Creating a Giant Component*

Tom Bohman [†], David Kravitz [‡]
Department of Mathematical Sciences
Carnegie Mellon University
Pittsburgh, PA 15123

June 11, 2004

Abstract

Let c be a constant and $(e_1, f_1), (e_2, f_2), \ldots, (e_{cn}, f_{cn})$ be a sequence of ordered pairs of edges on vertex set $[n]$ chosen uniformly and independently at random. Let A be an algorithm for the on-line choice of one edge from each presented pair, and for $i = 1, \ldots, cn$ let $G_A(i)$ be the graph on vertex set $[n]$ consisting of the first i edges chosen by A. We prove that all algorithms in a certain class have a critical value c_A for the emergence of a giant component in $G_A(cn)$ (i.e. if $c < c_A$ then with high probability the largest component in $G_A(cn)$ has $o(n)$ vertices and if $c > c_A$ then with high probability there is a component of size $\Omega(n)$ in $G_A(cn)$). We show that a particular algorithm in this class with high probability produces a giant component before $0.385n$ steps in the process (i.e. we exhibit an algorithm that creates a giant component relatively quickly). The fact that another specific algorithm that is in this class has a critical value resolves a conjecture of Spencer.

In addition, we establish a lower bound on the time of emergence of a giant component in any process produced by an on-line algorithm and show that there is a phase transition for the off-line version of the problem of creating a giant component.

1 Introduction

Let c be a constant and $(e_1, f_1), (e_2, f_2), \ldots, (e_{cn}, f_{cn})$ be a sequence of pairs of edges on vertex set $[n] = \{1, 2, \ldots, n\}$ chosen uniformly and independently at random. Our goal is to choose exactly one edge from each pair to form a graph that contains a giant component (i.e. a component of size $\Omega(n)$). We consider both the on-line version (pairs appear sequentially and the choice of edge from

*Results similar to Theorems 5 and 6 were obtained independently by Flaxman, Gamarnik and Sorkin [10].

†Supported in part by NSF grant DMS-0100400.
‡Supported in part by a VIGRE grant from the NSF, DMS-9819950.

the pair (e_t, f_t) is made without knowledge of future edges) and off-line version (the choice of edge from the pair (e_t, f_t) is made with complete knowledge of the full set of edges) of the problem. It is well known, a classical result of Erdős and Rényi [7], that for any $c > \frac{1}{2}$ we can create a giant component **whp** by choosing e_i for all i and for any $c < \frac{1}{4}$ **whp** we can not create a giant even if we are allowed to choose all $2cn$ edges. Thus, the question is interesting for $c \in (\frac{1}{4}, \frac{1}{2})$.

Recently, work has been done on a very similar problem in which the object is to avoid a giant component instead of creating one. Here, the interesting case is $c > \frac{1}{2}$. Bohman and Frieze introduced an on-line algorithm and proved that **whp** it produces a graph with no giant component for any $c < 0.535$ [3]. Spencer and Wormald claim that $c = 0.89$ can be achieved by an on-line algorithm [11]. Bohman and Kim showed that the off-line version of the avoiding a giant component problem has a threshold at c_{off} roughly equal to 0.976 [5]. (If $c > c_{\text{off}}$ then **whp** every graph that consists of at least on edge from each pair $(e_1, f_1), \ldots, (e_{cn}, f_{cn})$ has a giant component and if $c < c_{\text{off}}$ then the **whp** there exists a choice of one edge from each pair that produces a graph in which the largest component has size $o(n)$.) This threshold is strictly greater than an upper bound on the on-line version of the avoiding a giant component problem that was established by Bohman, Frieze and Wormald [4]; in other words, there exist values of c for which any on-line algorithm **whp** produces a graph with a giant but there exists a way to make an off-line choice of one edge from each pair that succeeds **whp** in producing a graph with no giant component.

We begin with general on-line processes. Let A be an on-line algorithm for the choice of one edge from each presented pair (i.e. the choice of edge from the pair (e_i, f_i) is made without knowledge of pairs (e_j, f_j) such that $j > i$). Let $G_A(i)$ be the graph on vertex set $[n]$ consisting of the first i edges chosen by A. This produces a random graph process $G_A(1), G_A(2), \ldots, G_A(cn)$. Such a process is called an **Achlioptas process**, after Dimitris Achlioptas who posed the question of on-line avoidance of a giant component. Note that this class includes processes produced by algorithms that might be designed to influence the emergence of a giant component in a variety of ways.

Our main result is a general theorem that establishes the existence of a critical value for the emergence of a giant component for a class of Achlioptas processes. Before defining this class of processes, we discuss two broader classes of processes that were introduced by Spencer [11]. A **size algorithm** A makes the choice between edges e_{t+1} and f_{t+1} based on the sizes of the components in $G_A(t)$ that are joined by e_{t+1} and f_{t+1}. We will denote $e_{t+1} = \{u_{t+1}, v_{t+1}\}$ and $f_{t+1} = \{x_{t+1}, y_{t+1}\}$, for $t = 0, \ldots, cn-1$. We assume (for convenience) that $u_{t+1}, v_{t+1}, x_{t+1}, y_{t+1}$ is an ordered sequence of vertices and let $a_{t+1}, b_{t+1}, c_{t+1}$, and d_{t+1} be the sizes of the components containing $u_{t+1}, v_{t+1}, x_{t+1}$, and y_{t+1}, respectively, in $G_A(t)$. Formally, a size algorithm chooses the edge e_{t+1} if the 4-tuple $(a_{t+1}, b_{t+1}, c_{t+1}, d_{t+1})$ lies in some fixed set (this subset of $\{1, 2, \ldots\}^4$ defines the algorithm) and chooses f_i otherwise. A **bounded size algorithm** is a size algorithm that makes no distinction between components larger than some fixed constant m.

Spencer made a sequence of conjectures regarding these processes [11]. We state two of these here for reference. The issues here are the existence, location and nature of a phase transition in the size of the largest component of $G_A(t)$. For each bounded size algorithm there is a natural candidate for the location of this critical value: the blow-up point in the differential equation for the expected value of the sum of the squares of the component sizes (this quantity is known as the susceptibility in statistical physics). If A is a bounded size algorithm let this value be c_A (the formal definition of c_A is given in Section 2). Spencer and Wormald showed that for any bounded size algorithm and any $c < c_A$ **whp** the largest component in $G_A(cn)$ has $\log n^{O(1)}$ vertices [11].[1]

Conjecture 1 (Spencer). *Let A be the algorithm that takes e_{t+1} if it joins two isolated vertices in $G_A(t)$ and otherwise takes f_{t+1}. If $c > c_A$ then* **whp** *$G_A(cn)$ has a components of size $\Omega(n)$.*

Conjecture 2 (Spencer). *If A is a bounded size algorithm and $c > c_A$ then* **whp** *$G_A(cn)$ has a components of size $\Omega(n)$.*[1]

Conjectures 1 and 2 are Conjectures 2 and 6, respectively, of [11].

Now we are ready to state our main result. We call an algorithm A a **bounded first-edge algorithm** if it chooses between e_{t+1} and f_{t+1} by observing the sizes of the components in $G_A(t)$ connected by e_{t+1}, making no distinction between components larger than some fixed constant m. In other words, such an algorithm looks at the first edge in the pair of random edges and either accepts it based on the sizes of the components involved or rejects it in favor of the (not yet observed and therefore purely random) second edge. Note that the class of bounded first-edge algorithms is contained in the class of bounded size algorithms. We formally define a bounded first-edge algorithm A with a fixed set

$$\mathcal{S}_A \subseteq \{1, 2, \dots\}^2,$$

such that

$$\begin{aligned} (i,j) \in \mathcal{S}_A,\ m < i \text{ and } m < i' \quad &\Rightarrow \quad (i',j) \in \mathcal{S}_A\ , \quad \text{and} \\ (i,j) \in \mathcal{S}_A,\ m < j \text{ and } m < j' \quad &\Rightarrow \quad (i,j') \in \mathcal{S}_A \end{aligned}$$

The algorithm A accepts e_{t+1} if and only if $(a_{t+1}, b_{t+1}) \in \mathcal{S}_A$ (recall that a_{t+1} and b_{t+1} are the sizes of the components in $G_A(t)$ that contain the vertices in e_{t+1}).

Theorem 3. *Let A be a bounded first-edge algorithm. There exists a constant c_A such that*

(a) If $c < c_A$ then **whp** *the largest component in the graph $G_A(cn)$ has $O\left((\log n)n^{12/13}\right)$ vertices, and*

[1]While this paper was on the way to press Spencer and Wormald gave a proof of Conjecture 2 as well as a proof that for any bounded size algorithm A and $c < c_A$ **whp** the largest component of $G_A(cn)$ has $O(\log n)$ vertices [12].

(b) If $c > c_A$ then **whp** *the graph $G_A(cn)$ has a component of size $\Omega(n)$.*

The main tool in the proof of Theorem 3 is the differential equations method for random graph processes (see Wormald [13] for an excellent treatment of this method). The critical value c_A is the blow-up point of the differential equation for the sum of the squares of the component sizes, as above. Part (a) of Theorem 3 is stated for completeness as the stronger results of Spencer and Wormald have not yet appeared. Indeed, our focus here is on part (b) of Theorem 3 which establishes the creation of a giant component. Note that the algorithm in Conjecture 1 is a bounded first edge algorithm (defined by $\mathcal{S}_A = \{(1,1)\}$). Thus, Theorem 3 resolves Conjecture 1.

The machinery introduced in the proof of Theorem 3 can be applied to other situations. Note that bounded first-edge algorithms are static, that is, they employ the same rule throughout the process. The proof goes through, for example, for certain algorithms that always make the choice between e_{t+1} and f_{t+1} without observing f_{t+1} but allow the rule to change some bounded number of times during the process. The main limitation of the proof of Theorem 3 is that it requires a supply of chosen edges around the critical point that are purely random.

Since the critical value c_A given in Theorem 3 is given by a blow-up point in a system of differential equations, it can be estimated by numerically solving the system. In Section 5 we estimate the critical points for two bounded first-edge algorithms. The algorithm A_1 accepts e_{t+1} if neither u_{t+1} nor v_{t+1} is isolated in $G_A(t)$; that is,

$$\mathcal{S}_{A_1} = \{(i,j) : i,j \in \{2,3,\dots\}\}.$$

This algorithm creates a giant component relatively quickly.

Theorem 4. *If $c > 0.385$ then* **whp** *$G_{A_1}(cn)$ has a component of size $\Omega(n)$.*

In Section 5 we also estimate the critical value of the algorithm of Conjecture 1.

Besides analyzing the performance of on-line algorithms, we establish a lower bound on the on-line version of the creating a giant component problem. We show that all on-line algorithms **whp** fail to create a giant for c slightly larger than $\frac{1}{4}$ (recall that the interesting interval for the on-line version of creating a giant component is $\frac{1}{4} < c < \frac{1}{2}$). Define

$$f(d) \;=\; \frac{3 + 8d - 8de^{-4d} - 14de^{-12d} + 14de^{-16d}}{20 - 8e^{-4d} - 14e^{-12d} + 14e^{-16d}}. \qquad (1)$$

Note that $f(d)$ has a local maximum when $d \approx 0.019974$, where $f(d) \approx .2545$.

Theorem 5. *For any constant $c > 0$, if there exists $d \in (0,c)$ such that $f(d) > c$ then for any Achlioptas process* **whp** *all of the components of the graph created in cn steps will be of size $O(\log n)$.*

Finally, we consider the off-line version of the problem of creating a giant component. Here we are given pairs of random edges $(e_1, f_1), \dots, (e_{cn}, f_{cn})$, and we try to create a giant by choosing one edge from each pair. In this case we establish a phase transition.

Theorem 6. *Let c be a constant and let $(e_1, f_1), (e_2, f_2), \ldots, (e_{cn}, f_{cn})$ be a sequence of ordered pairs of edges on vertex set $[n]$ chosen independently and uniformly at random. If $c > \frac{1}{4}$ then* **whp** *there exists a collection of edges E such that $|E \cap \{e_i, f_i\}| \leq 1$ for all i and the graph $([n], E)$ has a component of size $\Omega(n)$.*

Note that it follows from Theorems 5 and 6 that there exist values of c for which any on-line algorithm **whp** produces a graph with no giant but there exists an off-line choice of one edge from each pair that succeeds **whp** in producing a graph with a giant component (which is analogous to the situation for the problem of avoiding a giant component).

The remainder of this paper is organized as follows. In Section 2 we give the proof of our main result, Theorem 3. In section 5, we discuss methods for estimating the critical values for bounded first-edge algorithms, prove Theorem 4, and make some comments about the smallest value of c for which an Achlioptas process can create a giant component. Section 3 consists of the proof of Theorem 5, and Section 4 consists of the proof of Theorem 6. Finally, in Section 6 we mention a very simple Achlioptas process that succeeds in creating a giant component for $c < 0.46$.

2 Bounded first-edge algorithms

Let A be a fixed bounded first edge algorithm. For notational convenience we work with $\mathcal{S}_A$ in a slightly different form. We introduce the symbol ℓ which represents 'all integers larger than m,' and we work with

$$\mathcal{S}_A \subseteq ([m] \cup \{\ell\})^2$$

where $(i, \ell) \in \mathcal{S}_A$ if A accepts e_{t+1} when $a_{t+1} = i$ and $b_{t+1} > m$, etc. We assume without loss of generality that $\mathcal{S}_A \neq ([m] \cup \{\ell\})^2$ (note that if $\mathcal{S}_A = ([m] \cup \{\ell\})^2$ then A always chooses e_{t+1} and we have the standard random graph). We will show (roughly) that the sum of the squares of the component sizes of $G_A(cn)$ is bounded for $c < c_A$ but goes to infinity as c approaches c_A. Furthermore, we will show that for $c < c_A$ the 'large' components do not make a significant contribution to the sum of the squares of the component sizes. Part (b) of Theorem 3 then follows from an application of Lemma 7.

Lemma 7. *Let G be a graph on n vertices and let τ be a constant such that G has $y_i n$ vertices in components of size i for $i = 1, 2, \ldots, \tau$, and $y_1 + \cdots + y_\tau = 1$. If η is a constant such that*

$$2\eta \sum_{i=1}^{\tau} i y_i \; > \; 1, \tag{2}$$

then the graph obtained by adding ηn random edges to G will **whp** *have a component of size $\Omega(n)$.*

Note that the summation in (2) is the sum of the squares of the component sizes of G. This quantity is also the expected component size of a vertex chosen at random, which is known as the susceptibility in statistical physics. The proof of Lemma 7 is given at the end of this section.

We track $m+2$ random variables over the evolution of the process using the differential equations method for random graph process (we follow the notation and use a variation of a general theorem of Wormald, who gives an excellent treatment of this method [13]). For $x = 1, \ldots, n$ let $Y_x(i)$ be the number of vertices in components of size x in $G_A(i)$. Furthermore, define

$$X(i) = \sum_{x=1}^{n} x Y_x(i)$$
$$Z(i) = \sum_{x=1}^{n} x^2 Y_x(i).$$

Note that X gives the sum of the squares of the component sizes while Z gives the sum of the cubes of the component sizes. We track the random variables $Y_1, Y_2, \ldots, Y_m, X, Z$ and show that $Y_1, Y_2, \ldots, Y_m, X$ are concentrated around an expected trajectory. We will only give an upper bound the growth of Z over the course of the algorithm. The relationship between the sum of the squares of the components sizes and the sum of the cubes of the component sizes near the critical value is also a key feature of a recent result of Aldous and Pittel [1] on the emergence of the giant component in a version of the random graph in which both vertices and edges appear as the process evolves.

We define a set of $m+2$ functions on $\mathbb{R}^{m+2}$. For $(z_1, \ldots, z_m, z_X, z_Z) \in \mathbb{R}^{m+2}$ set

$$z_\ell = 1 - z_1 - \cdots - z_m$$
$$\rho = \sum_{(x,y) \in \mathcal{S}_A} z_x z_y.$$

For $w = 1, \ldots, m$ define

$$\begin{aligned} f_w(z_1, \ldots, z_m, z_X, z_Z) = & \sum_{(x,y) \in \mathcal{S}_A : x+y=w} z_x z_y w \\ & - \sum_{(x,y) \in \mathcal{S}_A : y=w} z_x z_w w - \sum_{(x,y) \in \mathcal{S}_A : x=w} z_y z_w w \\ & + (1-\rho) w \left[\sum_{x=1}^{w-1} z_x z_{w-x} - 2 z_w \right]. \end{aligned}$$

Further define

$$\xi_w = w z_w \qquad \zeta_w = w^2 z_w \qquad \text{for } w = 1, \ldots, m$$
$$\xi_\ell = z_X - \sum_{x=1}^{m} \xi_x \qquad \zeta_\ell = z_Z - \sum_{x=1}^{m} \zeta_x$$

and

$$f_X(z_1, \ldots, z_m, z_X, z_Z) = \sum_{(x,y)\in\mathcal{S}_A} 2\xi_x\xi_y + (1-\rho)2z_X^2$$
$$f_Z(z_1, \ldots, z_m, z_X, z_Z) = \sum_{(x,y)\in\mathcal{S}_A} (3\xi_x\zeta_y + 3\xi_y\zeta_x) + (1-\rho)6z_X z_Z.$$

Note that $f_1, \ldots, f_m, f_X, f_Z$ are continuous and satisfy a Lipshitz condition on any bounded domain.

We are interested in the solution of the system of differential equations

$$\frac{dz_x}{dt} = f_x(z_1, \ldots, z_m, z_X, z_Z) \qquad x \in [m] \cup \{X, Z\}$$

with initial condition

$$z_1(0) = 1, \quad z_2(0) = \cdots = z_m(0) = 0, \quad z_X(0) = z_Z(0) = 1.$$

Note that $f_1, \ldots, f_m$ do not depend on z_X or z_Z. The solution of the system

$$\begin{aligned} \frac{dz_x}{dt} = f_x(z_1, \ldots, z_m, z_X, z_Z) = f_x(z_1, \ldots, z_m), \qquad & x \in [m] \\ z_1(0) = 1, \quad z_2(0) = \cdots = z_m(0) = 0 \qquad & \end{aligned} \tag{3}$$

can be viewed as a collection of function on the non-negative reals.

Claim 8. *The solution of (3) satisfies*

$$0 < z_1(t) < 1 \qquad \text{and} \qquad z_2(t), \ldots, z_m(t), z_\ell(t) > 0$$

for $t > 0$ in some neighborhood of 0.

A proof of Claim 8 is given at the end of this section. Since

$$\frac{dz_\ell}{dt} \geq 0 \quad \text{and}$$
$$(z_1, \ldots, z_m, z_\ell) \in (0,1)^{m+1} \quad \Rightarrow \quad \frac{dz_x}{dt} \geq -4xz_x \quad \text{for } x \in [m],$$

it follows from Claim 8 for any $T > 1/4$ there exists a constant $\delta > 0$ such that

$$z_1(t), z_2(t), \ldots, z_m(t), z_\ell(t) > \delta \qquad \text{for } \ t \in (1/4, T). \tag{4}$$

With the functions $z_1, z_2, \ldots, z_m$ in hand we can write

$$\frac{dz_X}{dt} = f_X(z_1, \ldots, z_a, z_X, z_Z) = g_1(t) + g_2(t)z_X + g_3(t)z_X^2 \tag{5}$$

where g_1, g_2, g_3 are bounded, smooth functions of t defined on $[0, +\infty)$. The differential equation (5) has a unique solution $z_X(t)$ passing through $z_X(0) = 1$. Note that this function blows up at some point (to see this, it may be easier to work with ξ_ℓ instead of z_X). We define c_A to be this blow-up point. With $z_X(t)$ in hand we can write

$$\frac{dz_Z}{dt} = f_Z(z_1, \ldots, z_a, z_X, z_Z) = g_4(t) + g_5(t) z_Z \tag{6}$$

where g_4 and g_5 are smooth, bounded, non-negative functions of t on intervals of the form $[0, s)$ where $s < c_A$. It follows that the unique solution $z_Z(t)$ of (6) passing through $z_Z(0) = 1$ exists for $0 \le t < c_A$. In other words z_X and z_Z blow up *at the same point in time*, c_A. Finally, we note that it follows immediately from (4) that there exists a constant $\delta > 0$ such that

$$z_1(t), z_2(t), \ldots, z_m(t), z_\ell(t) > \delta \tag{7}$$

for $1/4 \le t \le 2c_A$.

We add a small wrinkle to the graph process $G_A(1), G_A(2), \ldots$ to produce $G'_A(1), G'_A(2), \ldots$. Set $r(n) = n^{1/13}$. If A calls for the acceptance of an edge $\{x, y\}$ at round i and x or y is in a component of size greater than $r(n)$ then the edge is **not** added to the process (and no edge is added to the graph in the round). This produces the slightly altered graph process $G'_A(1), G'_A(2), \ldots$. The random variables $Y'_1, \ldots Y'_m, X', Z'$ refer to this altered graph process (i.e. $Y'_x(i)$ is the number of vertices in $G'_A(i)$ in components of size x, etc).

Lemma 9. *Let $0 < s < c_A$ be fixed. With probability $1 - O(n^{4/13}\exp(-n^{1/13}))$ we have*

$$\begin{aligned} Y'_x(i) &= nz_x(i/n) + O\left(n^{12/13}\right) && \forall x \in [m] \\ X'(i) &= nz_X(i/n) + O\left(n^{12/13}\right) && , \textit{ and} \\ Z'(i) &\le nz_Z(i/n) + O\left(n^{12/13}\right) \end{aligned}$$

uniformly for all $0 \le i \le sn$.

The proof of Lemma 9 is given below.

In order to complete the proof of Theorem 3 we must make an observation about the relationship between $G_A(i)$ and $G'_A(i)$. Note that $G'_A(i)$ is *not* simply a subgraph of $G_A(i)$: the edges we neglect in the formation of G'_A may influence future decisions. We use the symbol Δ to denote symmetric difference.

Lemma 10. *Let $0 < s < c_A$ be fixed.*

$$Pr\left[\left|E\left(G_A(sn)\right) \Delta E\left(G'_A(sn)\right)\right| = O\left(\frac{n \log n}{r^2(n)}\right)\right] = 1 - o(1).$$

Proof. For $i = 1, \ldots, sn$ let $L(i)$ be the set of vertices $x \in [n]$ such that the component of $G'_A(i)$ containing x has more than $r(n)$ vertices. Let $D(i)$ be the set of vertices $x \in [n]$ such that the component of $G_A(i)$ containing x is different than the component of $G'_A(i)$ containing x *and the size of at least one of these components is at most* m. Let B be the set of rounds $i \le sn$ such that that

$$E(G_A(sn)) \cap \{e_i, f_i\} \ne E(G'_A(sn)) \cap \{e_i, f_i\}.$$

Let $\mathcal{A}$ be the event that there exists $i < sn$ such that $|L(i)| > \frac{Kn}{r^2(n)}$, where $K = 2z_Z(s)$. Note that it follows from Lemma 9 that the probability of $\mathcal{A}$ is exponentially small. We have

$$Pr(i+1 \in B) \le 4\frac{|L(i)| + |D(i)|}{n}.$$

Furthermore, $|D(i+1)| \le |D(i)| + 3m$. We have

$$\begin{aligned} E[|D(i+1)|] &\le \sum_{k=0}^{n} Pr[D(i) = k]\left(k + \frac{12m(k + Kn/r^2(n))}{n}\right) + nPr(\mathcal{A}) \\ &= E[|D(i)|]\left(1 + \frac{12m}{n}\right) + \frac{12Km}{r^2(n)} + nPr(\mathcal{A}) \\ &\le E[|D(i)|]\left(1 + \frac{12m}{n}\right) + \frac{24Km}{r^2(n)} \end{aligned}$$

for n sufficiently large. It follows that

$$\begin{aligned} E[|D(i)|] &\le \frac{24Km}{r^2(n)} \sum_{j=0}^{i} \left(1 + \frac{12m}{n}\right)^j \\ &= \frac{2Kn}{r^2(n)}\left[\left(1 + \frac{12m}{n}\right)^{i+1} - 1\right]. \end{aligned}$$

A similar calculation gives the bound $E[|B|] = O\left(n/r^2(n)\right)$. □

Let $c < c_A$. The largest component in $G'_A(cn)$ has at most $2r(n)$ vertices. It then follows from Lemma 10 that **whp** the largest component in $G_A(cn)$ is of size $O(n \log n / r(n))$. This establishes part (a) of Theorem 3.

We now turn to part (b) of Theorem 3. Let $\epsilon > 0$ and $c = c_A + \epsilon$. Let K_1 be a constant such that

$$\epsilon\delta^2 K_1 > 1. \tag{8}$$

(Recall that δ is defined in (7).) There exists $t < c_A$ and a constant K_2 such that

$$z_X(t) > K_1 + 3 \quad \text{and} \quad z_Z(t) < K_2 - 1.$$

It follows from Lemma 9 that **whp** we have

$$X'(tn) > (K_1 + 2)n \quad \text{and} \quad Z'(tn) < K_2 n.$$

Thus

$$\sum_{x=K_2}^{\infty} xY'_x(tn) \le \frac{1}{K_2} \sum_{x=K_2}^{\infty} x^2 Y'_x(tn) < \frac{1}{K_2} Z'(tn) < n.$$

It follows that

$$\sum_{x=1}^{K_2} xY'_x(tn) > (K_1+1)n.$$

Now, it follows from Lemma 10 that $o(n)$ of the components in $G'_A(tn)$ intersect edges in $E\left(G_A(tn)\right) \Delta E\left(G'_A(tn)\right)$. Therefore,

$$\sum_{x=1}^{K_2} xY_x(tn) > K_1 n.$$

We note that $\left(\frac{i}{n}, \frac{Y_1(i)}{n}, \ldots, \frac{Y_m(i)}{n}\right)$ follows $(i/n, z_1(i/n), \ldots, z_m(i/n))$ well past the critical value.

Lemma 11. *Let $0 < s < 2c_A$ be fixed. With probability*

$$1 - O\left(n^{1/4} \exp\left(\frac{-n^{1/4}}{8m^3}\right)\right)$$

we have

$$Y_x(i) = nz_x(i/n) + O\left(n^{3/4}\right)$$

for all $x \in [m]$, uniformly for all $0 \le i \le sn$.

Lemma 11 follows from a routine application of Theorem 5.1 of [13]. It follows from Lemma 11 and (7) that the probability that f_i is chosen by A is at least δ^2 for $i = tn, \ldots, cn$. Therefore, **whp** at least

$$\frac{3}{4}\delta^2(c-t)n > \frac{3}{4}\delta^2 \epsilon n$$

edges f_i such that $tn < i < cn$ are chosen by A. As these are purely random edges, part (b) of Theorem 3 follows from an application of Lemma 7.

It remains to prove Lemma 7, Claim 8, and Lemma 9.

Proof of Lemma 9. We apply Theorem 5.1 of [13]. Let D be the domain

$$D = [0, c_A] \times [0,1]^m \times [0, 2z_X(s)] \times [0, 2z_Z(s)].$$

The stopping time T is the smallest i for which

$$\left(\frac{i}{n}, \frac{Y'_1(i)}{n}, \ldots, \frac{Y'_m(i)}{n}, \frac{X'(i)}{n}, \frac{Z'(i)}{n}\right)$$

does not lie in the domain D.

For note keeping purposes we set

$$Y'_\ell(i) = \sum_{x=m+1}^{2r(n)} Y'_x(i) = n - \sum_{x=1}^{m} Y'_x(i)$$

$$Y'_r(i) = \sum_{x=r(n)+1}^{2r(n)} Y'_x(i)$$

$$X'_\ell(i) = \sum_{x=m+1}^{2r(n)} xY'_x(i) = X'(i) - \sum_{x=1}^{m} xY'_x(i)$$

$$X'_r(i) = \sum_{x=r(n)+1}^{2r(n)} xY'_x(i)$$

$$X'_x(i) = xY'_x(i), \;\; Z'_x(i) = x^2 Y'_x(i) \qquad \text{for } \; x = 1, \ldots, m$$

$$p(i) = \sum_{(x,y)\in\mathcal{S}_A} \frac{Y'_x(i)Y'_y(i)}{n^2}.$$

Note that $p(i)$ gives the probability that the edge e_{t+1} is chosen. Note further that concentration of $X'_\ell(i)$ and $Y'_\ell(i)$ around some expected trajectory will follow from the concentration results we obtain for $Y'_1(i), Y'_2(i), \ldots, Y'_a(i), X'(i)$. We will not establish concentration of $Y'_r(i)$ or $X'_r(i)$. However, by bounding $Z'(i)$ we will have a trivial upper bound on $Y'_r(i)$ and $X'_r(i)$.

We are now ready to consider the expected changes in our random variables that result from one step of the process. We use h_i to denote the history of the process up to time i (this is just the sequence of edges $e_1, f_1; e_2, f_2; \ldots; e_i, f_i$). For $1 \le u \le m$ we have

$$\begin{aligned}
E[Y'_u(i+1) - Y'_u(i)|h_i] = & \sum_{(x,y)\in\mathcal{S}_A : x+y=u} \frac{Y'_x(i)Y'_y(i)u}{n^2} \\
& - \sum_{(x,y)\in\mathcal{S}_A : y=u} \frac{Y'_x(i)Y'_u(i)u}{n^2} - \sum_{(x,y)\in\mathcal{S}_A : x=u} \frac{Y'_u(i)Y'_y(i)u}{n^2} \\
& + (1-p(i))u \left[\sum_{v=1}^{u-1} \frac{Y'_v(i)Y'_{u-v}(i)}{n^2} - 2\frac{Y'_u(i)}{n} \right] \\
& - \mathbf{1}_{(u/2,u/2)\in\mathcal{S}_A} \frac{Y'_{u/2}(i)u^2/2}{n^2} + 2\mathbf{1}_{(u,u)\in\mathcal{S}_A} \frac{Y'_u(i)u^2}{n^2} \\
& + \mathbf{1}_{(u,\ell)\in\mathcal{S}_A} \frac{Y'_u(i)Y'_r(i)u}{n^2} + \mathbf{1}_{(\ell,u)\in\mathcal{S}_A} \frac{Y'_u(i)Y'_r(i)u}{n^2} \\
& + (1-p(i))u \left[-\mathbf{1}_{u \text{ even}} \frac{Y'_{u/2}(i)u/2}{n^2} \right. \\
& \qquad \left. + 2\frac{Y'_u(i)u}{n^2} + 2\frac{Y'_u(i)Y'_r(i)}{n^2} \right]
\end{aligned}$$

The first three lines of this expression give the expected change in Y'_u under the assumption that e_{i+1} and f_{i+1} neither fall within a connected component nor touch a component of size greater than $r(n)$ (the first two lines give the change when e_{i+1} is chosen while the third line gives the change when f_{i+1} is chosen). The other lines account for the these other possibilities. It follows that for $i < T$ we have

$$\begin{aligned}\left|E[Y'_u(i+1) - Y'_u(i)|h_i] - f_u\left(\frac{Y'_1(i)}{n}, \ldots, \frac{Y'_m(i)}{n}, \frac{X'(i)}{n}, \frac{Z'(i)}{n}\right)\right| \\ = O\left(\frac{1}{n}\right) + O\left(\frac{Y'_r(i)}{n}\right) \\ = O\left(\frac{1}{n}\right) + O\left(\frac{1}{r(n)^2}\right).\end{aligned} \tag{9}$$

Note that we use the fact that $i < T$ implies $Z'(i) < 2z_Z(s)n$ and therefore $Y'_r(i) = O(n/r(n)^2)$.

The expected changes for X' and Z' are more delicate. The key to the calculation for X' is the following observation: If H is an arbitrary graph with connected components $C_1, \ldots C_m$, we set $X(H) = \sum_{i=1}^m |C_i|^2$ and we add a single random edge to H to form the graph H^+ then the expected value of $X(H^+) - X(H)$ is

$$\sum_{i \neq j} \frac{|C_i||C_j|}{n^2} 2|C_i||C_j| = 2\frac{X^2(H)}{n^2} - 2\sum_{i=1}^m \frac{|C_i|^4}{n^2}. \tag{10}$$

The idea is that $2X^2/n^2$ will be the main term and, so long as no large components have appeared, the other term is just a small error (this reasoning is attributed to Janson [11]). We have

$$\begin{aligned}E[X'(i+1) - X'(i)|h_i] = &\sum_{(x,y)\in\mathcal{S}_A} \frac{2X'_x(i)X'_y(i)}{n^2} + (1-p(i))\frac{2X'(i)^2}{n^2} \\ &- \sum_{x:x\neq\ell,(x,x)\in\mathcal{S}_A} \frac{2x^3Y'_x(i)}{n^2} - \mathbf{1}_{(\ell,\ell)\in\mathcal{S}_A}\sum_{x=a+1}^{r(n)} \frac{2x^3Y'_x(n)}{n^2} \\ &- (1-p(i))\sum_{x=1}^{r(n)} \frac{2x^3Y'_x(i)}{n^2} \\ &- \sum_{(x,y)\in\mathcal{S}_A:x=\ell} \frac{2X'_y(i)X'_r(i)}{n^2} - \sum_{(x,y)\in\mathcal{S}_A:y=\ell} \frac{2X'_x(i)X'_r(i)}{n^2} \\ &+ \mathbf{1}_{(\ell,\ell)\in\mathcal{S}_A}\frac{2X'_r(i)^2}{n^2} + (1-p(i))\frac{-4X'(i)X'_r(i) + X'_r(i)^2}{n^2}\end{aligned}$$

The second and third lines account for the sum of the fourth powers of the component sizes as pointed out in (10). The fourth and fifth lines account for the fact that we drop edges that touch components that have more than $r(n)$ vertices. It follows that for $i < T$ we have

$$\begin{aligned}\left| E[X(i+1) - X(i)|h_i] - f_X\left(\frac{Y_1'(i)}{n},\ldots,\frac{Y_m'(i)}{n},\frac{X'(i)}{n},\frac{Z'(i)}{n}\right)\right| \\ = O\left(\frac{r(n)}{n}\right) + O\left(\frac{X_r'(i)}{n}\right) \qquad (11) \\ = O\left(\frac{r(n)}{n}\right) + O\left(\frac{1}{r(n)}\right).\end{aligned}$$

Note that we use the fact that $i < T$ implies both $Z'(i) = O(n)$ and $Y_r'(i) = O(n/r(n)^2)$.

For Z', we note that if H is an arbitrary graph with connected components $C_1, \ldots C_m$, we define $Z(H)$ to be $\sum_{i=1}^m |C_i|^3$ and we add a single random edge to H to form the graph H^+, then the expected value of $Z(H^+) - Z(H)$ is

$$\begin{aligned}\sum_{i\neq j} \frac{|C_i||C_j|}{n^2}\left(3|C_i|^2|C_j| + 3|C_i||C_j|^2\right) &= 6\frac{X(H)Z(H)}{n^2} - 6\sum_{i=1}^m \frac{|C_i|^5}{n^2} \\ &\leq 6\frac{X(H)Z(H)}{n^2}.\end{aligned}$$

It follows that we have

$$\begin{aligned}E[Z'(i+1) - Z'(i)|h_i] \leq \sum_{(x,y)\in\mathcal{S}_A} &\frac{3X_x'(i)Z_y'(i) + 3X_y'(i)Z_x'(i)}{n^2} \\ &+ (1-p(t))\frac{6X'(t)Z'(t)}{n^2}.\end{aligned}$$

Note that we do not have to take into account the edges that touch vertices in components having $r(n)$ or more vertices, as doing so would only result in a stronger upper bound. Also note that this function is *increasing* in $Z'(i)$. We have

$$E[Z'(i+1) - Z'(i)|h_i] \leq f_Z\left(\frac{Y_1'(i)}{n},\ldots,\frac{Y_m'(i)}{n},\frac{X'(i)}{n},\frac{Z'(i)}{n}\right). \qquad (12)$$

We now apply Theorem 5.1 of [13] (conforming to the notation there as much as possible). We set

$$\beta(n) = 8r(n)^3 = 8n^{3/13}, \qquad \lambda(n) = \frac{1}{r(n)} = \frac{1}{n^{1/13}} \qquad \text{and} \qquad \gamma(n) = 0.$$

Note that the boundedness hypothesis follows from the fact that we do not add edges that touch the component that have $r(n)$ or more edges. The trend hypothesis for variable $Y_1', \ldots, Y_m', X'$ follows from (9) and (11). Note that we have only a one-sided trend hypothesis for Z'. A minor alteration of the proof of Theorem 5.1 in [13] can account for this because

(a) z_Z does not appear in any of the functions other than f_Z, and

(b) f_Z is increasing in z_Z.

It follows from (a) that a one-sided bound on Z' does not influence the bounds for any other variable. It follows from (b) that having only a one-sided bound on Z' suffices to establish a one-sided bound on Z' in future iterations.

Finally, we note that in our application of Theorem 5.1 of [13] we violate one of the stated conditions: There is not a constant C_0 such that $|y_l(h_i)| < C_0 n$ for all h_i. However, this condition is not necessary as we have $\gamma(n) = 0$.

□

Proof of Lemma 7. Let $G = ([n], E)$ and $E' \subseteq \binom{[n]}{2}$ be a set of ηn edges chosen uniformly at random. Let $C_1, \ldots, C_\kappa$ be the connected components of G. Let H be the graph with vertex set $\{v_{C_1}, \ldots, v_{C_\kappa}\}$ and an edge between vertices v_{C_i} and v_{C_j} if there is an edge in E' between components C_i and C_j. We have

$$Pr\left(deg(v_C) = i\right) = \binom{2\eta n}{i}\left(1 - \frac{|C|}{n}\right)^{2\eta n - i}\left(\frac{|C|}{n}\right)^{i}.$$

It follows that **whp** H has

$$d_n(i) := \sum_{k=1}^{\tau} \frac{a_k n}{k} \frac{(2\eta k)^i}{i!} e^{-2\eta k} + o\left(n^{2/3}\right) \tag{13}$$

vertices of degree i for each $i \leq O(\log n / \log\log n)$ and no vertices of higher degree.

If we condition on the degree sequence, then we can view H as a graph chosen uniformly at random from the collection of all graphs having the given degree sequence. When (13) holds we apply a theorem of Molloy and Reed on the giant component in graphs with a fixed degree sequence. Following the notation of [9], we set

$$\lambda_i = \sum_{k=1}^{\tau} \frac{a_k}{k} \frac{(2\eta k)^i}{i!} e^{-2\eta k}$$

and $\mathcal{D} = d_n(1), d_n(2), \ldots$. If (13) holds then we have

$$\begin{aligned}
Q(\mathcal{D}) &= \sum_{i \geq 1} i(i-2)\lambda_i \\
&= \sum_{i \geq 1} i(i-2)\left[\sum_{k=1}^{\tau} \frac{a_k}{k} \frac{(2\eta k)^i}{i!} e^{-2\eta k}\right] \\
&= 2\eta\left[\sum_{k=1}^{\tau} (2\eta k - 1) a_k\right] \\
&> 0.
\end{aligned}$$

It follows from Theorem 1(a) of [9] that **whp** H has a component of size $\Omega(n)$. Therefore, $G + E'$ has a component of size $\Omega(n)$. □

Proof of Claim 8. It follows from a simple induction argument that

$$z_k^{(i)}(0) = 0 \quad \text{for} \quad i = 0, \ldots, k-2 \quad \text{and} \quad 2 \leq k \leq m. \tag{14}$$

Define $w_k = 1 - z_1 - z_2 - \cdots - z_k$ for $k = 1, \ldots, m$. Note that $z_\ell = w_m$. Again by induction we have

$$w_k^{(i)}(0) = 0 \quad \text{for} \quad i = 0, \ldots, k-1 \quad \text{and} \quad 1 \leq k \leq m. \tag{15}$$

It follows from (14) and (15) that for all $\delta > 0$ there exists $\epsilon > 0$ such that

$$|z_k(t)| < \frac{\delta}{(k-2)!} t^{k-2} \quad \text{for} \quad 0 \leq t \leq \epsilon \quad \text{and} \quad 2 \leq k \leq m \tag{16}$$

$$|w_k(t)| < \frac{\delta}{(k-1)!} t^{k-1} \quad \text{for} \quad 0 \leq t \leq \epsilon \quad \text{and} \quad 1 \leq k \leq m \tag{17}$$

Let m' be the smallest integer greater or equal to 2 such that there does not exist $(x, y) \in \mathcal{S}_A$ such that $x + y = m'$. Another inductive argument gives (choosing ϵ sufficiently small)

$$z_k(t) = \Theta(t^{k-1}) \quad \text{for} \quad 0 \leq t \leq \epsilon \quad \text{and} \quad 1 \leq k < m' \tag{18}$$

$$w_k(t) = \Theta(t^k) \quad \text{for} \quad 0 \leq t \leq \epsilon \quad \text{and} \quad 1 \leq k < m' - 1. \tag{19}$$

Note that this gives the full Claim if $m' \geq m + 2$.

Suppose that $m' \leq m+1$. A careful analysis of ρ (using (16), (18) and (19)) gives

$$1 - \rho = \Theta\left(t^{m'-2}\right) \quad \text{for} \quad 0 \leq t \leq \epsilon.$$

It then follows that

$$z_k(t) > 0 \quad \text{for } m' \leq k \leq m \quad \text{and } 0 < t \leq \epsilon.$$

Finally,

$$z_\ell(t) > 0 \quad \text{for } 0 < t < \epsilon.$$

□

3 On-line Lower Bound

Let c be a constant and let $(e_1, f_1), (e_2, f_2), \ldots, (e_{cn}, f_{cn})$ be a sequence of ordered pairs of edges on vertex set $[n]$ chosen independently and uniformly at random. We will show that if there exists $d \in (0, c)$ such that $c < f(d)$ where $f(d)$ is the function defined in (1) (i.e. if $c \leq .2545\ldots$) then any on-line algorithm for the sequential choice of one edge from each presented pair (e_i, f_i), $i = 1, 2, \ldots, cn$, **whp** will create a graph whose components are all of size $O(\log n)$. To do this, we will use the fact that on-line algorithms usually cannot make "good" choices early in the process. For small i, there is a good chance that all four vertices from (e_i, f_i) are appearing for the first time. In this case any on-line algorithm must make an essentially arbitrary choice, which could be a costly mistake if, for example, f_i is an isolated edge in the graph consisting of all $2cn$ edges but e_i is a bridge in the giant component.

Fix some constant $d \in (0, c)$. We will determine its value later. Divide the process into two parts, the first dn pairs and the last $(c - d)n$ pairs. Given the pairs $(e_1, f_1), \ldots, (e_{cn}, f_{cn})$, we create an auxiliary graph G on vertex set $[n]$ using the following steps:

(i.) For all $i < dn$, if all four vertices in (e_i, f_i) are occurring for the first time in the process, then randomly eliminate one of the edges with probability $\frac{1}{2}$.

(ii.) Take **all** remaining edges, including the last $2(c - d)n$ edges, and let this be the edge set of G.

In other words, G will contain every edge except for some that were chosen with probability $\frac{1}{2}$. It follows from symmetry that for any fixed algorithm A, the probability that A produces a giant is bounded above by the probability that G has a giant. Indeed, if we condition on the first dn rounds of the process then there is a permutation of $[n]$ that maps the graph produced by A in the first dn rounds to a subgraph of the graph consisting of the edges in G that come from $(e_1, f_1), \ldots, (e_{dn}, f_{dn})$.

In order to analyze G we consider a slightly different probability space. Let $a_1, b_1, c_1, d_1; a_2, b_2, c_2, d_2; \ldots$ be a sequence of vertices chosen uniformly and independently at random from $[n]$ with replacement (e.g. we allow $u_i = v_i$). Setting $e_i = \{a_i, b_i\}$ and $f_i = \{c_i, d_i\}$, we get a sequence of pairs of random edges. Of course, this model allows loops and multiple edges, but since the expected number of each over the cn rounds of our process is bounded by an absolute constant, **whp** the total number of loops and multiple edges is at most $\log(n)$. Thus, we can accommodate these flaws in the process by adding an extra $\log(n)$ rounds (which have no impact on the rest of argument).

The following is a variation of the branching process proof of the classical giant component results for $G_{n,p}$ which is given in [8, p. 109]. We generate an upper bound on the size of the component of G containing a fixed vertex v with the following process.

We begin by observing the locations of all appearances of v in the sequence $a_1, b_1, c_1, d_1; a_2, b_2, c_2, d_2; \ldots$. Note that we have not yet observed any whole edge and, conditioning on these locations, the rest of the random vertices in the sequence are uniform in $[n] \setminus \{v\}$. The edge-partners of each occurrence of v are now **active positions** (we still haven't looked at the vertices in these positions). An active position in round i such that $i \le dn$ is called an **early** active position. An active position in round i such that $i > dn$ is called a **late** active position.

Now, if the first appearance of v is in round i and $i \le dn$ then we reveal the other vertices in round i. Let x be the edge-partner of v in this round and let y and z be the other two vertices that appear in this round. We check to see if there is appearance of vertices x, y or z *before* round i. In particular, we determine the first occurrence of a vertex from the set $\{x, y, z\}$. If x, y and z do not appear before round i then we delete this active position with probability $\frac{1}{2}$. If x is the first vertex from the set $\{x, y, z\}$ to appear and its first appearance occurs before round i then the position remains active and we call it a **heavy** active position. If y or z is the first vertex from the set $\{x, y, z\}$ to appear then the position remains an early active position. The vertices x, y and z are now **sussed** and the vertex v is **saturated**. Note that so-far this process has simply determined the degree of v and checked to see if the edge containing the first occurrence of v is deleted in the formation of G while revealing as little information about the random graph as possible.

Now suppose that after some number of steps in this process we have given sets of saturated and sussed vertices, and sets of early, heavy, and late active positions. The saturated vertices are vertices in the component containing v whose complete neighborhood has already been determined, and the active positions correspond to vertices in the component containing v for which we have not yet observed the full neighborhood. We now consider the vertex, say vertex u, in an arbitrary active position π. We reveal all other occurrences of u in the sequence $a_1, b_1, c_1, d_1; a_2, b_2, c_2, d_2; \ldots$ and the edge-partners of all occurrences of u (other than the edge-partner of π which is an already saturated vertex) become active positions (early or late depending on the round in which they occur). If π is an early or heavy active position then we may have already checked to see if the first occurrence of u is in a deleted edge and we cannot make that check again. If π is a late active position and the first occurrence of u is in a round i such that $i \leq dn$, then we suss out the other vertices in round i. If all vertices appearing in round i are making their first appearance, then we delete the active position in round i with probability $\frac{1}{2}$. If, on the other hand, the edge-partner of the first occurrence of u appears before any of the other vertices that appear in round i then this position becomes a heavy active position. This step is repeated until the set of active positions is exhausted.

Before turning to a thoroughly rigorous analysis of this process, it will be helpful to note that it can be loosely modeled with a branching process in which there are three types of offspring: types 1,2 and 3 corresponding to late, early and heavy active positions, respectively. Note that the probability that a late active position generates an early active position that is deleted can (roughly speaking) be bounded from below by

$$\tfrac{1}{2}\left(1 - \left(1 - \tfrac{1}{n}\right)^{4dn}\right)\left(1 - \tfrac{3}{n}\right)^{4dn} \quad \approx \quad \frac{1 - e^{-4d}}{2e^{12d}} \quad =: p.$$

Furthermore, the probability that a late active position generates a heavy active position is (roughly speaking) at most

$$\tfrac{1}{3}\left(1 - \left(1 - \tfrac{1}{n}\right)^{4dn}\right)\left(1 - \left(1 - \tfrac{3}{n}\right)^{4dn}\right) \approx \tfrac{1}{3}(1 - e^{-4d})(1 - e^{-12d}) =: q.$$

The i, j position in the matrix

$$A = \begin{bmatrix} 4(c-d) & 4d-p & q \\ 4(c-d) & 4d & 0 \\ 4(c-d) & 4d+1 & 0 \end{bmatrix}$$

gives the expected number of offspring of type j of an animal of type i in this multi-type branching process. The largest eigenvalue of A is less than 1 if and only if $c < f(d)$, where $f(d)$ is defined in (1). Since this branching process dies out with probability one (see [2, page 186]) if the largest eigenvalue of the matrix A is less than one, we expect $c < f(d)$ to imply that **whp** the component of G containing v is small.

We now make a rigorous argument. Let $\epsilon > 0$ be a constant such that

$$4c + 4(c-d)(2q-p) + 8\epsilon < 1,$$

note that such a constant exists as the condition we impose on c in the statement of the theorem is equivalent to $c < (\frac{1}{4} + d(2q-p))/(1+2q-p))$. Also, set

$$m := \frac{4(4c+\epsilon)}{\epsilon^2} \log n.$$

We consider the first m steps in the process. We begin by noting that it may happen that when we reveal the vertex in some position we will find that it has already been sussed (note that the vertices in some of the early active positions and all of the heavy active positions have already been sussed but this fact is built-in to the process). This 'bad' event occurs with probability at most $\frac{3m}{n-m}$ at each position. Since the number of positions that we inspect is at most $4m$, the probability that the bad event occurs more than 1 time in the process is at most $\binom{4m}{2}(\frac{3m}{n-m})^2 = o(\frac{1}{n})$. Furthermore, when this bad event occurs it introduces at most 2 'extra' active positions, which will have no affect on the rest of the argument. We henceforth assume that no vertex that is revealed has been previously sussed.

It may also happen that the first appearance of some vertex that is revealed in a late active position is in a round that also contains a previously viewed vertex (either saturated or sussed). Of course, this event will change the probability of both the deletion of this active position and the probability that this position becomes a heavy active position. The probability of this event is at most $\frac{3}{n-4m}$ times the number of rounds that hold previously viewed vertices. Since the probability that there exists a vertex that appears more than $2\log n$ times in the sequence $a_1, b_1, c_2, d_1; a_2, b_2, c_2, d_2; \ldots$ is $o(\frac{1}{n})$, the probability of this second type of 'bad' event in any step of the process is at most $12m\frac{\log n}{n-4m}$. Again, we see that we may assume that this event occurs at most once during the process, with no consequence for the rest of the proof.

Let X be the number of late active positions that are introduced in the first m steps of the process. Since the number of late active positions introduced in any one of these steps is dominated by $\mathrm{Bi}\left(4(c-d)n, \frac{1}{n-m}\right)$, X is dominated by

Bi $\left(4(c-d)mn, \frac{1}{n-m}\right)$. It follows from the Chernoff bound ([8], p. 26), that

$$\Pr\left(X \leq (4(c-d)+\epsilon)m\right) \;=\; 1 - o(\tfrac{1}{n}). \tag{20}$$

Note that we do not necessarily saturate all of the late active positions during the first m steps of the process: if there are many active positions it may be the case that some of the late active positions are left over after the m^{th} step. Let X' be the number of late active positions that are actually saturated during the first m steps.

Let Z be the number of heavy active positions that are generated during this process. The probability that a late active position generates a heavy active position is at most

$$\tfrac{1}{3}\left(1-\left(1-\tfrac{1}{n-4m}\right)^{4dn}\right)\left(1-\left(1-\tfrac{3}{n-4m}\right)^{4dn}\right) = q + O\left(\tfrac{m}{n}\right)$$

So, conditioning on (20), Z is dominated by Bi$\Big((4(c-d)+\epsilon)m, q+O\left(\frac{m}{n}\right)\Big)$. It follows from the Chernoff bound that

$$\Pr\left(Z < (4(c-d)q + 2\epsilon)\, m\right) \;=\; 1 - o\left(\tfrac{1}{n}\right). \tag{21}$$

It remains to bound the number of early active positions that are generated in the first m steps of the process. The probability that a late active position generates an early active position that is deleted is at least

$$\tfrac{1}{2}\left(1-\left(1-\tfrac{1}{n}\right)^{4dn-8dm\log n}\right)\left(1-\tfrac{3}{n-4m}\right)^{4dn} = p - O\left(\tfrac{m\log n}{n}\right)$$

Let W be the number of deleted early positions in the first m steps of the process. W dominates Bi$\Big(X', p - O(\frac{m\log n}{n})\Big)$, and hence

$$\Pr\left(W \geq X'(p-\epsilon)\right) \;=\; 1 - o\left(\tfrac{1}{n}\right). \tag{22}$$

Let Y be the sum of m i.i.d. Bi $\left(4dn, \frac{1}{n-4m}\right)$ random variables. We have

$$\Pr\left(Y \leq (4d+\epsilon)m\right) \;=\; 1 - o(\tfrac{1}{n}) \tag{23}$$

The number of early active positions that are generated (and not deleted) in this process is dominated by $Y + Z - W$.

Now, in the event the component containing v has more than m vertices, the process does not terminate until after the m^{th} step and we have

$$m \leq X' + Z + (Y + Z - W). \tag{24}$$

It follows from (20), (21), (22) and (23) that, with probability $1 - o(\frac{1}{n})$, the right hand side of (24) is at most

$$\begin{aligned} X'(1-p+\epsilon) &+ 2m(4(c-d)q) + 4dm + 5\epsilon m \\ &\leq 4(c-d)(1-p)m + 2m(4(c-d)q) + 4dm + 8\epsilon m \\ &< m. \end{aligned}$$

Thus, the probability that the component containing v has more than m vertices is $o(\frac{1}{n})$, and the probability that G contains a component having more than m vertices is $o(1)$, and we have proved Theorem 5.

4 Off-line Phase Transition

Fix $c > \frac{1}{4}$ and suppose $(e_1, f_1), \ldots, (e_{cn}, f_{cn})$ are pairs of random edges. In this section we prove that there exists a set E of edges such that $|E \cap \{e_i, f_i\}| \leq 1$ for $i = 1, \ldots, cn$ and the graph $([n], E)$ has a component of size $\Omega(n)$.

Let G be the graph with vertex set $[n]$ and all $2cn$ edges. For any tree T within graph G, we will say that T **survives** if there is no $i \in \{1, 2, \ldots, cn\}$ such that $\{e_i, f_i\} \subseteq E(T)$. In other words, T survives if no two edges in T were paired together. Clearly, if T is a tree in G which survives, then it is possible to make choices so all of T ends up in the final graph.

If T is a tree in G with t vertices, let $\phi(t)$ be the probability that it survives. A straightforward calculation shows

$$\phi(t) = \prod_{i=1}^{t-2} \frac{2cn - 2i}{2cn - i} = \prod_{j=1}^{\frac{t}{2}-1} \frac{2cn - t - 2j}{2cn + 1 - 2j} = \prod_{j=1}^{\frac{t}{2}-1} \left(1 - \frac{t+1}{2cn + 1 - 2j}\right).$$

Bounding using the first and last terms in the product and using $\mathbf{e}^{-2x} \leq 1 - x \leq \mathbf{e}^{-x}$ for small x leads to

$$\exp\left(-\tfrac{t^2}{2cn-t}\right) \leq \phi(t) \leq \exp\left(-\tfrac{t^2}{4cn} + O(\tfrac{t}{n})\right). \tag{25}$$

With high probability, there exists a surviving tree with $\lfloor n^{1/3} \rfloor$ vertices because $\phi(\lfloor n^{1/3} \rfloor) \to 1$ and **whp** G has a component of size $\Omega(n) \geq \lfloor n^{1/3} \rfloor$.

If T is a surviving tree, we will say that T is **maximal** if there is no edge e such that $T \cup \{e\}$ is a surviving tree. So, if T is maximal then for every edge $\{u, v\} \in E(G)$ such that $u \in T$ and $v \notin T$, necessarily $\{u, v\}$ is paired with some edge in T, otherwise we could add it to T and create a larger surviving tree. Define

$$\mathcal{T} = \{t : \log^2 n \leq t \leq (c - \tfrac{1}{4})^2 n \} \tag{26}$$

We will prove:

$$\Pr(\exists \text{ a maximal surviving tree with size } t \in \mathcal{T}) \to 0. \tag{27}$$

Since $\lfloor n^{1/3} \rfloor \in \mathcal{T}$, (27) establishes the existence of a surviving tree of size at least $(c - \frac{1}{4})^2 n$ **whp**.

For the remainder of this section, note that any given edge appears in G with probability $\frac{2cn}{\binom{n}{2}} = \frac{4c}{n-1}$. Since the appearance of fixed edges makes others less likely, we can say

$$\Pr(\text{a set of } j \text{ edges appears}) \le (\tfrac{4c}{n-1})^j.$$

Lemma 12. *Let $P = \Pr(T$ maximal$|T$ survives$)$, where T is a tree on K_n with t vertices. We have*

$$P \;\le\; t\exp\left(-\tfrac{4c}{n-1}t(n-t)\left[1-\tfrac{t}{c(n-1)}-\tfrac{t}{2cn-t}\right]+O\left(\tfrac{t}{n}\right)\right). \tag{28}$$

Proof. To find the probability that T is maximal, note that there are $t(n-t)$ edges "leaving" T. Every one of these edges must either be left out of G or be paired with an edge from T. The probability that an edge is paired with something in T is bounded above by $\frac{t}{2cn-t}$. In the following sum, j is the exact number of the $t(n-t)$ edges "leaving" T that appear in G:

$$P \;\le\; \sum_{j=0}^{t-1} \tbinom{t(n-t)}{j}(\tfrac{4c}{n-1})^j(\tfrac{t}{2cn-t})^j(1-\rho)^{t(n-t)-j},$$

where ρ is a lower bound on the probability that a fixed edge appears in G, conditioned on all edges that have been observed. The probability ρ will take its smallest possible value when $2(t-1)$ edges have already been observed to exist in G while possibly zero edges have been left out. We can say

$$\rho \;\ge\; \frac{2cn-2(t-1)}{\binom{n}{2}-2(t-1)} \;\ge\; \frac{4c}{n-1}-\frac{4t}{(n-1)^2}.$$

Now we will find the maximum possible value of P. We avoid the sum by finding the maximum and multiplying by t.

$$\begin{aligned} P \;\le\; & t\exp\left(-t(n-t)\left(\tfrac{4c}{n-1}-\tfrac{4t}{(n-1)^2}\right)\right) \\ & \qquad \cdot \max_{0\le j<t}\left(t(n-t)\tfrac{4c}{n-1}\exp(\tfrac{4c}{n-1}-\tfrac{4t}{(n-1)^2})\tfrac{t}{2cn-t}\tfrac{\mathbf{e}}{j}\right)^j. \end{aligned}$$

The function $(\lambda\mathbf{e}/j)^j$ is maximized at $j=\lambda$ with maximum value $\exp(\lambda)$. After this and substituting $\exp(\frac{4c}{n-1}-\frac{4t}{(n-1)^2}) = 1+O(\frac{1}{n})$, then factoring out $\frac{4c}{n-1}$, we get (28). $\square$

Let E_t be the expected number of maximal surviving trees with t vertices. In order to prove (27), it is sufficient to show $nE_t \to 0$ for any $t \in \mathcal{T}$. First let us find the probability p_T that a given tree T with t vertices becomes a maximal surviving tree.

$$p_T \;=\; \Pr(T\subseteq G)\Pr(T \text{ survives}|T\subseteq G)\Pr(T \text{ maximal}|T \text{ survives})$$

It is easy to see that $\Pr(T \subseteq G) \le (\frac{4c}{n-1})^{t-1} \le \frac{n}{4c}(\frac{4c}{n-1})^t$. We have an upper bound for $\Pr(T \text{ survives} | T \subseteq G)$ from (25). Combining these with (28),

$$p_T \;\le\; \tfrac{n}{4c}(\tfrac{4c}{n-1})^t \; t \exp\left(-\tfrac{4c}{n-1}t(n-t)\left[1 - \tfrac{t}{c(n-1)} - \tfrac{t}{2cn-t}\right] - \tfrac{t^2}{4cn} + O(\tfrac{t}{n})\right)\Big).$$

Since there are $\binom{n}{t}t^{t-2} \le \frac{1}{t^2}(n\mathbf{e})^t$ trees on K_n with t vertices, we have $E_t \le \frac{1}{t^2}(n\mathbf{e})^t p_T$, so

$$nE_t \;\le\; \tfrac{n^2}{4ct}(4c\mathbf{e})^t \left\{\exp\left(-\tfrac{4c}{n-1}(n-t)t\left[1 - \tfrac{t}{c(n-1)} - \tfrac{t}{2cn-t}\right] - \tfrac{t^2}{4cn} + o(t)\right)\right\},$$

leading to

$$\begin{aligned} nE_t \;&\le\; \exp\left(2\log n + t\log(4c\mathbf{e})\right) \\ &\qquad\cdot \exp\left(-\tfrac{4c}{n-1}(n-t)t\left[1 - \tfrac{t}{c(n-1)} - \tfrac{t}{2cn-t}\right] - \tfrac{t^2}{4cn} + o(t)\right). \end{aligned}$$

We have $2\log n = o(t)$. Take a factor of t out of the exponent, then let $t = \alpha n$.

$$nE_t \;\le\; \exp\left(\log(4c\mathbf{e}) - 4c(1-\alpha)\left[1 - \tfrac{\alpha}{c} - \tfrac{\alpha}{2c-\alpha}\right] - \tfrac{\alpha}{4c} + o(1)\right)^t.$$

The inside is increasing over all values of $\alpha \ge 0$ provided $c > \frac{1}{4}$, and it is negative whenever $\alpha < (c - \frac{1}{4})^2$. Therefore, $nE_t \to 0$ whenever $t \in \mathcal{T}$. We have proved Theorem 6.

5 Two Algorithms

We begin with the bounded first-edge algorithm A_1. Using the notation of Section 2 we have $m = 1$ and $\mathcal{S}_{A_1} = \{(\ell, \ell)\}$. In words, A_1 chooses edge e_i if it includes no isolated vertices and otherwise chooses edge f_i. For this algorithm we have

$$\begin{aligned} \frac{dz_1}{dt} &= -2z_1(2z_1 - z_1^2) \\ \frac{dz_X}{dt} &= 2(z_X - z_1)^2 \;+\; 2z_X^2(2z_1 - z_1^2) \end{aligned}$$

with initial conditions $z_1(0) = 1$, $z_X(0) = 1$. In order to get an upper bound on the critical value c_{A_1}, we approximate the solution by implementing Euler's method, being sure to underestimate z_X. (The program for our approximation, written in C++, is available at `http://www.math.cmu.edu/~tbohman`.)

Let $\tilde{z}_1$ and $\tilde{z}_X$ denote our approximations. We use the standard Euler's method for $\tilde{z}_1$. For $\tilde{z}_X$ we set

$$\begin{aligned} \tilde{z}_X(t+h) &= \tilde{z}_X(t) + h\phi(\tilde{z}_X, \tilde{z}_1) - 2\epsilon, \quad \text{where} \\ \phi(\tilde{z}_x, \tilde{z}_1) &= 2\max\left(\tilde{z}_X - \tilde{z}_1 - 2\tfrac{\epsilon}{h}, 0\right)^2 + 2\tilde{z}_X^2(2\tilde{z}_1 - \tilde{z}_1^2 - \tfrac{2\epsilon}{h}), \end{aligned}$$

h is the step size, and ϵ accounts for the computational rounding errors. For our approximation we take $h = 10^{-7}$, and $\epsilon = 10^{-12}$ suffices. We claim that we have the following (a.) $\tilde{z}_X(t) < z_X(t)$

(b.) $\tilde{z}_X(0.3847) > 10^4$, and

(c.) $z_1(t) > 1/2$ for $t \in [0, 0.385]$.

It then follows from Lemmas 7 and 9 that $c_{A_1} < 0.3847 + 0.0001 < 0.385$. In order to establish these claims we first note that

(d.) $e(t) := |z_1(t) - \tilde{z}_1(t)| \leq \frac{2\epsilon}{h}$ for $t \in [0, 0.4]$.

To see this, let $f_1(z_1) = -2z_1(2z_1 - z_1^2)$. Since $|z_1''(t)| \leq 4.2$ and $|f_1(z_1(t)) - f_1(\tilde{z}_1(t))| \leq 3e(t)$ for all $t \in [0, 0.4]$, we have

$$e(t + h) \leq (1 + 3h)e(t) + 2.1h^2 + \epsilon.$$

This leads to (d.). We note that (d.) together with observation of $\tilde{z}_1$ suffices to establish (c.). Of course, (b.) follows from the observation of $\tilde{z}_X$ alone. It remains to prove (a.). To this end, we first note that

(e.) $z_X''(t) > 0$ for all $t > 0$.

We have

$$z_X'' = 2(z_X - z_1)(z_X' - z_1') + 2z_X z_X' z_1(2 - z_1) + 2z_X^2(1 - z_1)z_1',$$

and therefore we may use $-2 \leq z_1' \leq 0$, $z_X \geq 1$, and $z_1 \geq \frac{1}{2}$ to get

$$z_X'' \geq 2(1 - z_1)[z_X' + \tfrac{1}{2}z_X z_X' - 2z_X^2] \geq \tfrac{3}{2}z_X' - 2z_X^2.$$

It is easy to see from the original differential equation that $z_X' \geq \frac{3}{2}z_X^2$ whenever $z_1 \geq \frac{1}{2}$, thus $z_X'' \geq \frac{1}{4}z_X^2 > 0$. Finally, we note that

(f.) $\tilde{z}_X(t) \leq z_X(t)$ implies $\phi(\tilde{z}_X(t), \tilde{z}_1(t)) < z_X'(t)$.

Claim (a.) now follows from $z_X(0) = \tilde{z}_X(0)$, (e.) and (f.).

We now turn to the algorithm of Conjecture 1 and apply our machinery to determine the existence and location of a phase transition. Let the bounded first-edge algorithm A_2 defined by $m = 1$ and $\mathcal{S}_{A_2} = \{(1, 1)\}$. In words, this algorithm chooses edge e_i if it is an isolated edge and otherwise chooses f_i. For this algorithm we have

$$\frac{dz_1}{dt} = -2z_1(t)^2 - 2z_1(t)(1 - z_1(t)^2)$$
$$\frac{dz_X}{dt} = 2z_1(t)^2 + 2z_X^2(t)(1 - z_1(t)^2)$$

Using the methods above to approximate the solution of this system of differential equations we have $z_X(.5882) > 10^4$. This implies $c_{A_2} < .589$.

We close this section by noting that simulations suggest that the Achlioptas processes that chooses the edge from the pair (e_i, f_i) that results in the largest increase in the sum of the squares of the component sizes creates a giant component in as few as $0.34n$ rounds. However, this algorithm is not a first-edge algorithm: it depends on all four vertices in $e_i \cup f_i$. For such an algorithm it is a considerable challenge to bound the errors in the numerical simulations.

6 A simple Achlioptas process.

In this section we give a simple Achlioptas process that succeeds **whp** in creating a giant for any $c > \frac{3\sqrt{6}}{16}$.

Begin by fixing a set $S \in \binom{[n]}{\alpha n}$ for some $\alpha \in (0,1)$. During each round, choose only edges which are in $\binom{S}{2}$. If two such edges are presented, choose one at random. If no such edges are presented, choose neither.

In each round, the probability that we choose an edge from $\binom{S}{2}$ is $2\alpha^2 - \alpha^4$. Furthermore, in each round, any edge in $\binom{S}{2}$ is equally likely. Therefore, **whp** we will take $(2\alpha^2 - \alpha^4 + o(1))cn$ edges at random from $\binom{S}{2}$. So **whp** we have a giant component whenever

$$(2\alpha^2 - \alpha^4 + o(1))cn \quad > \quad \tfrac{1}{2}\alpha n. \tag{29}$$

To optimize (29), divide by αn and note that the left side is minimized when $\alpha^2 = \frac{2}{3}$. Therefore, we create a component of size $\Omega(n)$ **whp** whenever $c > \frac{3\sqrt{6}}{16} = 0.459\ldots$.

Acknowledgement. We thank the anonymous referee for comments and observations that improved the quality of the exposition of this paper. We also thank Shari Moskow for helpful discussions.

References

[1] D. Aldous, B. Pittel, On a random graph with immigrating vertices: Emergence of the giant component, *Random Structures and Algorithms* **17** (2000) 79-102.

[2] K.B. Athreya and P.E.Ney, *Branching Processes*, Springer-Verlag(1972).

[3] T. Bohman and A. Frieze, Avoiding a Giant Component, *Random Structures and Algorithms*, **19** (2001), 75–85.

[4] T. Bohman, A. Frieze and N. Wormald, Avoiding a Giant Component II, submitted.

[5] T. Bohman and J.H. Kim, A Phase Transition for Avoiding a Giant Component, submitted.

[6] B. Bollobás, *Random Graphs*, Second Edition, Academic Press(2001).

[7] P. Erdős and A. Rényi, On the Evolution of Random Graphs, *Publ. math. Inst. Hungar. Acad. Sci.*, **5** (1960), 17–61.

[8] S. Janson, T. Łuczak and A. Ruciński, *Random Graphs*, John Wiley & Sons, Inc.(2000).

[9] M. Molloy and B. Reed, A critical point for random graphs with a given degree sequence, *Random Structures and Algorithms* **6** (1995) 161-180.

[10] G. Sorkin, Embracing the giant component, manuscript.

[11] J. Spencer, Percolating Thoughts, manuscript dated January 31, 2001.

[12] J. Spencer and N. Wormald, Birth control for giants, manuscript.

[13] N. Wormald, The Differential Equation Method For Random Graph Processes and Greedy Algorithms, in *Lectures on Approximation and Randomized Algorithms*, (M. Karonski and H.J. Promel, eds) (1999) 73–155.