Avoiding a Giant Component

Tom Bohman* and Alan Frieze†

Department of Mathematical Sciences,
Carnegie Mellon University
Pittsburgh PA 15213.

August 2, 2001

Abstract

Let $e_1, e_1'; e_2, e_2'; \ldots; e_i, e_i'; \ldots$ be a sequence of ordered pairs of edges chosen uniformly at random from the edge set of the complete graph K_n (i.e. we sample with replacement). This sequence is used to form a graph by choosing at stage i, $i = 1, \ldots,$ one edge from e_i, e_i' to be an edge in the graph, where the choice at stage i is based only on the observation of the edges that have appeared by stage i. We show that these choices can be made so that **whp**[1] the size of the largest component of the graph formed at stage $.535n$ is polylogarithmic in n. This resolves a question of Achlioptas.

1 Introduction

Let $e_1, e_2, \ldots$ be a sequence of edges from the edge set of the complete graph K_n where e_i is chosen uniformly at random from the collection of edges that have not yet appeared. Let $E_m = \{e_1, e_2, \ldots, e_m\}$ and $G_m = ([n], E_m)$ be the mth random graph in this process. It is a classical theorem, due to Erdős and Rényi [3], that if $m = cn$, $c > 1/2$ a constant, then **whp** G_m contains a unique giant component of order $\Omega(n)$.

Achlioptas posed the following interesting question: Suppose that edges come in pairs $e_1, e_1'; e_2, e_2'; \ldots; e_i, e_i'; \ldots$ and we are allowed at stage i to choose one of e_i, e_i' to be edge in our graph. We must base our decision on $e_1, e_1', \ldots e_{i-1}, e_{i-1}'$ only; that is, the decision is made on-line. Does there exist an online algorithm A and a constant $c > 1/2$ such that the random graph $G_A(m)$ arising from the first $m = cn$ choices **whp** does not contain a giant component?

We describe a simple algorithm which proves that the answer to Achlioptas' question is yes.

*Supported in part by a grant from the NSA

†Supported in part by NSF grant CCR-9818411.

[1] A sequence of events $\mathcal{E}_n$ is said to occur with high probability (**whp**) if $\lim_{n\to\infty} \mathbf{Pr}(\mathcal{E}_n) = 1$

Theorem 1. *There is an algorithm A and positive constant $c_0 > .535$ such that* **whp** *the size of the largest component in $G_A = G_A(m), m = c_0 n$, is bounded by* $(\log n)^{O(1)}$.

Note that this says that a choice of two edges reduces the size of the largest component from order n to order polylog n, i.e. by an exponential factor. This is reminiscent of the beautiful result of Azar, Broder, Karlin and Upfal [1] where, in the framework of n balls in n boxes, the choice of one of two random boxes for each ball reduces the number of balls in the box containing the most balls from roughly $\frac{\log n}{\log\log n}$ to about $\log\log n$.

We prove Theorem 1 assuming edges are chosen with replacement. Since we choose less than $2n$ edges altogether, the probability that we do not replace any choice is bounded below by an absolute constant. Thus, the theorem remains true conditioning on this event, which extends the result to sampling without replacement.

We also prove the following converse to Theorem 1 which gives a limitation on our ability to avoid a giant component.

Theorem 2. *If $m = cn$, $c > 1$ constant, then* **whp** every *collection of m edges from $G_{n,2m}$ gives a graph with a component of size at least $\epsilon_c n$ where*

$$\epsilon_c = \frac{1}{2}(2^c e^{c+1})^{-1/(c-1)}.$$

It follows that any algorithm (including any off-line algorithm) that chooses one of two random edges through cn steps where $c > 1$ must produce a subgraph with a component of size at least $\epsilon_c n$.

The Algorithm A that we introduce to prove Theorem 1 is very simple: it takes e_t unless e_t' is disjoint from all previously examined edges. This introduces a small bias in favor of choosing isolated edges and against merging larger components. We shall see that this is enough to give the theorem. Let

$$m = \alpha n = \lfloor .535n \rfloor$$

(thus, $\alpha \approx .535$) and

$$V_t = [n] \setminus \bigcup_{i=1}^{t-1} (e_i \cup e_i').$$

In words, V_t is the set of vertices not incident with any edge in the first $t-1$ pairs of edges.

```
Algorithm $A$
begin
```

```
    For $t = 1$ to $m$ do
        $f_t := \begin{cases} e_t & e'_t \nsubseteq V_t & \text{Blue edge} \\ e'_t & e'_t \subseteq V_t & \text{Red edge} \end{cases}$
    $E(G_A) \leftarrow \{f_1, \dots, f_m\}$
end.
```

In the analysis of the algorithm it will be useful to make a distinction between chosen edges of the form $f_t = e'_t$ (i.e. edges that are chosen because they are disjoint from all previously examined edges) and chosen edges of the form $f_t = e_t$ (i.e. edges which are chosen because their partner is incident with some previously examined edges). We call the former *Red* edges and the latter *Blue* edges.

The remainder of the paper is organized as follows. In then next section we reduce the proof of Theorem 1 to 2 technical lemmas, which are proved in Sections 4 and 5. Section 3 is a brief discussion of some of the many open questions which follow naturally from this work. Finally, we prove Theorem 2 in Section 6.

2 Main Argument

We begin with some notational conventions. Let E_A be the collection of edges chosen; to be precise, $E_A = \cup_{i=1}^{m} f_i$. Let $G_A = ([n], E_A)$. For a tree T let $b(T)$ denote the number of branching nodes of T (i.e. nodes of degree 3 or more). If T has $|T|$ nodes and $\epsilon > 0$ is a positive constant then we say that T is *ϵ-bushy* if $b(T) \leq \epsilon |T|$. For positive integers K and n let $I_{K,n}$ be the interval

$$I_{K,n} = [K \log n, 2K \log n].$$

Our theorem follows from two lemmas:

Lemma 1. *Let G be a graph with vertex set $[n]$ which has maximum degree, $\Delta(G)$, at most $\log n$. Let K and ϵ be fixed positive constants and set $\xi = \lceil 2/\epsilon \rceil$. If n is sufficiently large and G contains no ϵ-bushy tree T for which $|T| \in I_{K,n}$ then G has no component with more than $k_0 = (\log n)^{\xi+1}$ vertices.*

Lemma 2. *There exist constants $K, \epsilon > 0$ such that* **whp** *G_A contains no ϵ-bushy tree T such that $|T| \in I_{K,n}$.*

Now, the random graph G_A is a subgraph of the random graph $G_{n,2n}$ (the graph produced by the first n pairs). It is well known (see for example Bollobás [2], Chapter III) that **whp** $\Delta(G_{n,2n}) \leq \log n$. Hence, Theorem 1 follows immediately from Lemma 1 and 2.

The proof of Lemma 1 is constructive. We prove Lemma 2 by conditioning on the times and colors of chosen edges. Consider the event E that a fixed pair of *incident* edges appears in $f_1, \ldots, f_m$. Note that at least one of these edges must be Blue (the set of Red edges forms a matching). Furthermore, if one of the edges is Red and the other is Blue then the Red edge must appear before the Blue edge, that is, the set of pairs of times for the occurrence of these edges is reduced by half. It follows from these observations that the probability of E is less than $1/n^2$. This is, very loosely speaking, the central idea of the proof.

Lemmas 1 and 2 are proved in Sections 4 and 5, respectively.

3 Open questions

Before turning to the proofs of Lemma 1 and 2, we list and discuss some questions which follow naturally from Theorem 1.

There are a number of ways in which Theorem 1 might be extended or applied:

- What is the maximum value of c_0 for which the theorem remains valid? It seems to us that .535 is not the maximum.
- How well can an off-line algorithm (i.e. an algorithm that chooses one of each e_t, e'_t *after* it has seen all edges) do in avoiding a giant component?
- We can also consider the converse problem. Is it possible to create a giant component with $(\frac{1}{2} - \epsilon)n$ edges if we have a choice of two at each stage?
- Are there any algorithmic implications, as there are for [1]?

Many interesting questions analogous to the original conjecture of Achlioptas can be asked; that is, for any graph property we can ask if the choice of one of two random edges at each stage allows us to delay (or advance) the appearance of the property. We can also adapt the question to other random structures. At first glance, we see no reason why statements analogous to Theorem 1 can not be proven for other properties. It seems more challenging to show that there are non-trivial graph properties for which the threshold is robust under the introduction of an on-line choice of one of two random edges at each stage. We suspect that the size of the components in $G_{n,p}$ for p near $1/n$ will not provide such examples; that is, we suspect that the answer to the following question is yes:

- Does there exists an algorithm A and a constant $\beta > 1/2$ such that size of the largest component of $G_A(\beta n)$ is $O(\log n)$ (rather than $O(poly(\log n))$). In other words, can we keep components of size $(\log n)^2$ from occurring until significantly after the $(n/2)^{\text{nd}}$ edge?

In fact, it may be the case that components of size even small than $\log n$ can be delayed past $p = 1/n$.

- Does there exists an algorithm A and a constant $\beta > 1/2$ such that size of the largest component of $G_A(\beta n)$ is $o(\log n)$?

4 Proof of Lemma 1

Let G be a graph with vertex set $[n]$ such that $\Delta(G) \leq \log n$ which contains no ϵ–bushy tree T such that $|T| \in I_{K,n}$. Assume for the sake of contradiction that G contains a component with more than k_0 vertices. Let T_1 be a subtree of this component such that $|T_1| = \lceil (\log n)^{\xi+1} \rceil$. We show that T_1 contains an ϵ-bushy tree $T^\star$ with $|T^\star| \in I_{K,n}$. We achieve this by constructing two sequences of trees:

$$T_1 \supseteq T_2 \supseteq \cdots \supseteq T_p$$

and

$$\hat{T}_1 \subseteq \hat{T}_2 \subseteq \cdots \subseteq \hat{T}_p$$

such that

(i) $\hat{T}_i$ is a subtree of T_i for each i,

(ii) $\hat{T}_i$ is ϵ–bushy for each i, and

(iii) $|\hat{T}_p| \in I_{K,n}$.

The idea is to 'trim' subtrees of T_i that do not contain any long paths while introducing a long path (length at least ξ) to $\hat{T}_i$ every time a new branch vertex is introduced to $\hat{T}_i$.

At each stage of this process some vertices are deleted and some vertices will be *chosen*. Vertices deleted at stage i do not appear in T_{i+1}. The vertices chosen during stage i are the vertices in $\hat{T}_{i+1} \setminus \hat{T}_i$. In other words, chosen vertices will be vertices in $T^\star = \hat{T}_p$. The collection of vertices that have not yet been either chosen or deleted will be either *touched* or *untouched*.

Initially all vertices are untouched and $P_1 = x_1, x_2, \ldots, x_k$ is a maximal path of T_1. We may assume P_1 has fewer than $K \log n$ vertices (n.b. if T_1 contains a path of length $K \log n$ then we simply take this path as $T^\star$). The vertices of P_1 become chosen. Fix $1 \leq j \leq k$ and let E_j denote the set of edges which are incident with x_j but are not part of P_1. Consider $e = \{x_j, y\} \in E_j$ and the subtree T' containing y which is connected to the rest of T_1 by e. Let P' be a longest path in T' which has y as one endpoint. If P' has fewer than ξ edges than we delete the whole of T'. Otherwise the vertices of P' become touched. We do this for every j and every edge of E_j. The paths that comprise the collection of touched vertices become chosen one path at a time. If at any stage in this process the number of chosen vertices lies in $I_{K,n}$ then we stop and the collection of chosen vertices is the desired ϵ-bushy tree $T^{\star}$. If the process does not terminate then the subtree of vertices that have not been deleted is denoted by T_2.

In general, after $j-1$ steps we have a tree T_j with a subtree $\hat{T}_j$ of chosen vertices, and the remaining vertices are all untouched. For each chosen vertex v we consider the set of edges E_v which join it to the untouched vertices (note that we need only consider the vertices that were chosen in the previous round). Each $e \in E_v$ defines a tree which is deleted if it has no long path to its root and otherwise it produces a path which becomes touched. At the end of the round the paths that comprise the collection of touched vertices become chosen one path at a time, and we stop the process if the number of chosen vertices lies in $I_{K,n}$.

We must show that the process terminates before all vertices are either chosen or deleted. We can assume that we never see a path of length $K \log n$ or more, otherwise we can use (part of) this path as our ϵ-bushy tree. Thus, if the process fails to terminate then the collection of chosen vertices never exceeds $K \log n$. Now, the deleted vertices make up a collection of trees which can be rooted at a chosen vertex and have depth at most ξ. Thus the number of deleted vertices is at most $\Delta + \Delta^2 + \cdots + \Delta^{\xi-1}$ times the number of chosen vertices. If the process did not terminate properly then T_p would be an ϵ-bushy tree and would satisfy

$$|T_p| \geq \frac{|T_1|}{1+\Delta+\cdots+\Delta^{\xi-1}} \geq \frac{|T_1|}{2\Delta^{\xi-1}} > K \log n,$$

for n sufficiently large. This is a contradiction.

It only remains to prove Lemma 2.

5 Proof of Lemma 2

Recall that the chosen edge f_t, $t = 1, 2, \ldots, m$ is assigned the colour Red if $f_t = e'_t$ and is assigned Blue otherwise. By this colouring E_A is partitioned into E_B, E_R, the sets of Blue and Red edges respectively. We prove Lemma 2 by conditioning on the times when edges appear and the colors they are assigned.

We begin by defining a generic event $\mathcal{E}$. Consider collections of edges B_0, B_1, R, Q which satisfy

(i) R is a matching.

(ii) $B_1 \cup R$ can be decomposed into $|R|$ edge-disjoint paths of length 2, each containing one edge from B_1 and one edge from R.

(iii) $0 \leq |Q| \leq 3$.

Let $B = B_0 \cup B_1$ and let $\mathcal{E} = \mathcal{E}(B, R, Q)$ be the event that we have

$$B \subseteq E_B, \quad R \subseteq E_R, \quad Q \subseteq E_A.$$

Condition (i) is motivated by the fact that the set of Red edges always forms a matching. Condition (ii) is motivated by the fact that, since Red edges appear as isolated edges, any edge in G_A incident with a Red edge is Blue.

While it is not a formal necessity of this definition, whenever we consider the event $\mathcal{E}(B, R, Q)$ the edge set $B \cup R \cup Q$ forms a connected component (in fact, this edge set is usually a tree and on rare occasion a unicyclic connected component). Suppose we have a set of Red and Blue edges that forms a connected component. We determine sets B_0, B_1, R, Q that satisfy conditions (i)-(iii) as follows. Choose a spanning tree T of the connected component and an arbitrary root of T. Form paths consisting of one Blue and one Red edge by 'matching' each Red edge to the first edge on the path from that Red edge to the root. This procedure fails to 'match' any Red edge incident with the root. The set Q holds this 'unmatched' Red edge, when it exists.

We now bound the probability of such an event. Let $s = |B| + |R|, q = |Q|$.

Lemma 3. *If* $s \leq (\log n)^2$ *then*

$$\mathbf{Pr}(\mathcal{E}) \leq (1 + o(1)) \frac{2^{s+q}}{n^{s+q}} \left(\alpha - \frac{1}{8}(1 - e^{-8\alpha}) \right)^{|B|-|R|} \times \left(\frac{1}{128}(16\alpha + 4e^{-8\alpha} - (3 + e^{-16\alpha})) \right)^{|R|}.$$

Proof We prove the Lemma by conditioning on the times when the given edges appear in our random process. Fix a collection of times $\mathbf{t} = \{t_f : f \in R \cup B \cup Q\}$ such that every pair of adjacent edges f_1, f_2 with $f_1 \in B \cup Q$ and $f_2 \in R$ we have $t_{f_1} > t_{f_2}$ (we will use this property only for the pairs given by condition (ii)). Suppose that $\{t_f : f \in S \cup Q\} = \{\tau_1 < \tau_2 < \cdots < \tau_{s+q}\}$. Define events

$$\mathcal{A}_i = \begin{cases} \{e_{t_f} = f, e'_{t_f} \not\subseteq V_{t_f}\} & \text{if } \tau_i = t_f,\ f \in B \\ \{e'_{t_f} = f, f \subseteq V_{t_f}\} & \text{if } \tau_i = t_f,\ f \in R \\ \{f \in \{e_{t_f}, e'_{t_f}\}\} & \text{if } \tau_i = t_f,\ f \in Q \end{cases}$$

Then we let

$$\mathcal{A}_{\mathbf{t}} = \bigcap_{i=1}^{s+q} \mathcal{A}_i.$$

Now let us estimate

$$\alpha_i = \mathbf{Pr}(\mathcal{A}_i \mid \mathcal{A}_1, \dots, \mathcal{A}_{i-1}) \tag{1}$$

for $i = 1, 2, \dots, s + q$. Note that the conditioning influences the randomness of the edges unfixed edges in $e_1, e'_1; \dots; e_{\tau_i - 1}, e'_{\tau_i - 1}$ in two ways:

(a) If the edge f is in R and $j \leq t_f < \tau_i$ then neither e_j nor e'_j is incident with f.

(b) If f is in B and $t_f < \tau_i$ then there exists a *first* edge e''_{t_f} that appears before time t_f and is incident with e'_{t_f}. In other words, the conditioning may force a number of paths on 2 edges in the set $e_1, e'_1; \ldots; e_{\tau_i-1}, e'_{\tau_i-1}$.

All told, very few edges in $e_1, e'_1; \ldots; e_{\tau_i-1}, e'_{\tau_i-1}$ are fixed and those that are not fixed are distributed nearly uniformly.

Now, we bound α_i by further conditioning on the edges e'_{t_f} and e''_{t_f} for $f \in B$ and the time of appearance of the edge e''_{t_f}, which we denote t''_f. With this additional conditioning in place the edge e_j or e'_j that is not fixed by the conditioning is chosen uniformly at random from the set of edges that does not intersect

$$U_j := \left(\bigcup_{f \in R: j < t_f < \tau_i} f \right) \cup \left(\bigcup_{f \in B: t_f < \tau_i \text{ and } j < t''_f} e'_{t_f} \right).$$

Note that the size of this vertex set is at most $2s$; that is, the edge is chosen uniformly at random from a collection of at least $\binom{n-2s}{2}$ edges. Also note that we have $U_1 \supseteq U_2 \supseteq \cdots \supseteq U_{\tau_i-1}$. We are now ready to bound α_i.

Suppose first that $\tau_i = t_f$, $f \in Q$, then

$$\alpha_i \le \frac{2}{N} \tag{2}$$

since the edges are chosen independently with replacement.

Suppose now that $\tau_i = t_f$, $f \in B$. Clearly, $\mathbf{Pr}(e_{t_f} = f) = 1/N$. We must multiply this by the probability that e'_{t_f} intersects some previously appearing edge (note that these two events are independent). The probability that e'_{t_f} intersects one of the fixed edges is at most $6sn/\binom{n}{2}$ (since there are at most $3s$ fixed edges). If e'_{t_f} intersects none of the fixed edges then e'_{t_f} does not intersect U_0, and the probability e'_{t_f} is incident with any given unfixed edge is at most $2n/\binom{n-2s}{2}$. Putting these observations together we have

$$\alpha_i \le \frac{1}{N}\left[O(s/n) + \left(1 - \left(1 - \frac{4n}{n^2 - O(sn)} \right)^{2t_f} \right) \right] = \frac{1}{N}\left(1 - e^{-8t_f/n} + O(s/n) \right). \tag{3}$$

Finally, assume that $\tau_i = t_f$, $f \in R$. By reasoning analogous to that for the previous case we have

$$\alpha_i \le \frac{1}{N}\left(1 - \frac{4n}{n^2 - O(sn)} \right)^{2t_f - O(s)} = \frac{1}{N}\left(e^{-8t_f/n} + O(s/n) \right). \tag{4}$$

Applying (2), (3) and (4) we see that

$$\mathbf{Pr}(\mathcal{A}_t) \le \left(\frac{2^q}{N^{s+q}} \prod_{f \in B} (1 - e^{-8t_f/n} + O(s/n)) \prod_{f \in R} (e^{-8t_f/n} + O(s/n) \right).$$

Now let $\mathcal{T}$ denote the set of feasible choices for $\mathbf{t}$. So

$$\begin{aligned}\mathbf{Pr}(\mathcal{E}) &\le \frac{2^q}{N^{s+q}}\left(\sum_{\mathbf{t}\in\mathcal{T}}\prod_{f\in B}(1-e^{-8t_f/n}+O(s/n))\prod_{f\in R}(e^{-8t_f/n}+O(s/n)\right)\\ &\le \frac{2^q}{N^{s+q}}\left(\sum_{i=1}^{m}(1-e^{-8i/n}+O(s/n))\right)^{|B|-|R|}\\ &\qquad\times\left(\sum_{i=1}^{m-1}\sum_{j=i+1}^{m}(e^{-8i/n}+O(s/n))(1-e^{-8j/n}+O(s/n))\right)^{|R|}.\end{aligned}$$

Now

$$\sum_{i=1}^{m}e^{-8i/n}\le 1+\int_{x=0}^{m}e^{-8x/n}dx=1+\frac{n}{8}(1-e^{-8m/n})$$

and

$$\begin{aligned}\sum_{i=1}^{m-1}\sum_{j=i+1}^{m}e^{-8i/n}(1-e^{-8j/n}) &\le m+\int_{x=0}^{m}\int_{y=x}^{m}e^{-8x/n}(1-e^{-8y/n})dydx\\ &=m+\int_{x=0}^{m}e^{-8x/n}\left(m-x+\frac{n}{8}(e^{-8m/n}-e^{-8x/n})\right)dx\\ &=m+\frac{mn}{8}+\frac{n^2}{32}e^{-8m/n}-\frac{n^2}{128}(3+e^{-16m/n})\end{aligned}$$

and the lemma follows. □

We also need a bound the number of ϵ-bushy trees on a fixed set. Let $\mathcal{B}_\epsilon$ denote the set of ϵ-bushy trees spanning the set $[k]$.

Lemma 4.

$$|\mathcal{B}_\epsilon|\le k!e^{\phi(\epsilon)k}$$

where $\phi(\epsilon)\to 0$ *as* $\epsilon\to 0$.

Proof We use the well-known Prüfer correspondence between labelled trees on k vertices and sequences of length $k-2$ over $[k]$, (see for example Lovász [4], Problem 4.5). Suppose that we fix $b(T)=t$ and the set of t branching nodes and the degrees $b_1,b_2,\dots,b_t$ of these nodes. If $b_1+b_2+\cdots+b_t=B+t$ then we have at most

$$\binom{k-t}{k-2-B}\binom{k-2}{B}(k-2-B)!\frac{B!}{(b_1-1)!(b_2-1)!\cdots(b_t-1)!}=$$
$$\binom{k-t}{k-2-B}\frac{(k-2)!}{(b_1-1)!(b_2-1)!\cdots(b_t-1)!}$$

choices for the tree.

Thus the total number of ϵ-bushy trees is at most

$$(k-2)!\sum_{t=0}^{\epsilon k}\binom{k}{t}\sum_{B\geq 2t}\binom{k-t}{k-2-B}\sum_{b_1+\cdots+b_t=B+t}\frac{1}{(b_1-1)!(b_2-1)!\cdots(b_t-1)!}$$
$$=(k-2)!\sum_{t=0}^{\epsilon k}\binom{k}{t}\sum_{B\geq 2t}\binom{k-t}{k-2-B}[x^B]e^{tx}$$
$$=(k-2)!\sum_{t=0}^{\epsilon k}\binom{k}{t}\sum_{B\geq 2t}\binom{k-t}{k-2-B}\frac{t^B}{B!}.$$

Now let $u_B=\binom{k-t}{k-2-B}\frac{t^B}{B!}$. If $B\geq \epsilon^{1/2}k$ then

$$\frac{t^B}{B!}\leq\left(\frac{te}{B}\right)^B\leq\left(\frac{e\epsilon k}{B}\right)^B\leq(e\epsilon^{1/2})^B.$$

So

$$\sum_{B\geq\epsilon^{1/2}k}u_B\leq(e\epsilon^{1/2})^{t-2}\sum_{B\geq 2t}\binom{k-t}{B+2-t}(e\epsilon^{1/2})^{B+2-t}$$
$$\leq(e\epsilon^{1/2})^{t-2}(1+e\epsilon^{1/2})^{k-t}=e^{O(\epsilon^{1/2}k)}. \quad (5)$$

On the other hand

$$\sum_{B\leq\epsilon^{1/2}k}u_B\leq\binom{k-t}{k-2-\epsilon^{1/2}k}\sum_{B\leq\epsilon^{1/2}k}\frac{t^B}{B!}$$
$$\leq\binom{k-t}{k-2-\epsilon^{1/2}k}e^{\epsilon k} \quad (6)$$
$$\leq e^{O((\epsilon^{1/2}\log(1/\epsilon))k)}.$$

So, (5) and (6) imply that the total number of ϵ-bushy trees is at most

$$(k-2)!e^{O((\epsilon^{1/2}\log(1/\epsilon))k)}\sum_{t=0}^{\epsilon k}\binom{k}{t}$$

and the lemma follows. □

We now prove Lemma 2. Let T, $|T|=k\in I_{K,n}$ be an ϵ-bushy tree. We choose a root r of T and let $\mathcal{M}$ denote the set of matchings of T. We have

$$Pr(T\subseteq G_A)=\sum_{M\in\mathcal{M}}Pr(M\subseteq E_R,E(T)\setminus M\subseteq E_B).$$

Note that in this event, for a fixed matching M, every edge incident with an edge in M must occur in E_B. For each $M\in\mathcal{M}$ let Q denote the set of edges of M

which are incident with r (of course $|Q| \leq 1$). Form paths of length two consisting of one edge from E_B and one edge from E_R by associating each edge $e \in M \setminus Q$ with the first edge in the path from e to the root r. We apply Lemma 3:

$$\begin{aligned}
\mathbf{Pr}(T \subseteq G_A) &\leq \sum_{M \in \mathcal{M}} \mathbf{Pr}(M \setminus Q \subseteq E_R, E(T) \setminus M \subseteq E_B,\ Q \subseteq E_A) \\
&\leq (1+o(1)) \frac{2^{k-1}}{n^{k-1}} \left(\alpha - \frac{1}{8}(1 - e^{-8\alpha}) \right)^{k-1} \qquad (7) \\
&\quad \times \sum_{M \in \mathcal{M}} \left(\frac{16\alpha + 4e^{-8\alpha} - (3 + e^{-16\alpha})}{128 \left(\alpha - \frac{1}{8}(1 - e^{-8\alpha}) \right)^2} \right)^{|M|-1}
\end{aligned}$$

We observe that the number of l-edge matchings in a tree T on k vertices is at most the number of l-edge matchings in a path on k vertices. Indeed, let $m_{k,l}$ be the number of l-edge matchings on a path with k vertices and let $m^*_{k,l}$ be the maximum over all trees T with k vertices of the number of l-edge matchings in T. Then for $k \geq 2$,

$$m_{k,\ell} = m_{k-2,\ell-1} + m_{k-1,\ell} \text{ and } m^*_{k,\ell} \leq m^*_{k-2,\ell-1} + m^*_{k-1,\ell}.$$

The equality comes from considering whether or not an l-matching contains a particular end edge of a path with k vertices. The inequality comes from considering, in a fixed tree T, whether or not an l-matching contains a particular pendant edge $\{x, y\}$ such that $T - \{x, y\}$ leaves a tree plus isolated vertices. Here we have to use the fact that $m^*_{k,l}$ is monotone non-decreasing with k. A simple induction then shows that, in fact, $m^*_{k,l} = m_{k,l}$.

Now, if we let

$$\psi(x, y) = \sum_{k=0}^{\infty} \sum_{l=0}^{k} m_{k,l} x^k y^l$$

where $m_{k,l}$ is the number of l-edge matchings on a path with k vertices, then

$$\begin{aligned}
\psi(x, y) = \frac{1}{1 - x - x^2 y} = \\
\frac{1}{\sqrt{1+4y}} \sum_{k=0}^{\infty} \left(\left(\frac{1 + \sqrt{1+4y}}{2} \right)^{k+1} - \left(\frac{1 - \sqrt{1+4y}}{2} \right)^{k+1} \right) x^k.
\end{aligned}$$

Indeed, we have the recurrence $m_{k,\ell} = m_{k-2,\ell-1} + m_{k-1,\ell}$ for $k \geq 2$, and we get the claimed expression for $\psi(x, y)$ by multiplying each such equation by $x^k y^\ell$ and summing in the usual way.

This implies that for any k vertex tree and any $y > 0$,

$$\sum_{M \in \mathcal{M}} y^{|M|} \leq \sum_{\ell=0}^{k} m_{k,\ell} y^\ell \leq 2 \left(\frac{1 + \sqrt{1+4y}}{2} \right)^k.$$

Plugging this into (7) we get

$$\mathbf{Pr}(T \subseteq G_A) =$$
$$O\left(\frac{1}{n^{k-1}}\left(\alpha - \frac{1}{8}(1-e^{-8\alpha})\right)^k \left(1+\sqrt{1+\frac{16\alpha+4e^{-8\alpha}-(3+e^{-16\alpha})}{32\left(\alpha-\frac{1}{8}(1-e^{-8\alpha})\right)^2}}\right)^k\right)$$

Thus, by Lemma 3, the probability that G_A contains an ϵ-bushy tree with $k \in I$ vertices is

$$O\left(\binom{n}{k} k! e^{\phi(\epsilon)k}\frac{1}{n^{k-1}}\left(\alpha - \frac{1}{8}(1-e^{-8\alpha})\right)^k\right.$$
$$\left.\times\left(1+\sqrt{1+\frac{16\alpha+4e^{-8\alpha}-(3+e^{-16\alpha})}{32\left(\alpha-\frac{1}{8}(1-e^{-8\alpha})\right)^2}}\right)^k\right)$$

Putting $\alpha \approx .535$ and ϵ sufficiently small, we see that

$$\mathbf{Pr}(T \subseteq G_A) = O(n\zeta^k)$$

where $\zeta < 1$ is constant. Thus **whp** G_A contains no ϵ-bushy tree of size k in the range $I_{K,n}$ for $K = 2\log\zeta^{-1}$ and Lemma 2 follows immediately.

Remark This argument can be used to show that **whp** all the components of G_A are trees or unicyclic. Consider a minimal complex connected graph. This consists of a path $P = (x_1, x_2, \ldots, x_k)$ plus two edges contained in $\{x_1, x_2, \ldots, x_k\}$. Following the above argument we see that for $k \in I_{K,n}$, the probability of such a subgraph is $O(n\zeta^k n^{-2}) = o(1)$ where the two extra edges are placed in Q, when we apply Lemma 3. This accounts for the extra factor n^{-2}. For $k > 2K\log n$, Lemma 2 rules out the existence of P.

6 Proof of Theorem 2

Consider the random graph $G = G_{n,p}$, $p = 4c/n$. We first prove that **whp**

$$\not\exists S \subseteq [n], s = |S| \leq \epsilon_c n \text{ such that } S \text{ contains at least } cs \text{ edges of } G. \qquad (8)$$

Indeed, the probability that there exists an S violating the density condition of (8) is at most

$$\sum_{s=4}^{\epsilon_c n}\binom{n}{s}\binom{\binom{s}{2}}{cs}\left(\frac{4c}{n}\right)^{cs} \leq \sum_{s=4}^{\epsilon_c n}\left(\frac{ne}{s}\left(\frac{es}{2c}\cdot\frac{4c}{n}\right)^c\right)^s$$
$$= \sum_{s=4}^{\epsilon_c n}\left(2^c e^{c+1}\left(\frac{s}{n}\right)^{c-1}\right)^s = o(1).$$

Note that, by monotonicity, we can replace $G_{n,p}$ by $G_{n,2m}$, $m = cn$ in the conclusion of (8).

Now, suppose that there are no sets violating the density condition of (8) and assume for the sake of contradiction that $G = G_{n,2m}$ contains a set of m edges X such that the edge induced subgraph $G(X)$ has no component of size $\epsilon_c n$ or more. Let the components of $G(X)$ be $C_1, C_2, \ldots, C_\rho$. Then the number of edges in $G(X)$ is less than

$$\sum_{i=1}^{\rho} c|C_i| = cn$$

contradiction. □

Acknowledgement We thank Dimitris Achlioptas for thinking of this lovely problem and Mike Molloy for some earlier discussions.

References

[1] Y. Azar, A. Z. Broder, A. R. Karlin and E. Upfal, *Balanced Allocations*, Proceedings of the 26th Annual ACM Symposium on the Theory of Computing, (1994) 593–602.

[2] B. Bollobás, *Random Graphs*, Academic Press 1985.

[3] P. Erdős and A. Rényi, *On the evolution of random graphs*, Publ. Math. Inst. Hungar. Acad. Sci. 5 (1960) 17-61.

[4] L. Lovász, *Combinatorial problems and exercises*, North-Holland 1993.

This research was sponsored in part by National Science Foundation (NSF) grant no. CCR-0122581.