Critical random graphs and the structure of a minimum spanning tree

L. Addario-Berry* N. Broutin*† B. Reed*

May 11, 2008

Abstract

We consider the complete graph on n vertices whose edges are weighted by independent and identically distributed edge weights and build the associated minimum weight spanning tree. We show that if the random weights are all distinct, then the expected diameter of such a tree is $\Theta(n^{1/3})$. This settles a question of Frieze and McDiarmid [15]. The proofs are based on a precise analysis of the behaviour of random graphs around the critical point.

Keywords: Minimum Spanning Trees, Random Graphs, Random Walks, Random Trees

1 Introduction

Given a connected graph $G = (V, E)$, $E = \{e_1, \ldots, e_{|E|}\}$, together with edge weights $W = \{w(e) : e \in E\}$, a *minimum weight spanning tree* of G is a spanning tree $T = (V, E')$ that minimizes

$$\sum_{e \in E'} w(e).$$

As we show below, if the edge weights are distinct then this tree is unique; in this case we denote it by $MWST(G, W)$ or simply $MWST(G)$ when W is clear.

Minimum spanning trees are at the heart of many combinatorial optimization problems. In particular, they are easy to compute [5, 21, 22, 30], and may be used to approximate hard problems such as the minimum weight traveling salesman tour [35]. (A complete account on the history of the minimum spanning tree problem may be found in the surveys of Graham and Hell [17], and Nešetřil [27].) As a consequence, much attention has been given to studying their structure, especially in random settings and under various models of randomness. For instance, Frieze [14] determined the weight of the $MWST$ of a complete graph whose edges have been weighted by independent and identically distributed (i.i.d.) $[0, 1]$-random variables. This result has been reproved and generalized by Frieze and McDiarmid [16], Aldous [1], Steele [34] and Fill and Steele [11]. Under the same model, Aldous [1] derived the degree distribution of the $MWST$. These results concern *local* properties of minimum spanning trees. Questions concerning the *global* structure of the minimum spanning tree remain mostly untouched.

*School of Computer Science, McGill University, Montreal, Canada. Email: louigi@gmail.com, nbrout@cs.mcgill.ca, breed@cs.mcgill.ca

†Projet Algorithms, INRIA Rocquencourt, 78153 Le Chesnay, France

The *distance* between vertices x and y in a graph H is the length of the shortest path from x to y. The *diameter* $diam(H)$ of a connected graph H is the greatest distance between any two vertices in H. We are interested in the diameters of the minimum weight spanning trees of a clique K_n on n vertices whose edges have been assigned i.i.d. real weights. We use $w(e)$ to denote the weight of edge e. In this paper we prove the following theorem, answering a question of Frieze and McDiarmid [15], Research Problem 23:

Theorem 1. *Let $K_n = (V, E)$ be the complete graph on n vertices, and let $\{X_e : e \in E\}$ be independent identically distributed edge-weights. Then conditional upon the event that $X_e \neq X_f$ for distinct edges e and f, it is the case that the expected value of the diameter of $MWST(K_n)$ is $\Theta(n^{1/3})$.*

In the remainder of this section, we give some general properties of minimum spanning trees and explain informally the intuition behind Theorem 1. Let T be some minimum weight spanning tree of G. If e is not in T then the path between its endpoints in T consists only of edges with weight at most $w(e)$. If $e = xy$ is in T then every edge f between the component of $T - e$ containing x and the component of $T - e$ containing y has weight at least $w(e)$, since $T - e + f$ is also a spanning tree. Thus, if the edge weights are distinct, e is in T precisely if its endpoints are in different components of the subgraph of G with edge set $\{f : w(f) < w(e)\}$. It follows that if the edge weights are distinct, $T = MWST(G)$ is unique and the following greedy algorithm generates $MWST(G)$:

(1) Order E as $\{e_1, \ldots, e_m\}$ so that $w(e_i) < w(e_{i+1})$ for $i = 1, 2, ..., m - 1$.

(2) Let $E_T = \emptyset$, and for i increasing from 1 to m, add edge e_i to E_T unless doing so would create a cycle in the graph (V, E_T). The resulting graph (V, E_T) is the unique $MWST$ of G.

Kruskal's algorithm [22] above lies at the heart of the proof of Theorem 1. It provides a way to grow the minimum spanning tree that is perfectly suited to keeping track of the evolution of the diameter of E_T as the edges are processed. We now turn our attention to this forest growing process and review its useful properties.

Observe first that, if the weights $w(e)$ are distinct, one does not need to know $\{w(e), e \in E\}$ to determine $MWST(G)$, but merely the ordering of E in (1) above. If the $w(e)$ are i.i.d. random variables, then conditioning on the weights being distinct, this ordering is a random permutation. Thus, for any i.i.d. random edge weights, conditional upon all edge weights being distinct, the distribution of $MWST(G)$ is the same as that obtained by weighting E according to a uniformly random permutation of $\{1, \ldots, m\}$.

This provides a natural link between Kruskal's algorithm and the $G_{n,m}$ random graph evolution process of Erdős and Rényi [10]. This well-known process consists of an increasing sequence of $|E| = \binom{n}{2}$ random subgraphs of K_n defined as follows. Choose a uniformly random permutation $e_1, \ldots, e_{|E|}$ of the edges, and set $G_{n,m}$ to be the subgraph of K_n with edge set $\{e_1, \ldots, e_m\}$. If we let e_i have weight i, $1 \leq i \leq \binom{n}{2}$, then $e_m \in MWST(K_n)$ precisely if e_m is a cutedge of $G_{n,m}$. It is well known that the random graph process is easier analyzed via the model $G_{n,p}$ where each edge is present with probability p, independently of the others [3, 20]. This is mostly because $G_{n,p}$ is amenable to an analysis based on branching processes.

The graph on n vertices and with edge set $\{e : w(e) \leq p\}$ is distributed as $G_{n,p}$, and we now consider this particular coupling between the Kruskal construction and the random graph process. Recall that $e \in MWST(G)$ precisely if e is a cutedge of $G_{n,w(e)}$, hence the

component structure of the forest $F_{n,p} = MWST(K_n) \cap \{e : w(e) \leq p\}$ built by Kruskal's algorithm corresponds to that of $G_{n,p}$: for every $p \in [0,1]$, the connected components of $F_{n,p}$ and $G_{n,p}$ have the same vertex sets. Observe that the minimum spanning tree is nothing else than $F_{n,1}$. The evolution of the diameter of the minimum spanning forest $F_{n,p}$ exhibits three distinct phases that reflect the subcritical, critical and supercritical phases of the random graph $G_{n,p}$. These three phases correspond to $p \sim c/n$ for c lower than, equal to, or larger than one, respectively [3, 10, 20]. In the following a.a.s. stands for *asymptotically almost surely*, i.e., with probability tending to one as $n \to \infty$. We first discuss the subcritical and supercritical phases, as the effect of these phases on the expected diameter of $MWST(K_n)$ is small and easily analyzed.

THE SUBCRITICAL PHASE. For $p \sim c/n$, $c < 1$, $G_{n,p}$ consists a.a.s. of small components of size at most $O(\log n)$. In this phase, the diameter of the minimum spanning forest is at most $O(\log n)$.

THE SUPERCRITICAL PHASE. For $p \sim c/n$, $c > 1$, a.a.s.,

($\star$) the largest component $H_{n,p}$ of $G_{n,p}$ has size $\Theta(n)$ and all the other have size $O(\log n)$.

In this phase, $H_{n,p}$ is called the *giant component*. More precisely, a.a.s. ($\star$) holds for *all* $p > c/n$. This implies that a.a.s., for all $p' > p > (1+\epsilon)/n$, $H_{n,p} \subseteq H_{n,p'}$. Restating this in colourful but imprecise language, which component is the largest never changes, and simply expands by gobbling up smaller ones. It turns out that in this last phase, the diameter of the minimum spanning forest does not substantially change. To understand why, observe that every edge added is uniform among those which do not create a cycle. In other words, the places where the small components hook up to the giant are uniform, and we can think of the *increase* in the diameter as roughly explained by a simplified process similar to the construction of a *random recursive tree*: one starts with a single metanode (representing the component $H_{n,c/n}$ for any fixed $c > 1$); then further metanodes (representing the small components at the moment they connect to the component containing $H_{n,c/n}$) are added one at a time by choosing a uniformly random point of connection. (An intermediate stage of this such a process is depicted in Figure 1.) Such a tree on n metanodes has diameter $O(\log n)$ [9, 29, 33]. The discrepancy between this idealized process and the real one relies in that the merging process involves components that are of order greater than one, which biases the process. Second, the diameter in the actual tree is made of edges between the metanodes, and edges internal to the metanodes. However, since the non giant components have size and diameter $O(\log n)$ at the moment they connect to the giant, this should nonetheless convince the reader that the *increase* in the diameter during the supercritical phase is an additive term of order $O(\log^\alpha n)$, for some constant α. A similar argument is made precise later (Lemma 4), in a weaker form that still suffices to prove the upper bound we require.

THE CRITICAL PHASE. The previous two paragraphs suggest that the contributions to $MWST(K_n)$ from the sub- and supercritical phases are polylogarithmic. Thus, it is around $p \sim 1/n$ that we must look for the explanation of Theorem 1. For $p = 1/n$, $G_{n,p}$ contains a tree component T whose size is between $n^{2/3}/2$ and $2n^{2/3}$ with positive probability (see [20], Theorem 5.20). This tree is a subtree of $MWST(K_n)$, so $diam(MWST(K_n)) \geq diam(T)$. Conditioned on its size, such a tree is a Cayley tree (uniform labeled tree), and hence has expected diameter $\Theta(n^{1/3})$, as proved by Rényi and Szekeres [31] or Flajolet and Odlyzko [12]. Therefore, $\mathbf{E}\left[diam(MWST(K_n))\right] = \Omega(n^{1/3})$. Proving the matching upper bound is far more technical, but the reader should be convinced by the following heuristic argument: once $p > 1/n$, the pool of components of order $\Theta(n^{2/3})$ is essentially at its peak: these

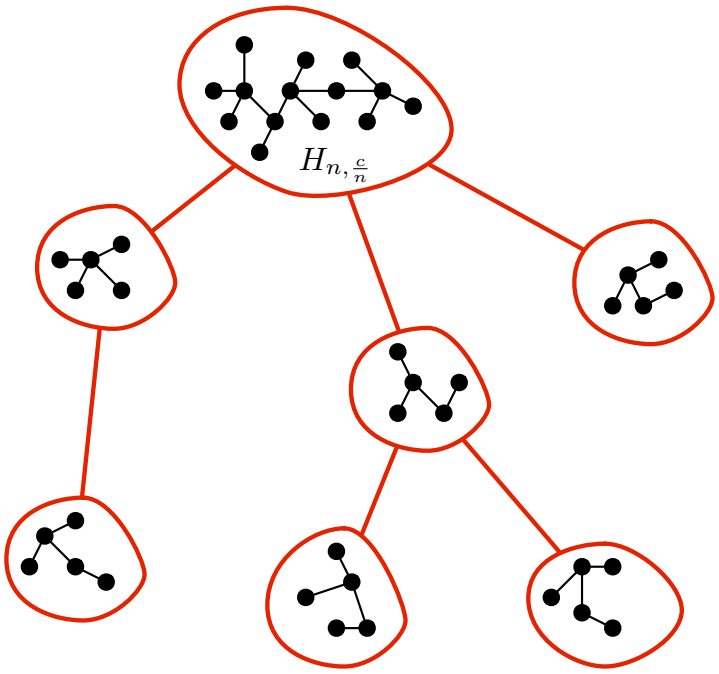

Figure 1: The "random recursive tree" picture of the growth of $T_{n,p}$ in the supercritical phase. The circled trees represent the metanodes of the idealized process.

components hook up together at a very fast pace, and the small ones are much more likely to get glued to larger ones than merging to create new large ones. On the other hand, the number of these component is never larger than ω_n, for any function $\omega_n \to \infty$. In other words, we have a constant number of these components, and even if they were hooking up in the worst possible way, the diameter of the resulting tree would still be $\Theta(n^{1/3})$.

2 Towards the upper bound

In this section, we give a detailed plan of our proof of the upper bound of Theorem 1. In particular, we make formal the intuitive arguments presented above. We also discuss how our proof techniques related to the work of other authors.

2.1 The critical phase

As suggested by the previous arguments, it turns out that when tracking the diameter of $F_{n,p}$, $0 < p < 1$, the action essentially occurs in the "critical window" around $p = 1/n$. The correct parametrization to examine the critical window is $p = 1/n + \lambda n^{-4/3}$. Łuczak [23] showed that for any function $\lambda \to \infty$ a.a.s. for all $p > 1/n + \lambda n^{-4/3}$,

- $|H_{n,p}| = \omega(n^{2/3})$, and all other components have size $o(n^{2/3})$, and
- for all $p' > p$, $H_{n,p} \subseteq H_{n,p'}$.

This fact is crucial to our analysis. Essentially, rather than looking at the forest $F_{n,p}$, we focus on the minimum spanning tree of the giant $MWST(K_n) \cap H_{n,p}$ for $p = 1/n + \Omega(n^{-4/3})$. To track the diameter of this increasing (for inclusion) sequence of graphs, we use the

following fact. For a graph $G = (V, E)$, we write $lp(G)$ for the length of the longest path of G. The subgraph of G induced by a vertex set $U \subset V$ is denoted $G[U]$.

Lemma 2. *Let G, G' be graphs such that $G \subset G'$. Let $H \subset H'$ be connected components of G, G' respectively. Then* $\mathrm{diam}(H') \leq \mathrm{diam}(H) + 2\mathrm{lp}(G'[V - V(H)]) + 2$.

Proof. For any w_1 and w_2 in H', let P_i be a shortest path from w_i to H $(i = 1, 2)$, and let P_3 be a shortest path in H joining the endpoint of P_1 in H to the endpoint of P_2 in H. Then $P_1 \cup P_2 \cup P_3$ is a path of H' from w_1 to w_2 of length at most $diam(H)+2lp(G'[V-V(H)])+2$ (See Figure 2.1). □

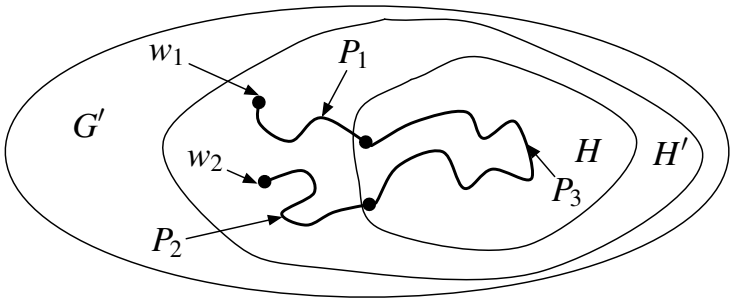

Figure 2: The path $P = P_1 \cup P_2 \cup P_3$ from w_1 to w_2 in H'.

We consider an increasing sequence $1/n < p_0 < p_1 < \ldots < p_t < 1$ of values of p at which we take a snapshot of the random graph process. Specifically, we fix some large constant f_0, and for $i \geq 1$, set $f_i = (5/4)^i f_0$, stopping at the first integer t for which $f_t \geq n^{1/3}/\log n$, and choose $p_i = 1/n + f_i/n^{4/3}$ (the reason for this choice will become clear). This is similar to Łuczak's method of considering "moments" of the graph process [23]. For each p_i, we consider the largest component H_{n,p_i} of G_{n,p_i}. We define D_i to be the diameter of $MWST(K_n) \cap H_{n,p_i}$. We intend to control the increase in diameter between any two successive p_i and p_{i+1} using Lemma 2. For $1 \leq i < t$, we say G_{n,p_i} is *well-behaved* if

(I) $|H_{n,p_i}| \geq (3/2)n^{2/3} f_i$ and the longest path of H_{n,p_i} has length at most $f_i^4 n^{1/3}$, and

(II) the longest path of $G_{n,p_{i+1}}[V - V(H_{n,p_i})]$ has length at most $n^{1/3}/\sqrt{f_i}$.

If G_{n,p_i} is well-behaved then by Lemma 2, $D_{i+1} - D_i \leq 2n^{1/3}/\sqrt{f_i}$. Let $i^\star$ be the smallest integer for which G_{n,p_j} is well-behaved for all $i^\star \leq j < t$ or $i^\star = t$ if $G_{n,p_{t-1}}$ is not well-behaved. By Lemma 2, we have deterministically that

$$D_t - D_{i^\star} \;\leq\; 2\sum_{i=i^\star}^{t-1} n^{1/3}/\sqrt{f_i} \;\leq\; 2f_0 n^{1/3} \sum_{i=1}^{t-1} (4/5)^{i/2} \;=\; O(n^{1/3}). \tag{1}$$

By definition, we have $p_t = 1/n + 1/(n \log n)$, so p_t is not quite supercritical in the sense that $np_t \to 1$, as $n \to \infty$. For such p_t, we cannot prove the polylogarithmic bound we claimed holds if p_t had been c/n for $c > 1$. However, p_t is far enough from $1/n$ that we are able to prove that $\mathbf{E}\,[MWST(K_n) - D_t] = O(n^{1/3})$ (Lemma 4 below). This slight modification in the extents of the phases permits us to even their contributions, and keep

p_t within the range $1/n + o(1/n)$, which happens to be crucial to analyze the events (I) and (II). It follows that

$$\mathbf{E}\left[MWST(K_n)\right] = \mathbf{E}\left[D_{i^\star}\right] + O(n^{1/3}), \tag{2}$$

and the last stage in proving Theorem 1 consists in bounding $\mathbf{E}\left[D_{i^\star}\right]$. The key to doing so is to show that for all j between 0 and $t-1$

$$\mathbf{P}\left\{i^\star = j+1\right\} \;\leq\; 6e^{-\sqrt{f_j/8}}. \tag{3}$$

Using (3) together with (1) and the fact that the longest path has length no longer than n yields that

$$\begin{aligned}
\mathbf{E}\left[D_{i^\star}\right] &\leq f_0^4 n^{1/3} + n\mathbf{P}\left\{i^\star = t\right\} + \sum_{i=1}^{t-1} f_i^4 n^{1/3} \mathbf{P}\left\{i^\star = i\right\} \\
&\leq f_0^4 n^{1/3} + n \cdot 6e^{-(n^{1/3}/8\log n)^{1/2}} + \sum_{i=1}^{t-1} f_i^4 n^{1/3} \cdot 6e^{-(f_{i-1}/8)^{1/2}} \\
&\leq f_0^4 n^{1/3} + O(\frac{1}{n}) + 6n^{1/3} \sum_{i=1}^{t-1} f_i^4 e^{-(f_{i-1}/8)^{1/2}} \\
&= O(n^{1/3}).
\end{aligned}$$

Combining this with (2) completes the upper bound of Theorem 1. To prove (3), we note that if $i^* = j+1$ and $j > 0$, then one of (I) or (II) fails for G_{n,p_j}. We shall show that the probability of any of these events happening is small enough:

Lemma 3. *The following bounds hold for the events (I) and (II) defined above:*
(a) $\mathbf{P}\left\{\text{(I) fails for } G_{n,p_j}\right\} \leq e^{-\sqrt{f_j}}$
(b) $\mathbf{P}\left\{\text{(II) fails for } G_{n,p_j}\right\} \leq 5e^{-\sqrt{f_j/8}}$.

Observe that Lemma 3 implies (3), and hence Theorem 1, since $5e^{-\sqrt{f/8}} + e^{-\sqrt{f}} < 6e^{-\sqrt{f/8}}$ for all $f > 0$.

The key to Lemma 3 (b) is to prove tail bounds on the size of the giant $H_{n,p}$. We then use existing knowledge about the diameter of $G_{n,p}$ for *subcritical* p [25], together with the fact that for $p > 1/n$, $p - 1/n = o(1/n)$, the structure of $G_{n,p}$, minus its giant component, is very similar to the structure of a subcritical random graph [4, 23]. (Our choice of $p_t = 1/n + 1/(n\log n)$ takes advantage of this fact.) To prove Lemma 3 (a), in addition to tail bounds on the size of $H_{n,p}$ we need information about its structure.

Given a connected graph G, the *excess* of G is the quantity $|E(G)| - |V(G)|$; trees, for example, have excess -1. We study $H_{n,p}$ by analyzing a branching process for growing $G_{n,p}$ and a random walk which can be associated to this branching process. Using this approach, we are able to prove upper and lower tail bounds on the size and excess of $H_{n,p}$. Rényi and Szekeres [31], Flajolet and Odlyzko [12], Flajolet et al. [13] have studied the moments of the height of uniformly random labeled trees, and Łuczak [24] has provided information about the precise number of such trees with a given height. The latter result can be used to prove tail bounds on the lengths of longest paths in uniformly random labeled graphs with small excess. We will prove Lemma 3 (a) by combining the bounds on the size and excess of $H_{n,p}$ with these latter bounds.

The size and excess of the components of $G_{n,p}$ for p near $1/n$ have been addressed by several authors [2, 18, 19, 23, 26]. However, these authors were mostly interested in limit probabilities, and did not derive the tail bounds we need. In this sense, our approach can be viewed as providing complementary information on the birth of the giant component. Further results on the size of the $H_{n,p}$ during the critical phase appear in the recent paper of Pittel [28], which proves a central limit theorem for the random variable measuring the size of $H_{n,p}$ and asymptotics for the tails of this random variable, together with an interesting account of known results. Also, Armendáriz (personal communication) has recently generalized the results of Aldous [2], proving that his construction is essentially valid as a process as p varies inside the critical window, and not only for fixed p.

In order to analyze randomized algorithm for sparse instances of Max Cut and Max 2-CSP, Scott and Sorkin [32] elegantly derived *upper* bounds on the size and excess of $G_{n,p}$. They also proceed by analyzing Karp's process, and the resulting bounds are quite similar to ours but are stated for the component containing a given vertex rather than the largest component. Starting from their results, we could easily derive the upper bounds we require. However, deriving corresponding lower bounds using their method would require a substantial reworking of their arguments (Scott, personal communication), and to do so would neither shorten nor conceptually simplify our paper.

2.2 The supercritical phase

We now return to a description of the final stage of the proof, and establish the following lemma.

Lemma 4. *Let t be the first integer for which $f_0(5/4)^t \geq n^{1/3}/\log n$. Then,*

$$\mathbf{E}\left[MWST(K_n)\right] - \mathbf{E}\left[D_t\right] = O(n^{1/6}(\log n)^{7/2}).$$

It is convenient to think of growing the $MWST$ in a different fashion at this point. Consider an arbitrary component C of $G_{n,p_t}[V - V(H_{n,p_t})]$. The edge e with one endpoint in C and the other endpoint in some other component of G_{n,p_t} and which minimizes $w(e)$ subject to this is a cutedge of $G_{n,w(e)}$, and therefore e is necessarily an edge of $MWST(K_n)$.

Let E be the event that $|H_{n,p_t}| > n/\log n$ and every other component of H_{n,p_t} has longest path of length at most $n^{1/6}\sqrt{\log n}$. If E does not occur then one of (I) or (II) fails for G_{n,p_t}, so Lemma 3 tells us that

$$\mathbf{P}\left\{\bar{E}\right\} \leq e^{-(n^{1/3}/\log n)^{1/2}} + 3e^{-(n^{1/3}/8\log n)^{1/2}} = O(1/n). \tag{4}$$

If E holds, then since the edge weights are i.i.d., the second endpoint of e is uniformly distributed among vertices not in C. In particular, with probability at least $|H_{n,p_t}|/n > 1/\log n$, the second endpoint is in H_{n,p_t}, and C gets hooked to the largest component. If the second endpoint is not in H_{n,p_t}, we can think of C joining another component to create C'. The component C' has longest path of length at most $2n^{1/6}\sqrt{\log n}$.

(As an aside, note that $MWST(C')$ is not necessarily a tree created by Kruskal's algorithm, as there may well be edges leaving C' which have weight less than $w(e)$. The technique of growing the $MWST$ of a graph by focussing on the cheapest edge leaving a *specific* component, rather than the cheapest edge joining *any* two components, is known as Prim's tree growing method [21, 30].)

Conditional upon this choice of e, the edge e' leaving C' which minimizes $w(e')$ is also in $MWST(K_n)$. Again, with probability at least $1/\log n$ the second endpoint lies

in $H_{n,t}$. If not, C' joins another component to create C'' with longest path of length at most $3n^{1/6}\sqrt{\log n}$. Continuing in this fashion, we see that the probability the component containing C has longest path of length greater than $rn^{1/6}\sqrt{\log n}$ when it joins to H_{n,p_t} is at most $(1-1/\log n)^r$. In particular, the probability that it has length greater than $n^{1/6}(\log n)^{7/2}$ is at most $(1-1/\log n)^{\log^3 n} = O(1/n^2)$.

Since C was chosen arbitrarily and there are at most n such components, with probability $1-O(1/n)$ none of them has longest path of length greater than $n^{1/6}(\log n)^{7/2}$ before joining H_{n,p_t}. Recall that we are building a tree, and the diameter may only go through two such components. It follows from Lemma 2 that with probability $1-O(1/n)$, $diam(MWST(K_n)) - D_t \leq 2n^{1/6}(\log n)^{7/2} + 2$. Since $diam(MWST(K_n))$ never exceeds n, and recalling (4), we obtain

$$\begin{aligned} \mathbf{E}\left[diam(MWST(K_n)) - D_t\right] &\leq \mathbf{E}\left[\, diam(MWST(K_n)) - D_t \mid E\,\right] + n\mathbf{P}\left\{\bar{E}\right\} \\ &= O(n^{1/6}(\log n)^{7/2}) + O(1), \end{aligned} \tag{5}$$

proving Lemma 4. The remainder of the paper is devoted to proving Lemma 3, which is the only missing ingredient to the proof of Theorem 1. In Section 3 we explain a breadth-first-search-based method for creating $G_{n,p}$. This is the core of the paper, where we derive finer information about the component structure of $G_{n,p}$, improving on the previous known results. In Section 5 we use these results to prove Lemma 3 (b). In Section 6 we derive tail bounds on the diameters of random treelike graphs. Finally, in Section 7 we use these tail bounds to prove Lemma 3 (a).

3 Understanding $G_{n,p}$ through breadth-first search.

3.1 An exploration process

We analyze the component structure of $G_{n,p}$ using a process similar to breadth-first search (BFS) [8] and to a process used by Aldous [2] to study random graphs in the critical window from a weak limit point of view. We highlight that $G_{n,p}$ is a labeled random graph model with vertex set $\{v_1, v_2, \ldots, v_n\}$. For $i \geq 0$, we define the set $\mathcal{O}_i$ of open vertices at time i, and the set A_i of the vertices that have already been explored at time i. We set $\mathcal{O}_0 = v_1$, $A_0 = \emptyset$, and construct $G_{n,p}$ as follows:

Step i $(0 \leq i \leq n-1)$: Let v be an arbitrary vertex of $\mathcal{O}_i$ and let N_i be the random set of neighbours of v in $V \setminus A_i$. Set $\mathcal{O}_{i+1} = \mathcal{O}_i \cup N_i - \{v\}$ and $A_{i+1} = A_i \cup \{v\}$. If $\mathcal{O}_{i+1} = \emptyset$, then reset $\mathcal{O}_{i+1} = \{u\}$, where u is the element of $\{v_1, v_2, \ldots, v_n\} - A_i$ with the smallest index.

Each time $\mathcal{O}_{i+1} = \emptyset$ during some Step i, then a component of $G_{n,p}$ has been created. To get a handle on this process, we now further examine what may happen during Step i. The number of neighbours of v not in $A_i \cup \mathcal{O}_i$ is distributed as a binomial random variable $\mathrm{Bin}(n-i-|\mathcal{O}_i|, p)$. By the properties of $G_{n,p}$, the distribution of edges from v to $V - A_i$ is independent of what happens in the previous steps of the process. Furthermore, if $\mathcal{O}_{i+1} = \emptyset$ does not occur during Step i, then $w \in \mathcal{O}_{i+1} - \mathcal{O}_i$ precisely if $w \notin A_i \cup \mathcal{O}_i$ and we expose an edge from v to w during this step. It follows that $|\mathcal{O}_{i+1}|$ is distributed as $\max(|\mathcal{O}_i| + \mathrm{Bin}(n-i-|\mathcal{O}_i|, p) - 1, 1)$. An advantage of this method of construction is that if $\mathcal{O}_{j+1} = \emptyset$ during Step j, instead of thinking of the process continuing to construct $G_{n,p}$ we may think of *restarting* the process to construct $G_{n-j,p}$.

We can thus analyze the growth of the components of $G_{n,p}$, created by the above BFS-based process, by coupling the process to the following random walk. Let $S_0 = 1$. For $i \geq 0$, let

$$X_{i+1} = \text{Bin}(n - i - S_i, p) - 1 \qquad \text{and} \qquad S_{i+1} = \max(S_i + X_{i+1}, 1).$$

With this definition, for all $i < n - 1$, S_i is precisely $|\mathcal{O}_i|$, and any time $S_{i-1} + X_i = 0$, a component of $G_{n,p}$ has been created. We will sometimes refer to such an event as $\{S_i = 0\}$ or say that "S visits zero at time i".

An analysis of the height of the random walk S and its concentration around its expected value will form a crucial part of almost everything that follows. We will prove matching upper and lower bounds that more-or-less tie down the behavior of the random variable S_i for i in a certain key range, and thereby imply bounds on the sizes of the components of $G_{n,p}$. In analyzing this random walk, we find it convenient to use the following related, but simpler processes:

- S' is the walk with $S'_0 = 1$ and $S'_{i+1} = S'_i + X'_{i+1}$, where $X'_{i+1} = \text{Bin}(n-i-|\mathcal{O}_i|, p)-1$, for $i \geq 0$. This walk behaves like S_i but is allowed to take non-positive values.
- S^{ind} is the walk with $S^{ind}_0 = 1$ and $S^{ind}_{i+1} = S^{ind}_i + \text{Bin}(n - (i+1), p) - 1$, for $i \geq 0$.
- S^h is the walk with $S^h_0 = 1$ and $S^h_{i+1} = S^h_i + \text{Bin}(n - (i+1) - h, p) - 1$, for $i \geq 0$.

Note that *all* of these walks are allowed to go negative. *We couple all the above walks to* S; we emphasize that until the first visit of S to 0, S' agrees with S while S^{ind} strictly dominates it. Finally, S dominates S^h until the first time that S exceeds $h + 1$.

As a preliminary exercise, consider what happens to the random walk S which corresponds to the random graph process $G_{n,1/n}$. It is known that at this point, the largest component has size $\Theta(n^{2/3})$ [10]. In our warmup, we prove a bound of $O(n^{2/3} \log^{1/3} n)$. To do so, we focus on the size of the component containing v_1.

Note that this component has size greater than t precisely if $S_i \neq 0$ for all $i \leq t$. We know that the probability of this event is bounded above by the probability that S^{ind} exceeds 0 for all $i \leq t$. This is at most the probability that $S^{ind}_t > 0$. Now,

$$S^{ind}_t \;=\; 1 + \sum_{i=1}^{t} (\text{Bin}(n - i, 1/n) - 1) \;=\; \text{Bin}\left(nt - \binom{t+1}{2}, \frac{1}{n}\right) - (t - 1).$$

Thus, its mean is $\mathbf{E}S^{ind}_t = 1 - \binom{t+1}{2}/n$. The following theorem states that binomial random variables are very concentrated around their mean:

Theorem 5 (6)**.** *If $Y = \text{Bin}(m, q)$, then denoting $\mathbf{E}[Y]$ (which is mq) by λ, we have:*

$$\mathbf{P}\{Y - \mathbf{E}[Y] > r\} \leq e^{-r^2/2(\lambda + r/3)}, \tag{6}$$

and

$$\mathbf{P}\{Y - \mathbf{E}[Y] < -r\} \leq e^{-r^2/2\lambda}. \tag{7}$$

Let α be a real number greater than 0. Setting $t = \alpha n^{2/3}$ and applying Theorem 5 to $S^{ind}_t + (t - 1)$ yields

$$\mathbf{P}\left\{S^{ind}_t > 0\right\} \;\leq\; 2\exp\left(-\frac{\left(\binom{t+1}{2}/n\right)^2}{2\left(t - \binom{t+1}{2}/n + \binom{t+1}{2}/3n\right)}\right) \;\leq\; 2e^{-\alpha^3/8}.$$

In particular, letting $r = n^{2/3}(16 \log n)^{1/3}$, it follows that $\mathbf{P}\left\{S_r^{ind} > 0\right\} \leq 2/n^2$. Since v_1 is arbitrary, the size of the largest component $H_{n,1/n}$ may be bounded using the union bound:

$$\mathbf{P}\left\{|H_{n,1/n}| > r\right\} \quad \leq \quad \frac{2}{n}.$$

As a consequence, the expected size of the largest component is

$$\mathbf{E}|H_{n,1/n}| \;\leq\; r + n\mathbf{P}\left\{|H_{n,p}| > r\right\} \;=\; O(n^{2/3}\log^{1/3} n).$$

We could prove that the expected size of the largest component at the threshold $p = 1/n$ is $O(n^{2/3})$ by strengthening our argument in either of the following ways:

- by considering the probability not just that S^{ind} exceeds 0 at time t, but the probability that in addition S^{ind} has remained above 0 until time t. It turns out that this latter probability is about $\frac{1}{t}$ times the former in the crucial range, which can be used to obtain the desired sharpening. Or,
- by noting that as our process continues, the probability that we get large components decreases. Specifically, if $S_i = 0$ and $\mathcal{O}_{i+1} = \{u\}$, then the probability that the component containing u has size at least t is at most the probability that $S_{j+t}^{ind} - S_j^{ind} > 0$. This is also the probability that $\mathrm{Bin}(tn - tj - \binom{t+1}{2}, \frac{1}{n})$ exceeds $t-1$.

The preceding argument and remarks lend credence to two claims: first that the largest component of $G_{n,1/n}$ has size $O(n^{2/3})$, and second that any component of this size must arise early in the branching process. The main goal of the rest of this section is to state and prove precise versions of these claims which hold for $G_{n,p}$ when $p - 1/n = \Omega(1/n^{4/3})$ and $p - 1/n = o(1/n \log n)$. To do so, we need to tie down the behavior of S. First, however, we analyze S^{ind},S^h and S', as they buck a little less wildly.

3.2 The height of the tamer walks

We can handle S^{ind} for $p = 1/n + \delta$ using the analysis discussed above, which consists of little more than standard results for the binomial distribution. Specifically, we have that for $t \geq 1$, $S_t^{ind} + (t-1)$ is distributed like $\mathrm{Bin}(nt - \binom{t+1}{2}, p)$, so by linearity of expectation, we have:

Fact 6. *For $p = 1/n + \delta$ with $\delta < 1/n$,*

$$\mathbf{E}S_t^{ind} = \delta nt - \frac{t(t+1)}{2n} - \frac{t(t+1)\delta}{2} + 1 \leq t + 1.$$

Using the fact that the variance of a $\mathrm{Bin}(m,p)$ random variable is $m(p-p^2)$, we have:

Fact 7. *For $p = 1/n + \delta$ with $\delta = o(1/n)$ and $t = o(n)$,*

$$\mathbf{Var}\left\{S_t^{ind}\right\} = \mathbf{Var}\left\{S_t^{ind} + (t-1)\right\} = (1 + o(1))t$$

Intuitively, S_t^{ind} has a good chance of being negative if the variance exceeds the square of the expectation and a tiny chance of being negative if the expectation is positive and dwarfs the square root of the variance. Indeed, we can formalize this intuition using the [6] bounding method.

We are interested in the critical range, $p = 1/n+\delta$ for $\delta = o(1/n)$. For such δ, $t(t+1)\delta/2$ is $o(t(t+1)/2n)$, so we see that $\mathbf{E}S_t^{ind}$ goes negative when $\delta nt \simeq t(t+1)/2n$, i.e., when $t \simeq 2\delta n^2$. Furthermore, for any $\alpha \in (0,1)$, there exist $a_1 = a_1(\alpha) > 0$ and $a_2 = a_2(\alpha) > 0$ such that $\mathbf{E}S_t^{ind}$ is sandwiched between $a_1\delta nt$ and $a_2\delta nt$, for $\alpha\delta n^2 \leq t \leq (2-\alpha)\delta n^2$. As a consequence, $(\mathbf{E}S_t^{ind})^2 = \Theta(\delta^2 n^2 t^2) = \Theta(\delta^3 n^4 t)$ for such p and t.

Also, Fact 7 states that $\mathbf{Var}\left\{S_t^{ind}\right\} = (1+o(1))t$, so the square of the expectation dwarfs the variance in this range provided $\delta^3 n^4$ is much greater than 1, i.e., provided δ is much greater than $1/n^{4/3}$.

Writing $\delta = f/n^{4/3} = f(n)/n^{4/3}$, we will focus on the case where $f > 1$ and $f = o(n^{1/3})$. We assume for the remainder of Section 3, and in particular as a hypothesis in all lemmas and theorems of this section, that $p = 1/n + f/n^{4/3}$ and that f satisfies these constraints. In the lemma that follows we use Chernoff bounds to show that S_t^{ind} is close to its expected value for all such f.

Lemma 8. *For all $1 \leq t \leq n-1$ and $0 \leq x \leq t$,*

$$\mathbf{P}\left\{\left|S_t^{ind} - \mathbf{E}S_t^{ind}\right| > x\right\} \leq 2e^{-x^2/5t}.$$

Furthermore, for any $1 \leq i < j \leq t$,

$$\mathbf{P}\left\{\left|(S_j^{ind} - S_i^{ind}) - \mathbf{E}\left[S_j^{ind} - S_i^{ind}\right]\right| > x\right\} \leq 2e^{-x^2/5t}.$$

Proof. The tail bound on $S_j^{ind} - S_i^{ind}$ follows by applying Theorem 5 to $(S_j^{ind} - S_i^{ind}) + (j-i)$, which is a binomial random variable. Before applying it, we observe that by Fact 6, $\mathbf{E}\left[S_j^{ind} - S_i^{ind} + (j-i)\right] \leq 2j \leq 2t$. Thus

$$\mathbf{P}\left\{\left|(S_j^{ind} - S_i^{ind}) - \mathbf{E}\left[S_j^{ind} - S_i^{ind}\right]\right| > x\right\} \leq 2e^{-x^2/(4t+2x/3)} \leq 2e^{-x^2/5t},$$

which establishes the latter claim. The former is obtained by applying an identical argument to $S_t + (t-1)$, which is also a binomial random variable with mean at most $2t$. □

We turn now to S^h, which is also easier to handle than S.

Lemma 9. *For all $1 \leq t \leq n-1$,*

$$\mathbf{E}S_t^h = \frac{tf}{n^{1/3}} - \frac{t(t+1+2h)}{2n} - \frac{t(t+1+2h)f}{2n^{4/3}} + 1.$$

Furthermore, for all integers $0 \leq i < j \leq t$ and for all $0 \leq x \leq t$,

$$\mathbf{P}\left\{\left|(S_j^h - S_i^h) - \mathbf{E}\left[S_j^h - S_i^h\right]\right| > x\right\} \leq 2e^{-x^2/5t}.$$

We omit the proof of this lemma as it is established just as Fact 6 and Lemma 8. The above lemmas yield tail bounds on the value of some of the random walks associated with S at some *specific* time t. These bounds rather straightforwardly yield bounds on the probability that S is far from its expected value at *any* time up to some fixed time t:

Lemma 10. *Fix $1 \leq t \leq n-1$ and $1 \leq x \leq t$. Then*

$$\mathbf{P}\left\{|S_i^{ind} - \mathbf{E}S_i^{ind}| \geq x \text{ for some } 1 \leq i \leq t\right\} \leq 4e^{-x^2/5t}.$$

Furthermore, an identical bound holds for S^h, for any h for which $t + h \leq n$.

Proof. Let A be the event that there is $i \leq t$ for which $|S_i^{ind} - \mathbf{E}S_i^{ind}| \geq 2x$ - we aim to show bounds on $\mathbf{P}\{A\}$. We consider the first time $i^\star$ at which $|S_{i^\star}^{ind} - \mathbf{E}S_{i^\star}^{ind}| \geq 2x$ (or $i^\star = t+1$ if this never occurs).

For $i \leq t$, let A_i be the event that $i^\star = i$ and let B_i be the event that A_i occurs and $|S_t^{ind} - \mathbf{E}S_t^{ind}| \leq x$. Finally, let B be the event that $|S_t^{ind} - \mathbf{E}S_t^{ind}| > x$. If A occurs then either one of the events B_i occurs or B occurs. As $i^\star$ is a stopping time, for any $i \leq t$,

$$\mathbf{P}\{B_i \mid A_i\} \leq \mathbf{P}\left\{|(S_t^{ind} - S_i) - (\mathbf{E}S_t^{ind} - \mathbf{E}S_i)| \geq x\right\}.$$

Furthermore, the A_i are disjoint so the $\mathbf{P}\{A_i\}$ sum to at most 1. It follows that

$$\begin{aligned}
\mathbf{P}\{A\} &\leq \mathbf{P}\{B\} + \sum_{i=1}^{t} \mathbf{P}\{B_i\} = \mathbf{P}\{B\} + \sum_{i=1}^{t} \mathbf{P}\{B_i \mid A_i\}\,\mathbf{P}\{A_i\} \\
&\leq \mathbf{P}\{B\} + \max_{1\leq i\leq t} \mathbf{P}\left\{\left|(S_t^{ind} - S_i) - \mathbf{E}\left[S_t^{ind} - S_i\right]\right| \geq x\right\} \\
&\leq 2\max_{1\leq i\leq t} \mathbf{P}\left\{\left|(S_t^{ind} - S_i) - \mathbf{E}\left[S_t^{ind} - S_i\right]\right| \geq x\right\} \\
&\leq 4e^{-x^2/5t},
\end{aligned}$$

by applying Lemma 8. An identical bound holds for S^h by mimicking the above argument but applying Lemma 9 at the last step. □

3.3 The height of S

We now turn to the walk we are really interested in. For all i it is deterministically the case that $S_i \leq S_i^{ind} + i$, so we may use Lemma 10 to bound S_i (equivalently, $|\mathcal{O}_i|$) for $1 \leq i \leq t \leq n-1$. Letting $x = \epsilon t$ in Lemma 10 yields:

Corollary 11. *Fix $1 \leq t \leq n-1$ and $0 < \epsilon \leq 1$. Then*

$$\mathbf{P}\{|\mathcal{O}_i| \geq (1+\epsilon)t \text{ for some } 1 \leq i \leq t\} \leq 4e^{-\epsilon^2 t/5}.$$

The above crude bound on $|\mathcal{O}_i|$ is a result of bounds on S^{ind} and the fact that $S_i - S_i^{ind} \leq i$. To get more precise information on the height of S_i, we need to improve our bound on $S_i - S_i^{ind}$. To this end, we note that letting Z_t be the number of times that S_i hits zero up to time t, we have $S_t = S'_t + Z_t$. Since S'_t hits a new minimum each time S_t hits zero, $Z_t = -\min\{S'_i - 1 | 1 \leq i \leq t\}$. Since S^{ind} strictly dominates S', we thus have $S_i \leq S_i^{ind} + Z_i$ for all $1 \leq i \leq n-1$, which will turn out to yield considerably better bounds than $S_i \leq S_i^{ind} + i$ once we have obtained bounds on Z_i. Such bounds follow from the following lemma:

Lemma 12. *For all $1 \leq t \leq n-1$*

$$\mathbf{P}\left\{S'_i \leq \frac{if}{n^{1/3}} - \frac{2t^2}{n} \text{ for some } 1 \leq i \leq t\right\} \leq 8e^{-t^3/100n^2}.$$

Proof of Lemma 12. By Corollary 11, the probability $|\mathcal{O}_i| \geq (5/4)t$ for some $1 \leq i \leq t$ is at most $4e^{-t/80}$. On the other hand, as long as $|\mathcal{O}_i| \leq 5t/4$ for all $1 \leq i \leq t$, $S'_{i+1} - S'_i \geq \mathrm{Bin}(n - i - 5t/4, p) - 1$, so $S'_i \geq S_i^{5t/4}$ for all $1 \leq i \leq t$. Furthermore, it follows from

Lemma 9 and the fact that $f = o(n^{1/3})$ that for any $\epsilon > 0$, for n large enough, for all $1 \le i \le t$,

$$\mathbf{E}S_i^{5t/4} \ge \frac{if}{n^{1/3}} - \left(1 + \frac{f}{n^{1/3}}\right)\frac{i(i+1+5t/2)}{2n} \ge \frac{if}{n^{1/3}} - \frac{(7/4+\epsilon)t^2}{n}.$$

Thus, if $S'_i \le if/n^{1/3} - 2t^2/n$ for some $1 \le i \le t$, then either

(a) $|\mathcal{O}_j| \ge 5t/4$ for some $1 \le j \le t$, or

(b) $\mathbf{E}S_i^{5t/4} - (1/4-\epsilon)t^2/n \ge S'_i \ge S_i^{5t/4}$.

We have already seen that (a) occurs with probability at most $4e^{-t/80} \le 4e^{-t^3/100n^2}$. By choosing $\epsilon = 1/40$, say, and applying Lemma 10 with $x = (1/4-\epsilon)t^2/n = 9t^2/40n$, it follows that for n large enough (b) occurs for some $1 \le i \le t$ with probability at most $4e^{-81t^3/8000n^2} \le 4e^{-t^3/100n^2}$. The lemma follows. □

Corollary 13. *Let $t = 3fn^{2/3}$ – then for n large enough, the probability that $Z_t > 18f^2n^{1/3}$ is at most $8e^{-f^3/4}$.*

Proof. This follows from Lemma 12 applied with $t = 3fn^{2/3}$. □

We are now able to derive much stronger upper tail bounds on S_i:

Theorem 14. *For n large enough, the probability that $S_i > 20f^2n^{1/3}$ for some $1 \le i \le 3fn^{2/3}$ is at most $12e^{-f^3/15}$.*

Proof. Let $t = 3fn^{2/3}$. If $S_i \ge 20f^2n^{1/3}$ for some $1 \le i \le t$ then either $Z_t \ge Z_i \ge 18f^2n^{1/3}$ or $S_i^{ind} \ge 2f^2n^{1/3}$. Corollary 13 yields that the former event has probability at most $8e^{-f^3/4}$. Furthermore, using Fact 6 it is straightforward to see that $\mathbf{E}S_i^{ind} \le f^2n^{1/3}$ for all i, so applying Lemma 10 with $x = f^2n^{1/3}$ yields that the probability S_i^{ind} is more than $2f^2n^{1/3}$ for some $1 \le i \le t$ is at most $4e^{-x^2/5t} = 4e^{-f^3/15}$. Combining these bounds yields the result. □

4 The structure of $G_{n,p}$ inside the critical window

Using these bounds information about S we have gathered in Section 3, we are able to determine the structure of the giant component of $G_{n,p}$ for p in the range we are focussing on. Recall that the excess of a connected graph H is equal to $|E(H)| - |V(H)|$. In this section we prove:

Theorem 15. *There is $F > 1$ such that for $f > F$ and n large enough, with probability at least $1 - e^{-f}$, the random graph $G_{n,p}$ contains a component H of size between $(3/2)fn^{2/3}$ and $(5/2)fn^{2/3}$ and of excess at most $150f^3$.*

The proof of Theorem 15 goes in two natural steps. We first bound the probability that we obtain a component of the desired size (Section 4.1), and then bound the excess of this component (Section 4.2). Observe that Theorem 15 is actually a result about the giant component. Indeed, with high probability there is only one component of such a size, which is therefore, the giant. To prove Theorem 15, rather than considering the largest component, we consider $H_p^\star$, the component $G_{n,p}$ alive at time $fn^{2/3}$. Properties of $H_p^\star$ are easier to derive, and as it will become clear later, $H_p^\star$ happens to be the largest component $H_{n,p}$ with high probability (Theorem 20 below). We start by giving bounds on the size and excess of $H_p^\star$, then we prove that $H_p^\star$ is actually the giant component with high probability.

4.1 The size of the giant component

To begin, we strengthen the argument used in Lemma 12 to tighten the bounds on the size of $H_p^\star$. This is done by using the stronger bound on the height of S given by Theorem 14. Theorem 16 provides a lower bound on the size of $H_p^\star$. This lower bound is completed by an upper bound of Theorem 17.

Theorem 16. *Fix $0 < \alpha \leq 1$. Then for n large enough, the probability that $S_i = 0$ for some $\alpha f n^{2/3} \leq i \leq (2-\alpha) f n^{2/3}$ is at most $13e^{-\alpha^4 f^3/50}$. Hence,*

$$\mathbf{P}\left\{|H_p^\star| \leq (2-2\alpha) f n^{2/3}\right\} \leq 13e^{-\alpha^4 f^3/50}.$$

Proof. As $S_i \geq S_i'$ for all i, it suffices to prove that the probability $S_i' \leq 0$ for some such i is at most $e^{-\alpha^4 f^3/50}$. Letting $h = 20 f^2 n^{1/3}$, we have that S' is at least S^h until the first time i that $S_i \geq h$. From Lemma 9,

$$\mathbf{E}S_i^h \;=\; \mathbf{E}S_i^{ind} - \left(1 + \frac{f}{n^{1/3}}\right)\frac{hi}{n} \;\geq\; \mathbf{E}S_i^{ind} - \frac{40 f^2 i}{n^{2/3}},$$

for n large enough. For $\alpha f n^{2/3} \leq i \leq (2-\alpha) f n^{2/3}$ and n large enough, it follows from Fact 4 that $\mathbf{E}S_i^{ind} \geq (\alpha^2/2) f^2 n^{1/3}$, so

$$\mathbf{E}S_i^h \geq \frac{\alpha^2 f^2 n^{1/3}}{2} - \frac{40 f^2 i}{n^{2/3}}.$$

Furthermore, since $f = o(n^{1/3})$, for n large enough and $i \leq (2-\alpha) f n^{2/3}$ we have $i/n^{2/3} \leq (2-\alpha) f \leq \alpha^2 n^{1/3}/800$, so

$$\mathbf{E}S_i^h \;\geq\; \frac{\alpha^2 f^2 n^{1/3}}{2} - \frac{40\alpha^2 f^2 n^{1/3}}{800} = \frac{9\alpha^2 f^2 n^{1/3}}{20}.$$

Therefore, if $S_i' \leq 0$ for some $\alpha f n^{2/3} \leq i \leq (2-\alpha) f n^{2/3}$, either $S_j \geq h$ for some $j \leq i$ or $S_i^h \leq \mathbf{E}S_i^h - 9\alpha^2 f^2 n^{1/3}/20$. By Theorem 14, the former event has probability at most $12e^{-f^3/15} < 12e^{-\alpha^4 f^3/50}$. Letting $t = (2-\alpha) f n^{2/3}$ and $x = 9\alpha^2 f^2 n^{1/3}/20$ and applying Lemma 10 yields that the latter event has probability at most

$$4e^{-x^2/5t} \leq 4e^{-81\alpha^4 f^4 n^{2/3}/2000(2-\alpha) f n^{2/3}} \leq e^{-\alpha^4 f^3/50}.$$

This completes the proof. □

The above theorem tells us that with high probability, S does not visit zero between times $\alpha f n^{2/3}$ and $(2-\alpha) f n^{2/3}$. Furthermore, $S_i \leq S_i^{ind} + Z_{\alpha f n^{2/3}}$ until the first time after $\alpha f n^{2/3}$ that S visits zero. Combining this fact with our tail bounds on S^{ind} and $Z_{\alpha f n^{2/3}}$, we can show that S very likely *does* visit zero around time $2 f n^{2/3}$. This entails the following upper bound on the size of $H_p^\star$.

Theorem 17. *Fix $0 < \alpha \leq 1$. Then for n large enough, the probability that S does not visit zero between time $(2-\alpha) f n^{2/3}$ and $(2+2\alpha) f n^{2/3}$ is at most $23e^{-\alpha^4 f^3/100}$. As a consequence,*

$$\mathbf{P}\left\{|H_p^\star| \geq (2+2\alpha) f n^{2/3}\right\} \leq 23e^{-\alpha^4 f^3/100}.$$

Theorem 15 is very similar to Corollary 5.6 of Scott and Sorkin [32], though, as mentioned in the introduction, that result is stated for the component containing a specified vertex.

Proof. For simplicity, let $\underline{t} = (2-\alpha)fn^{2/3}$, $\bar{t} = (2+2\alpha)fn^{2/3}$ and let N be the event that S does not visit zero between time $\underline{t}$ and $\bar{t}$. If $S^{ind}_{\bar{t}} < -Z_{\underline{t}}$ then $S'_{\bar{t}} < -Z_{\underline{t}}$, so S has visited 0 between times $\underline{t}$ and $\bar{t}$. Therefore,

$$\mathbf{P}\{N\} \leq \mathbf{P}\left\{S^{ind}_{\bar{t}} \geq -Z_{\underline{t}}\right\}.$$

We bound the right hand side of this equation by writing

$$\mathbf{P}\left\{S^{ind}_{\bar{t}} \geq -Z_{\underline{t}}\right\} \leq \mathbf{P}\left\{S^{ind}_{\bar{t}} \geq -r\right\} + \mathbf{P}\left\{Z_{\bar{t}} > r\right\}, \tag{8}$$

and deriving bounds on the two terms of the right-hand-side of (8) for suitably chosen r. For any r,

$$\mathbf{P}\left\{Z_{\underline{t}} > r\right\} \quad \leq \quad \mathbf{P}\left\{Z_{\alpha fn^{2/3}} > r\right\} + \mathbf{P}\left\{Z_{\underline{t}} > Z_{\alpha fn^{2/3}}\right\}. \tag{9}$$

Since $Z_{\underline{t}} > Z_{\alpha fn^{2/3}}$ occurs precisely if S visits zero between times $\alpha fn^{2/3}$ and $\underline{t}$, Theorem 16 yields that

$$\mathbf{P}\left\{Z_{\underline{t}} > Z_{\alpha fn^{2/3}}\right\} \leq 13e^{-\alpha^4 f^3/50}. \tag{10}$$

By its definition, $Z_{\alpha fn^{2/3}} > 2\alpha^2 f^2 n^{1/3}$ precisely if $S_i \leq -2\alpha^2 f^2 n^{1/3}$ for some $i \leq \alpha fn^{2/3}$. Applying Lemma 12 with $t = \alpha fn^{2/3}$ thus yields that

$$\begin{aligned}
\mathbf{P}\left\{Z_{\alpha fn^{2/3}} > 2\alpha^2 f^2 n^{1/3}\right\} &\leq \mathbf{P}\left\{S_i \leq \frac{if}{n^{1/3}} - \frac{2t^2}{n} \text{ for some } 1 \leq i \leq t\right\} \\
&\leq 8e^{-t^3/100n^2} \\
&\leq 8e^{-\alpha^3 f^3/100}.
\end{aligned} \tag{11}$$

Letting $r = 2\alpha^2 f^2 n^{1/3}$, (9), (10), and (11) yield

$$\begin{aligned}
\mathbf{P}\left\{Z_{\underline{t}} > 2\alpha^2 f^2 n^{1/3}\right\} &\leq \mathbf{P}\left\{Z_{\underline{t}} > Z_{\alpha fn^{2/3}}\right\} + \mathbf{P}\left\{Z_{\alpha fn^{2/3}} > 2\alpha^2 f^2 n^{1/3}\right\} \\
&\leq 13e^{-\alpha^2 f^3/50} + 8e^{-\alpha^3 f^3/100} < 21e^{-\alpha^4 f^3/100}.
\end{aligned} \tag{12}$$

Furthermore,

$$\mathbf{E}S^{ind}_{\bar{t}} \leq \frac{\bar{t}f}{n^{1/3}} - \frac{\bar{t}^2}{2n} \leq -(2\alpha + 2\alpha^2)f^2 n^{1/3},$$

so $-2\alpha^2 f^2 n^{1/3} \geq \mathbf{E}S^{ind}_{\bar{t}} + 2\alpha f^2 n^{1/3}$. By applying Lemma 8 with $x = 2\alpha f^2 n^{1/3}$, it follows that

$$\mathbf{P}\left\{S^{ind}_{\bar{t}} \geq -2\alpha^2 f^2 n^{1/3}\right\} < 2e^{-x^2/5\bar{t}} = 2e^{-4\alpha^2 f^4 n^{2/3}/5\bar{t}} \leq 2e^{-\alpha^2 f^3/5}. \tag{13}$$

Combining (8), (12) and (13) yields that

$$\mathbf{P}\left\{S^{ind}_{\bar{t}} \geq -Z_{\underline{t}}\right\} \leq 23e^{-\alpha^4 f^3/100}.$$

This completes the proof. □

4.2 The excess of the giant

Theorems 16 and 17 tell us about the size of the giant component of $G_{n,p}$. We now turn to its excess. Rather than studying the actual largest component, We prove:

Lemma 18. *Let Exc be the event that $H_p^\star$ has excess at most $150f^3$. Then for n large enough,*

$$\mathbf{P}\left\{\overline{Exc}\right\} \leq 32e^{-f^3/(2^4\cdot 100)}. \tag{14}$$

Proof. For simplicity in coming calculations, we define the *net excess* of a connected graph H to be equal to the excess of H, plus 1. The net excess of components of $G_{n,p}$ can be analyzed much as we have just analyzed their size. In the process defined at the beginning of Section 3, each element of the random set N_i of neighbours of v_i that is in the set $\mathcal{O}_i$ contributes exactly 1 to the net excess of the component alive at time i. Thus, if a component is created between times t_1 and t_2 of the process (precisely, if $S_{t_1} - 1 = 0$ and the first time greater than $t_1 - 1$ at which S visits 0 is t_2), then the net excess of this component is precisely $\sum_{i=t_1}^{t_2-1} \mathrm{Bin}(|\mathcal{O}_i| - 1, p) = \mathrm{Bin}(\sum_{i=t_1}^{t_2-1} S_i - 1, p)$. Our upper bound on S in Theorem 14 can be thus used to prove upper bounds on the net excess of $H_p^\star$.

Let $\alpha = 1/2$ and let *Big* be the event that $H_p^\star$ has size more than $3fn^{2/3}$, let *High* be the event that $S_i \geq 20f^2n^{1/3}$ for some $i \leq 3fn^{2/3}$. If *Big* occurs then S does not return to zero between time $(2-\alpha)fn^{2/3}$ and time $(2+2\alpha)fn^{2/3}$, so by Theorem 17,

$$\mathbf{P}\left\{Big\right\} \;\leq\; 23e^{-\alpha^4 f^3/100} \;=\; 23e^{-f^3/(2^4\cdot 100)}.$$

By Theorem 14, $\mathbf{P}\left\{High\right\} \leq 8e^{-f^3/15}$. If neither *Big* nor *High* occurs, then the net excess of $H_p^\star$ is at most $\mathrm{Bin}(M,p)$, where $M = \sum_{i=1}^{3fn^{2/3}}(S_i - 1) \leq 60f^3n$. For any $m \leq 60f^3n$, $\mathbf{E}\left[\mathrm{Bin}(m,p)\right] \leq 120f^3$, so by Theorem 5, $\mathbf{P}\left\{\mathrm{Bin}(m,p) \geq 150f^3\right\} \leq e^{-f^3}$. Combining these bounds yields

$$\mathbf{P}\left\{\overline{Exc}\right\} \;\leq\; \mathbf{P}\left\{Big\right\} + \mathbf{P}\left\{High\right\} + e^{-f^3} \;\leq\; 32e^{-f^3/(2^4\cdot 100)}. \qquad \square$$

4.3 The Giant Towers Over the Others

As discussed in the introduction, the probability of growing a large component which starts in iteration t of the process decreases as t increases. This is what allows us to show that very likely there is a unique giant component and all the other components are much smaller. In the following, we call a connected component *complex* if it contains at least two cycles.

To be precise, let T_1 be the first time that S visits zero after time $(2-\alpha)fn^{2/3}$. Then the remainder of $G_{n,p}$ has $n' = n - T_1$ vertices and each pair of vertices is joined independently with probability p. If $\alpha = 1/4$, say, then

$$\begin{aligned} p &= \frac{1}{n} + \frac{(2-\alpha)f}{n^{4/3}} \leq \frac{1}{n'}\left(1 - \frac{(2-\alpha)f}{n^{1/3}}\right) + \frac{f}{n^{4/3}} \\ &< \frac{1}{n'} - \frac{(f/2)}{(n')^{4/3}}. \end{aligned} \tag{15}$$

Thus the final stages of the process look like a subcritical process on n' vertices. We could analyze how this procedure behaves by looking at the behaviour of our random walks as in the last three subsections but instead we find it convenient to quote results of Łuczak who did obtain tail bounds for the subcritical process.

The following theorem is a reformulation of Łuczak [23], Lemma 1 and Łuczak [25], Theorem 11. Those results are stated for the case $f = f(n) \to \infty$, but in both cases the proof is easily adapted to our formulation; the details are omitted.

Theorem 19. *For all fixed $K > 1$, there exists $F > 1$ such that for all $f > F$, n large enough and $p = 1/n - f/n^{4/3}$, for all $k > K$ the probability that $G_{n,p}$ contains a component of size larger than $(k + \log(f^3))n^{2/3}/f^2$ or a complex component of size larger than $2k$ is at most $3e^{-k}$. Furthermore, the probability there is a tree or unicyclic component of $G_{n,p}$ with size at most $n^{2/3}/f$ and longest path at least $12n^{1/3}\log f/\sqrt{f}$ is at most $e^{-\sqrt{f}}$.*

Using this result we easily obtain the following two theorems:

Theorem 20. *There is $F > 1$ such that for $f > F$ and n large enough, with probability at least $1 - e^{-f}$, the component alive at time $fn^{2/3}$ is the largest component, i.e., $H_p^\star = H_{n,p}$.*

Theorem 21. *For any $\epsilon > 0$ there is $F = F(\epsilon) > 1$ so that for all $f > F$ and $p = 1/n + f/n^{4/3}$, the expected number of components of $G_{n,p}$ of size exceeding $(3/2)fn^{2/3}$ is at most $1 + \epsilon$.*

Theorems 20 and 21 are simple consequences of Theorem 19 and of the above branching process. Letting T_1 be the first time after $(7/4)fn^{2/3}$ that S visits 0, there is at most one component of size at least $(3/2)fn^{2/3}$ that is grown up to time T_1, and with high probability there is exactly one such component. Note that any such component must be alive at time $fn^{2/3}$. We restart the branching process to grow the graph $G_{n-T_1,p}$. Theorem 19 guarantees that the probability a component of size exceeding $n^{2/3}$ ever occurs is at most $e^{-f^2/2}$, which is at most e^{-f} for f large enough. This proves Theorem 20.

If a component of size exceeding $n^{2/3}$ does occur after time T_1, then once it dies we again restart the branching process to grow the remainder of the graph; again, and independently, the probability a component of size exceeding $n^{2/3}$ ever occurs is at most $e^{-f^2/2}$. Continuing in this manner yields the geometric upper bound $(e^{-f^2/2})^i$ on the probability there are precisely i large components grown after time T_1; by making f large we may thus make the expected number of such components arbitrarily close to zero, which proves Theorem 21

We are now ready to prove Theorem 15.

Proof of Theorem 15. Theorems 16 and 17, applied with $\alpha = 1/4$, yield that

$$\mathbf{P}\left\{\frac{3}{2}fn^{2/3} \le |H_p^\star| \le \frac{5}{2}fn^{2/3}\right\} \ge 1 - 36e^{-f^3/(4^4\cdot 100)}.$$

On the other hand, Lemma 18 shows that excess of $H_p^\star$ is at most $150f^3$ with probability at least $1 - 32e^{-f^3/(2^4\cdot 100)}$. Thus, the probability both hold is at least $1 - 68e^{-f^3/(4^4\cdot 100)} \ge 1 - e^{-f}$ for f large enough. (We note that this establishes something slightly stronger than Theorem 15; namely, we have shown that *the component* $H_p^\star$ has such size and excess with the desired probability.) □

5 The proof of Lemma 3 (b)

Let $\mathcal{H}$ be the set of all labeled connected graphs H with vertex set $V(H) \subset \{v_1, \ldots, v_n\}$ for which H has between $(3/2)fn^{2/3}$ and $(5/2)fn^{2/3}$ vertices. For $H \in \mathcal{H}$, let C_H be the event that in the random graph process, H is a connected component of $G_{n,p}$, and let *Bad*

be the event that in the random graph process, *no* element of $\mathcal{H}$ is a connected component of $G_{n,p}$. For *any* event E we may write

$$\mathbf{P}\{E\} \;\leq\; \mathbf{P}\{Bad\} + \sum_{H\in\mathcal{H}} \mathbf{P}\{E \mid C_H\}\,\mathbf{P}\{C_H\}.$$

If Bad occurs then $G_{n,p}$ has *no* component of size between $(3/2)fn^{2/3}$ and $(5/2)fn^{2/3}$, so by Theorem 15, $\mathbf{P}\{Bad\} \leq e^{-f}$ for f large enough. Therefore,

$$\begin{aligned}\mathbf{P}\{E\} &\leq e^{-f} + \sum_{H\in\mathcal{H}} \mathbf{P}\{E \mid C_H\}\,\mathbf{P}\{C_H\}.\\ &\leq e^{-f} + \mathbf{E}\left[|\{H \;:\; C_H \text{ holds }\}|\right]\cdot \max_{H\in\mathcal{H}} \mathbf{P}\{E \mid C_H\}.\end{aligned}$$

Applying Theorem 21 with $\epsilon = 1$ to bound the above expectation, we have that for f and n large enough,

$$\mathbf{P}\{E\} \leq e^{-f} + 2\max_{H\in\mathcal{H}} \mathbf{P}\{E \mid C_H\}. \tag{16}$$

Let $p = 1/n + f/n^{4/3}$ and let $p' = 1/n + (5/4)f/n^{4/3}$. Recall also that $H_{n,p}$ is the largest component of $G_{n,p}$. We will apply equation (16) to the event *Long* that some component of $G_{n,p'}[V - V(H_{n,p})]$ has longest path of length at least $n^{1/3}/f^{1/4}$.

For any graph $H \in \mathcal{H}$, the graph $G_{n,p'}[V-V(H)]$ is $G_{n',p'}$ for some $n' \leq n-(3/2)fn^{2/3}$, and so

$$\begin{aligned}p' &= \frac{1}{n} + \frac{(5/4)f}{n^{4/3}} \;\leq\; \frac{1}{n'}\left(1 - \frac{(3/2)f}{n^{1/3}}\right) + \frac{(5/4)f}{n^{4/3}}\\ &< \frac{1}{n'} - \frac{(1/4)f}{n^{4/3}} \;\leq\; \frac{1}{n'} - \frac{(1/8)f}{(n')^{4/3}},\end{aligned} \tag{17}$$

for n large enough. Let $Large(H)$ be the event that either

(a) $G_{n,p'}[V-V(H)]$ has a component of size larger than $8n^{2/3}/f$, or a complex component of size greater than f, or

(b) $G_{n,p'}[V - V(H)]$ has a tree or unicyclic component of size at most $8n^{2/3}/f$ and longest path of length at least $36n^{1/3}\log f/\sqrt{f}$.

$G_{n,p'}[V - V(H)]$ is a subcritical random graph by (17). For f large enough, Theorem 19 applied with $k = f/10$, say, therefore yields that (a) occurs with probability at most $3e^{-f/10}$. Theorem 19 also yields that (b) occurs with probability at most $e^{-\sqrt{f/8}}$. As $3e^{-f/10} \leq e^{-\sqrt{f/8}}$ for f large enough, this yields, for f large enough,

$$\mathbf{P}\{Large(H)\} \;\leq\; 2e^{-\sqrt{f/8}}.$$

If C_H occurs but (a) does not then $G_{n,p}$ certainly has no component of size larger than H so $H = H_{n,p}$. Also, note that for f large enough $36n^{1/3}\log f/\sqrt{f} < n^{1/3}/f^{1/4}$. For such f, assuming C_H occurs and $Large(H)$ does not occur then *Long* does not occur. To see this, observe that (a) does not occur, so no component other than H has size at least $8n^{2/3}/f$. Using this fact, and since (b) does not occur either and no tree or unicyclic component has a large enough longest path. Finally, the complex components have size at most f, hence the length of their longest path is also bounded by f. So overall, no component of $G_{n,p'}[V-V(H)]$ contains path with length at least $n^{1/3}/f^{1/4}$ and *Long* does

not occur. Furthermore, $Large(H)$ is independent of C_H as the two events are determined by disjoint sets of edges. Therefore,

$$\mathbf{P}\{Long \mid C_H\} \;\leq\; \mathbf{P}\{Large(H) \mid C_H\} \;=\; \mathbf{P}\{Large(H)\} \;\leq\; 2e^{-\sqrt{f/8}},$$

which combined with (16) applied with $E = Long$ yields

Lemma 22. *There exists $F > 1$ such that for $f > F$, for n large enough, $\mathbf{P}\{Long\} \leq 5e^{-\sqrt{f/8}}$.*

This proves the bound of Lemma 3 (b).

6 Longest paths in random treelike graphs

As mentioned before, information about the excess of a random connected graph gives us information about its diameter. This is, in essence, because a random graph with only a few more edges than vertices is "treelike"; in this section we make this idea precise.

6.1 The diameter of uniform trees

We first collect the required bounds on the diameter of trees. A uniform random rooted tree of size s is a tree chosen uniformly at random from among all rooted labeled trees with s nodes. Rényi and Szekeres [31] and Flajolet and Odlyzko [12] have calculated the asymptotics of the moments of the height H_s of a uniform random rooted tree R_s of size s and provided sharp information about the number of uniformly random rooted trees of size s and height $c\sqrt{s}$ for constant c. (This notation H_s is only used in this section and should not affect the readability of the remainder of the document where $H_{n,p}$ denotes the largest component of $G_{n,p}$.) Through combinatorial arguments, Łuczak [24] has extended these results to count the number of such trees when $c = c(s)$ is $\omega(1)$. The version of Łuczak's result that we need can be stated as:

Theorem 23 ([24], p. 299)**.** *There is $C > 0$ such that for s large enough, for all $t \geq C\sqrt{s}$,*

$$\mathbf{P}\{H_s = t\} \leq \frac{e^{-t^2/4s}}{\sqrt{s}}.$$

In fact, this theorem is weaker than what Łuczak proved, but it is easier to state and suffices for our purposes. We have as an immediate consequence:

Corollary 24. *There is $C > 0$, such that for s large enough, for all $c \geq C$,*

$$\mathbf{P}\left\{H_s \geq c\sqrt{s}\right\} \leq 2e^{-c^2/4}.$$

Proof. By Theorem 23, we have that for $c \geq C$,

$$\begin{aligned}
\mathbf{P}\left\{H_s \geq c\sqrt{s}\right\} &\leq \sum_{t=\lceil c\sqrt{s}\rceil}^{s} \frac{1}{\sqrt{s}} e^{-t^2/4s} \\
&\leq \sum_{i=0}^{\lceil\sqrt{s}-c\rceil} \sum_{t=\lceil (c+i)\sqrt{s}\rceil}^{\lceil (c+i+1)\sqrt{s}\rceil - 1} \frac{1}{\sqrt{s}} e^{-t^2/4s} \\
&\leq \frac{\sqrt{s}+1}{\sqrt{s}} \sum_{i=0}^{\lceil\sqrt{s}-c-1\rceil} e^{-(c+i)^2/4} \\
&\leq 2e^{-c^2/4},
\end{aligned}$$

as long as c and s are large enough. $\square$

There is a natural s-to-1 map from rooted trees of size s to unrooted trees of size s, obtained by "unrooting". Clearly, if T_s is an unrooted tree corresponding to R_s via this map, then $lp(T_s) = lp(R_s) \leq 2H_s$. As a consequence,

Lemma 25. *Let T_s be a uniformly random unrooted tree (a* Cayley *tree) on s nodes. Then there is $C > 0$ such that for s large enough, for all $c > C$*

$$\mathbf{P}\left\{\mathrm{lp}(T_s) \geq 2c\sqrt{s}\right\} \;\leq\; \mathbf{P}\left\{H_s > c\sqrt{s}\right\} \;\leq\; 2e^{-c^2/4}. \tag{18}$$

Lemma 25 is the key fact about random trees that allows us to bound the lengths of the longest paths of uniformly random connected tree-like graphs. We now focus our attention on bounding longest paths in such graphs. In doing so, it is useful to describe them in a way that emphasizes some underlying tree structures.

6.2 Describing graphs with small excess

Given a connected labeled graph G with excess q, define the *core* $C = C(G)$ to be the maximum induced subgraph of G which has minimum degree 2. To see that the core is indeed unique, we note that it is precisely the graph obtained by repeatedly removing vertices of degree 1 from G until no such vertices exist (so in particular, if G is a tree then C is empty). It is clear from the latter fact that $G[V - V(C)]$ is a forest, so if $v_i \in V - V(C)$, then there is a unique shortest path in G from v_i to some $v_j \in V(C)$. We thus assign to each vertex $v_j \in V(C)$ the set of labels

$$L_{v_j} = \{j\} \cup \{i : \text{the shortest path from } v_i \text{ to } C \text{ ends at } v_j\}.$$

We next define the *kernel* $K = K(G)$ to be the multigraph obtained from $C(G)$ by replacing all paths whose internal vertices all have degree 2 in C and whose endpoints have degree at least three in C by a single edge [see, e.g., 20]. If $q < 1$ we agree that the kernel is empty; otherwise the kernel has minimum degree 3 and precisely q more edges than vertices. It follows that the kernel always has at most $2q$ vertices and at most $3q$ edges. We denote the multiplicity of edge e in K by $m(e)$. We think of K as a simple graph in which edge e has positive integer weight $m(e)$, to emphasize the fact that parallel edges are indistinguishable. We may keep track of what vertices correspond to *edges* of $K(G)$ as we did for *vertices* of $C(G)$: if $P_1, \ldots, P_{m(e)}$ are paths of $C(G)$ corresponding to edge $e = xy$ of $K(G)$, we let $L_e^i = \bigcup_{v \in V(P_i) - x - y} L_v$ (if $P_i = xy$ then $L_e^i = \emptyset$) and assign a *set of sets of labels* $\{L_e^1, \ldots, L_e^{m(e)}\}$ to e. We emphasize that permuting the order of $P_1, \ldots, P_{m(e)}$ does not change the label of e.

Given a labeled graph G, the above reduction yields a labeled multigraph K and sets L_v for each vertex of K, $\{L_e^1, \ldots, L_e^{m(e)}\}$ for each edge of K. Conversely, any graph with nonempty labeled kernel K to which such sets have been assigned can be described uniquely in the following way:

- For all $v_i \in V(K)$, let T_{v_i} be a labeled tree with labels from L_{v_i}.
- For all $e = xy \in V(K)$, and all $i = 1, 2, \ldots, m(e)$, let T_e^i be a labeled tree with labels from L_e^i (if $L_e^i = \emptyset$ then $T_e = \emptyset$ - this can occur for at most one $i \in \{1, 2, \ldots, m(e)\}$). If $L_e^i \neq \emptyset$, our description depends on whether e is a loop, i.e., on whether $y = x$:

– If $x \neq y$ then mark an element of L_e^i with an $\mathbf{X}$ and mark an element of L_e^i with a $\mathbf{Y}$. We allow that the same element of L_e^i receives both markers.

– If $x = y$ then place two markers of type $\mathbf{X}$ on elements of L_e^i. Again, we allow that the same element of L_e^i receives both markers.

Observe that in marking elements of L_e^i, if $e = xy$ and $x \neq y$ then there are $|L_e^i|^2$ ways to place the markers. If $e = xx$ then there are $|L_e^i| + \binom{|L_e^i|}{2} = (|L_e^i| + 1)|L_e^i|/2$ ways to place the markers as we may either choose an element of L_e^i and place both $\mathbf{X}$ markers on it, or we may choose two distinct elements of L_e^i and place an $\mathbf{X}$ marker on each.

We obtain G from this description as follows:

1. for all $v_j \in V(K)$, identify the vertices $v_j \in V(K)$ and $v_j \in T_{v_j}$, then
2. for all loops $e = xx \in E(K)$, choose a copy of e for each nonempty tree T_e^i, $1 \leq i \leq m(e)$. Remove this edge and let x be adjacent to the vertices in T_e^i marked with $\mathbf{X}$.
3. for all $e = xy \in E(K)$ with $x \neq y$, choose an edge xy for each nonempty tree T_e^i, $1 \leq i \leq m(e)$. Remove this edge, then let x (respectively y) be adjacent to the vertex in T_e^i marked with $\mathbf{X}$ (respectively $\mathbf{Y}$).

Clearly labeled graphs with distinct labeled kernels are not identical. Now, let G, G' be graphs with the same labeled kernel K. If for some $v \in V(K)$, $L_v \neq L'_v$ or for some $e \in E(K)$, $\{L_e^1, \ldots, L_e^{m(e)}\} \neq \{L_e'^1, \ldots, L_e'^{m(e)}\}$, then G, G' are not identical. Hence, given a labeled kernel K, and sets of labels $\{L_v | v \in V(K)\}$, $\bigcup_{e \in E(K)} \{L_e^1, \ldots, L_e^{m(e)}\}$, and distinguished elements of the nonempty sets L_e^i as described above, there are

$$\prod_{v \in V(K)} |L_v|^{|L_v|-2} \prod_{e \in E(K)} \prod_{i : L_e^i \neq \emptyset} |L_e^i|^{|L_e^i|-2}$$

possible graphs, corresponding to the choices of a tree for each set L_v and for each set L_e^i. It follows that if G is a uniformly random connected labeled graph with p vertices and excess $q \geq 1$ specified by its kernel K and a description as above, then conditional on the sizes of their elements, the sets $\mathcal{T}_V = \{T_v | v \in V(K)\}$ and $\mathcal{T}_E = \bigcup_{e \in E(K)} \{T_e^1, \ldots, T_e^{m(e)}\}$ must be uniformly random amongst all such sets. As a consequence, conditional on their sizes, the unrooted labeled trees in $\mathcal{T}_V$ and in $\mathcal{T}_E$ must be uniformly random; i.e., they are simply Cayley trees.

Labeled unicyclic graphs (graphs with excess 1) have empty kernels but nonempty cores; they can be described in a similar but simpler way. Suppose we are given a labeled graph G with unique cycle C. We let T_1 be the unique maximal tree containing vertex v_1 and containing exactly one element v^* of C – set $\mathcal{T}_V = \{T_1\}$ and mark v^*. The vertex v^* has exactly two distinct neighbours w^*, x^* in the tree T_2 induced by the vertices in $V(G) - V(T_1)$; we let $\mathcal{T}_E = \{T_2\}$ and mark w^*, x^*. Given trees T_1, T_2 such that $v_1 \in V(T_1)$, T_1 contains one marked vertex v^* and T_2 contains two marked vertices w^*, x^*, we may construct a unicyclic graph G by letting w^* and x^* be adjacent to v^*. The only difference between this bijection and that given for graphs with nonempty kernel is that now we need to mark a vertex in the tree in $\mathcal{T}_V$. As above, this bijection shows that conditional on their sizes, the trees T_1 and T_2 are Cayley trees.

6.3 The diameter of graphs with small excess

With this latter fact in hand, it is easy to prove bounds on $lp(G)$. Recall that the *net excess* of a connected graph G is equal to the excess of G, plus 1.

Lemma 26. *Let G be a uniformly random labeled connected graph on s vertices and with net excess q. Then there is C such that for s large enough, for all $c \geq C$,*

$$\mathbf{P}\left\{\mathrm{lp}(G) \geq 2(5q+1)c\sqrt{s} + 10q\right\} \leq 10qe^{-c^2/4}. \tag{19}$$

Proof. The bound holds by (18) if $q = 0$. If $q > 0$ then let the sets $\mathcal{T}_V$ and $\mathcal{T}_E$ be defined as above, and let $\mathcal{T} = \mathcal{T}_V \cup \mathcal{T}_E$ - then $|\mathcal{T}| \leq 5q$ as if $q \geq 2$, the kernel has at most $2q$ vertices and at most $3q$ edges, counting multiplicity, and if $q = 1$ then $|\mathcal{T}_V| = |\mathcal{T}_E| = 1$. Trivially, any path P in G is composed of paths from the trees in $\mathcal{T}$ together with edges of G that are not edges of some tree in $\mathcal{T}$. For a given tree T, if P does not have an endpoint in T then it must enter and exit T at most once, i.e., the intersection of P with T, if nonempty, is itself a path. P may also enter one or two of the trees without leaving them – such trees must contain an endpoint of P. If the endpoints are in distinct trees then the intersection of these trees with P are both paths; if the endpoints are in the same tree then that tree's intersection with P consists of two paths.

(In fact, P can *not* enter every tree. If $q > 1$, for example, then the set of vertices and edges of the kernel that have trees intersecting P can not itself contain a cycle in the kernel. We crudely bound the length of P by supposing that it may contain a path from every tree and two paths from at most one tree, so at most $(5q+1)$ paths from trees of $\mathcal{T}$ in total.)

Each time the path P enters or exits a tree, it uses an edge of G that is not an edge of a tree in $\mathcal{T}$. By the definition of the trees in $\mathcal{T}$, there are precisely two such edges for each nonempty tree T_e^i; thus there are at most $10q$ such edges in total. We thus have

$$\mathbf{P}\left\{lp(G) \geq 2(5q+1)c\sqrt{s} + 10q\right\} \quad \leq \quad \mathbf{P}\left\{\max_{T\in\mathcal{T}} lp(T) \geq 2c\sqrt{s}\right\}. \tag{20}$$

We choose C large enough so that if $|T| \geq C\sqrt{s}$ and $c \geq C$ then Lemma 25 applies to T with this choice of c. For $c \geq C$, for all $T \in \mathcal{T}$ either $|T| < C\sqrt{s}$, in which case $\mathbf{P}\{lp(T) \geq 2c\sqrt{s}\} = 0$, or $|T| \geq C\sqrt{s}$, in which case since $|T| \leq s$, there is $c' \geq c$ such that $2c\sqrt{s} = 2c'\sqrt{|T|}$. In the latter case, $\mathbf{P}\{lp(T) \geq 2c\sqrt{s}\} = \mathbf{P}\left\{lp(T) \geq 2c'\sqrt{|T|}\right\} \leq 2e^{-c^2/4}$ by Lemma 25. Therefore, by a union bound applied to the right-hand-side of (20) we have

$$\mathbf{P}\left\{lp(G) \geq 2(5q+1)c\sqrt{s} + 10q\right\} \leq 5q(2e^{-c^2/4}) = 10qe^{-c^2/4}. \qquad \square$$

7 Proof of Lemma 3 (a)

We apply Lemma 26 to bound $lp(H_{n,p})$. First, let D be the event that $H_{n,p}$ has size greater than $(5/2)fn^{2/3}$, which we denote by s, or excess greater than $150f^3$, which we denote by q. If D occurs then either

(a) $H_p^\star$ has size greater than s or excess greater than q or

(b) $H_p^\star \neq H_{n,p}$.

The event (a) occurs with probability at most e^{-f}. By Theorem 20, the probability that (b) occurs is also at most e^{-f}, so $\mathbf{P}\{D\} \leq 2e^{-f}$. Letting E_p be the event that $lp(H_{n,p}) > f^4 n^{1/3}$ we have

$$\begin{aligned} \mathbf{P}\{E_p\} &\leq \mathbf{P}\left\{E_p \mid \overline{D}\right\} + \mathbf{P}\{D\} \\ &\leq \mathbf{P}\left\{E_p \mid \overline{D}\right\} + 2e^{-f}. \end{aligned}$$

Furthermore, $2(5q+1)c\sqrt{s}+10q < 5000cf^{7/2}n^{1/3}$ for f large enough. It follows by applying Lemma 26 with $c = f^{1/2}/5000$ (which is at least C for f large enough) that

$$\mathbf{P}\left\{E_p \mid \overline{D}\right\} \leq 1500 f_r^3 e^{-c^2/4} \leq 1500 f_r^3 e^{-f/(10000^2)} \leq e^{-f/2^{30}},$$

for f large enough. By this bound and (21), we have $\mathbf{P}\{E_p\} \leq 2e^{-f} + e^{-f/2^{30}}$, which is at most $e^{-\sqrt{f}}$ for f large enough. This proves Lemma 3 (a).

8 Conclusion

We have pinned down the growth rate of the diameter of the minimum spanning tree of K_n whose edges are weighted with i.i.d. $[0,1]$-uniform random variables. We did so using probabilistic arguments relying on a random walk approach to $G_{n,p}$. Theorem 1 raises a myriad of further questions. Two very natural questions are: does $\mathbf{E}\left[diam(MWST(K_n))\right]/n^{1/3}$ converge to a constant? What constant? What about the behavior of the random variable $diam(MWST(K_n))/n^{1/3}$? Theorem 1 seems related not only to the diameter of minimum spanning trees, but also to the diameter of $G_{n,p}$ itself. This latter problem still seems difficult when p gets closer to $1/n$ [7]. A key difference between the analysis required for the two problems is captured by the fact that there is some p^* such that for $p \geq p^*$, the expected diameter of $G_{n,p}$ is decreasing, whereas the diameter of $F_{n,p}$ is increasing for all $0 \leq p \leq 1$. At some point in the range $(p - 1/n) = o(1/n)$, the diameters $G_{n,p}$ and $F_{n,p}$ diverge; the precise behavior of this divergence is unknown. If the expected diameter of $G_{n,p}$ is unimodal, for example, then it makes sense to search for a specific probability p^{**} at which the expected diameters of $G_{n,p}$ and $F_{n,p}$ cease to have the same order. In this case, what can we say about $|p^* - p^{**}|$? For $p = (1+\epsilon)/n$ and $\epsilon > 0$ constant, the diameter of $G_{n,p}$ is concentrated on a finite number of values, whereas it follows from results of [26] that this is not the case in $G_{n,p}$ for $p = 1/n + O(1/n^{4/3})$. How does this behavior change as p increases through the critical window? Answering such questions would seem to be a prerequisite to a full understanding of the diameter of $G_{n,p}$ in the critical range.

References

[1] D. Aldous. A random tree model associated with random graphs. *Random Structures Algorithms*, 4:383–402, 1990.

[2] D. Aldous. Brownian excursions, critical random graphs and the multiplicative coalescent. *Ann. Probab.*, 25:812–854, 1997.

[3] B. Bollobás. *Random Graphs.* Cambridge Studies in Advanced Mathematics. Cambridge University Press, 2nd edition, 2001.

[4] B. Bollobás. The evolution of random graphs. *Trans. Amer. Math. Soc.*, 286(1): 257–274, 1984.

[5] O. Borůvka. O jistém problému minimálním. *Práce Mor. Přírodověd. Spol. v Brně (Acta Societ. Scient. Natur. Moravicae)*, 3:37–58, 1926.

[6] H. Chernoff. A measure of the asymptotic efficiency for tests of a hypothesis based on the sum of observables. *Ann. Math. Statist.*, 2:493–509, 1952.

[7] F. Chung and L. Lu. The diameter of random sparse graphs. *Adv. Appl. Math.*, 26: 257–279, 2001.

[8] T.H. Cormen, C.E. Leiserson, R.L. Rivest, and C. Stein. *Introduction to algorithms.* MIT Press, Cambridge, MA, 2nd edition, 2001.

[9] L. Devroye. Branching processes in the analysis of the heights of trees. *Acta Informatica*, 24:277–298, 1987.

[10] P. Erdős and A. Rényi. On the evolution of random graphs. *Publ. Math. Inst. Hungar. Acad. Sci.*, 5:17–61, 1960.

[11] J.A. Fill and J. M. Steele. Exact expectations of minimal spanning trees for graphs with random edge weights. In A.D. Barbour and L.H.Y. Chen, editors, *Stein's Method and Applications*, pages 169–180, Singapore, 2005. World Publications.

[12] P. Flajolet and A. Odlyzko. The average height of binary trees and other simple trees. *J. Comput. System Sci.*, 25(171–213), 1982.

[13] P. Flajolet, Z. Gao, A. Odlyzko, and B. Richmond. The distribution of heights of binary trees and other simple trees. *Combin. Probab. Comput.*, 2:145–156, 1993.

[14] A. M. Frieze. On the value of a random minimum spanning tree problem. *Discrete Appl. Math.*, 10:47–56, 1985.

[15] A. M. Frieze and C. McDiarmid. Algorithmic theory of random graphs. *Random Structures Algorithms*, 10(1-2):5–42, 1997.

[16] A. M. Frieze and C. J. H. McDiarmid. On random minimum length spanning trees. *Combinatorica*, 9:363–374, 1989.

[17] R. L. Graham and P. Hell. On the history of the minimum spanning tree problem. *IEEE Ann. Hist. Comput.*, 7:43–57, 1985.

[18] S. Janson and J. Spencer. A point process describing the component sizes in the critical window of the random graph evolution. *Combin. Probab. Comput.*, 16:631–658, 2007.

[19] S. Janson, D.E. Knuth, T. Luczak, and B. Pittel. The birth of the giant component. *Random Structures Algorithms*, 4:233–359, 1993.

[20] S. Janson, T. Łuczak, and A. Ruciński. *Random Graphs.* Wiley, New York, 2000.

[21] V. Jarńik. O jistém problému minimálním. *Práce Mor. Přírodověd. Spol. v Brně (Acta Societ. Scient. Natur. Moravicae)*, 6:57–63, 1930.

[22] J. B. Kruskal. On the shortest spanning subtree of a graph and the traveling salesman problem. *Proc. Amer. Math. Soc.*, 2:48–50, 1956.

[23] T. Łuczak. Component behavior near the critical point of the random graph process. *Random Structures Algorithms*, 1(3):287–310, 1990.

[24] T. Łuczak. The number of trees with large diameter. *J. Aust. Math. Soc.*, 58:298–311, 1995.

[25] T. Łuczak. Random trees and random graphs. *Random Structures Algorithms*, 13: 485–500, 1998.

[26] T. Łuczak, B. Pittel, and J.C. Weirman. The structure of a random graph at the point of the phase transition. *Trans. Amer. Math. Soc.*, 341(2):721–748, 1994.

[27] J. Nešetřil. A few remarks on the history of the MST-Problem. *Arch. Math. (Brno)*, 33:15–22, 1997.

[28] B. Pittel. On the largest component of the random graph at a nearcritical stage. *J. Combin. Theory Ser. B*, 82:237–269, 2001.

[29] B. Pittel. Note on the height of random recursive trees and m-ary search trees. *Random Structures Algorithms*, 5:337–347, 1994.

[30] R.C. Prim. Shortest connection networks and some generalizations. *Bell System Techn. J.*, 36:1389–1401, 1957.

[31] A. Rényi and G. Szekeres. On the height of trees. *J. Aust. Math. Soc.*, 7:497–507, 1967.

[32] A.D. Scott and G.B. Sorkin. Solving sparse random instances of max cut and max 2-CSP in linear expected time. *Combin. Probab. Comput.*, 15:281–315, 2006.

[33] R. T. Smythe and H. M. Mahmoud. A survey of recursive trees. *Theory Probab. Math. Statist.*, 51:1–27, 1995.

[34] J. M. Steele. Minimal spanning trees for graphs with random edge weights. In B. Chauvin, P. Flajolet, D. Gardy, and A. Mokkadem, editors, *Mathematics and Computer Science II: Algorithms, Trees, Combinatorics and Probability*, Boston, 2002. Birkhäuser.

[35] V. Vazirani. *Approximation Algorithms*. Springer Verlag, New York, 2001.