At this point the proof of Theorem 4 is complete.  $\square$

#### 6.4. A sharper estimate for the mass upstairs

In section 6.3 above, we have just proved that upstairs never becomes empty in the supercritical regime (Lemma 13). But, as was already mentioned earlier, we expect by analogy with the random graph that in fact a positive fraction of all  $n$  vertices stay upstairs. This is the content of Theorem 5, which we restate here for convenience and then prove.

**Theorem 5.** *For any  $c > 1$ , at time  $cn/2$  there are at least  $\theta(c)n - o(n)$  vertices located on large cycles (i.e., of size greater than or equal to  $n^a$ , for any  $a < 2/3$ ).*

*Proof.* In fact it is a simple consequence of Lemmas 10 and 11. Indeed, the mass missing upstairs must be a piece of the random graph's giant component fallen downstairs by fragmentation. Therefore either it is a medium cluster or it has experienced a consecutive fragmentation. But we now know that there are never more than  $n^a(\log n)^2$  vertices in medium clusters by Lemma 10. On the other hand, by Lemma 11, the expected number of vertices in clusters having experienced multiple fragmentation has to be smaller than

$$n^a \cdot Kn^{2a-1}(\log n)^2 = o(n)$$

as long as  $a < 2/3$ .  $\square$

## References

- Aldous, D.: Brownian excursions, critical random graphs and the multiplicative coalescent. *Ann. Prob.* **25**, 812–854 (1997)
- Aldous, D.: Deterministic and stochastic models for coalescence (aggregation and coagulation) : a review of the mean-field theory for probabilists. *Bernoulli*. **5**, 3–48 (1999)
- Angel, O.: Random infinite permutations and the cyclic time random walk. Pages 9–16 in Banderier and Krattenthaler (2003)
- Arratia, R., Barbour, A., Tavaré, S.: *Logarithmic combinatorial structures : a probabilistic approach*. European Math. Society Monographs, 1. (2003)

- Bafna, V., Pevzner, P.: Sorting by reversals: Genome rearrangement in plant organelles and evolutionary history of X chromosome. *Mol. Biol. Evol.* **12**, 239–246 (1995)
- Ball, F.: The threshold behaviour of epidemic models. *J. Appl. Prob.* **20**, 227–241 (1983)
- Banderier, C., Krattenthaler, C.: Proc. of the conference Discrete Random Walks. Discrete Math and Computer Science. [dmtcs.loria.fr/proceedings/dmACind.html](http://dmtcs.loria.fr/proceedings/dmACind.html) (2003)
- Berestycki, N., Durrett, R.: A phase transition in the random transposition random walk. Pages 17–26 in Banderier and Krattenthaler (2003)
- Bollobás, B.: The evolution of random graphs. *Trans. Amer. Math. Soc.* **286**, 257–274 (1984)
- Bollobás, B.: *Random Graphs*, Cambridge. University Press, 1985
- Borel, E.: Sur l’emploi du théorème de Bernoulli pour faciliter le calcul d’une infinité de coefficients. Application au problème de l’attente à un guichet. *C.R. Acad. Sci. Paris.* **214**, 452–456 (1942)
- Bourque, G., Pevzner, P. A.: Genome-scale evolution: reconstructing gene orders in the ancestral species. *Genome Research.* **12**, 26–36 (2002)
- Devroye, L.: The branching process method in the Lagrange random variate generation, [cgm.cs.mcgill.ca/~luc/branchingpaper.ps](http://cgm.cs.mcgill.ca/~luc/branchingpaper.ps) (1992)
- Diaconis, P., Mayer-Wolf, E., Zeitouni, O., Zerner, M.: Uniqueness of invariant distributions for split-merge transformations and the Poisson-Dirichlet law. *Ann. Prob.*, to appear (2003)
- Durrett, R.: *Probability: Theory and Examples*. Edition, Duxbury Press, 1996
- Durrett, R.: *Probability Models for DNA Sequence Evolution*. Springer-Verlag, New York, 2002
- Durrett, R.: Shuffling Chromosomes. *J. Theor. Prob.* **16**, 725–750 (2003)
- Durrett, R., Nielsen, R., York, T.L.: Bayesian estimation of genomic distance. *Genetics*, to appear (2003)
- Hannehalli, S., Pevzner, P.A.: Transforming cabbage into turnip (polynomial algorithm for sorting signed permutations by reversals). Proceedings of the 27<sup>th</sup> Annual Symposium on the Theory of Computing, 178–189. Full version in the *Journal of the ACM.* **46**, 1–27 (1995)
- Jacod, J., Shiryaev, A.: *Limit Theorems for Stochastic Processes*. Springer, New-York, 1987
- Janson, S., Knuth, D. E., Luczak, T., Pittel, B.: The birth of the giant component. *Rand. Struct. Algor.* **4**, 231–358 (1993)
- Janson, S., Luczak, T., Rucinski, A.: *Random Graphs*. Wiley-Interscience, New York, 2000
- Luczak, T., Pittel, B., Wierman, J. C.: The structure of a random graph near the point of the phase transition. *Trans. Amer. Math. Soc.* **341**, 721–748 (1994)
- Mayer-Wolf, E., Zeitouni, O., Zerner, M.: Asymptotics of certain coagulation-fragmentation processes and invariant Poisson-Dirichlet measures. *Electr. Journ. Prob.* **7**, 1–25 (2002)
- Pevzner, P.A.: *Computational Molecular Biology: An Algorithmic Approach*. MIT Press, Cambridge, 2000
- Pevzner, P.A., Tesler, G.: Genome rearrangement in mammalian evolution: lessons from human and mouse genomes. *Genome Research.* **13**, 37–45 (2003)
- Pitman, J.: Enumerations of trees and forests related to branching processes and random walks. *Microsurveys in Discrete Probability*, D. Aldous and J. Propp editors. DIMACS Ser. Discrete Math. Theoret. Comp. Sci no.41 163–180. Amer. Math. Soc. Providence RI. (1998)
- Pitman, J.: Coalescent random forests, *J. Comb. Theory A.* **85** 165–193 (1999)
- Pitman, J.: Poisson-Dirichlet and GEM invariant distributions for split-and-merge transformations of an interval partition. *Combin. Prob. Comput.* **11**, 501–514 (2002)
- Pitman, J.: Combinatorial stochastic processes. Lecture Notes for St. Flour Course. To appear, available at <http://stat-www.berkeley.edu/users/pitman/> (2003)
- Pittel, B.: On tree census and the giant component in sparse random graphs, *Rand. Struct. Algor.*, **1**, 311–342 (1990)

- Ranz, J.M., Casals, F., Ruiz, A.: How malleable is the eukaryotic genome? Extreme rate of chromosomal rearrangement in the genus *Drosophila*. *Genome Research*. **11**, 230–239 (2001)
- Revuz, D., Yor, M.: *Continuous martingales and Brownian Motion*, Springer-Verlag, New York, 1999
- Schramm, O.: Composition of random transpositions, *Israel J. Math.* to appear (2004)
- Tanner, J.C.: A derivation of the Borel distribution. *Biometrika* **48**, 222–224 (1961)
- York, T.L., Durrett, R., Nielsen, R.: Bayesian estimation of inversions in the history of two chromosomes. *J. Comp. Bio.* **9**, 808–818 (2002)