Fourth Colloquium on Mathematics and Computer Science DMTCS proc. **AG**, 2006, 237–248

The Diameter of the Minimum Spanning Tree of a Complete Graph

Louigi Addario-Berry[1], Nicolas Broutin[1] and Bruce Reed[1]

[1]*School of Computer Science, McGill University, Montréal, Canada*

Let $\{X_1, \ldots, X_{\binom{n}{2}}\}$ be independent identically distributed weights for the edges of K_n. If $X_i \neq X_j$ for $i \neq j$, then there exists a unique minimum weight spanning tree T of K_n with these edge weights. We show that the expected diameter of T is $\Theta(n^{1/3})$. This settles a question of Frieze and McDiarmid (1997).

Keywords: Minimum Spanning Trees, Random Graphs, Kruskal's Algorithm, Branching Processes, Ballot Theorem

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.

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 e. In this paper we prove the following theorem, answering a question of Frieze and McDiarmid (1997, 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 for all $e \neq f$, $X_e \neq X_f$, it is the case that the expected value of the diameter of* MWST(K_n) *is $\Theta(n^{1/3})$.*

We start with some general properties of minimum spanning trees. 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 $w(e)$. On the other hand, if $e = xy$ is in T then every edge between the component of $T - e$ containing x and the component of $T - e$ containing y has weight at least $w(e)$. 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 (Kruskal, 1956) 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.

Observe first that, if the $w(e)$ are distinct, one does not need to know the weights $\{w(e), e \in E\}$ 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 (1960). 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 $e_m \in \mathit{MWST}(K_n)$ precisely if e_m is a cutedge of $G_{n,m}$.

Using this link, the lower bound is easily obtained. It suffices to note that, with positive probability, $G_{n,n/2}$ contains a tree component T whose size is between $n^{2/3}/2$ and $2n^{2/3}$ (see Janson et al. (2000), Theorem 5.20). This tree is a subtree of $\mathit{MWST}(K_n)$, so $\mathit{diam}(\mathit{MWST}(K_n)) \geq \mathit{diam}(T)$. Conditioned on its size, such a tree is a Cayley tree (uniform labeled tree), and hence has expected diameter $\Theta(n^{1/3})$ (Rényi and Szekeres, 1967; Flajolet and Odlyzko, 1982). Therefore, $\mathbf{E}\{\mathit{diam}(\mathit{MWST}(K_n))\} = \Omega(n^{1/3})$.

The upper bound is much more delicate. To obtain it, we in fact study the random graph process $G_{n,p}$ (Erdős and Rényi, 1960; Janson et al., 2000; Bollobás, 2001): assign an independent $[0,1]$-uniform edge weight $w(e)$ to each edge e of K_n, and for all $p \in [0,1]$, set $G_{n,p} = \{f | w(f) \leq p\}$. Our preference for this model over $G_{n,m}$ is due to the fact that it can be analyzed via a branching process. For this edge weighting, $e \in \mathit{MWST}(G)$ precisely if e is a cutedge of $G_{n,w(e)}$. This implies that the vertex sets of the components of $G_{n,p}$ are precisely the vertex sets of the components of the forest $F_{n,p} = \mathit{MWST}(K_n) \cap \{e | w(e) \leq p\}$ built by Kruskal's algorithm. Actually it implies something stronger: $\mathit{MWST}(K_n) \cap \{e | w(e) \leq p\}$ consists exactly of the unique *MWST*s of the connected components of $G_{n,p}$ under the given weighting. It is this fact which allows us to determine the diameter of $\mathit{MWST}(K_n)$. We shall take a snapshot of $F_{n,p}$ for an increasing sequence of p and examine how this graph evolves via a branching process.

Erdős and Rényi (1960) showed that for any $\epsilon > 0$, if $p < (1-\epsilon)/n$ then a.a.s. (asymptotically almost surely, i.e., with probability tending to 1 as $n \to \infty$), the largest component of $G_{n,p}$ has size $O(\log n)$. If $p > (1+\epsilon)/n$ then a.a.s.:

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

More precisely, they showed that a.a.s. ($\star$) holds for *all* $p > (1+\epsilon)/n$. This implies that a.a.s., for all $p' > p > (1+\epsilon)/n$, $H_{n,p} \subseteq H_{n,p'}$. It turns out that when tracking the diameter of $F_{n,p}$, $0 < p < 1$, the range of interest is essentially the "critical window" *around* $p = 1/n$. For this, we need a refined analysis of the evolution of $G_{n,p}$ close to this critical probability, similar to that provided by Łuczak (1990). He showed that for any function $h(n) > 0$ which is $\omega(n^{-4/3})$ a.a.s. for all $p > 1/n + h(n)$,

(A) $|H_{n,p}| = \omega(n^{2/3})$, and all other components have size $o(n^{2/3})$, and

(B) for all $p' > p$, $H_{n,p} \subseteq H_{n,p'}$.

This fact is crucial to our analysis. Essentially, rather than looking at $F_{n,p}$, we focus on the diameter of $\mathit{MWST}(K_n) \cap H_{n,p}$ 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 $\mathit{lp}(G)$ 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* $\mathrm{diam}(H') \leq \mathrm{diam}(H) + 2\mathrm{lp}(G'[V - V(H)]) + 2$.

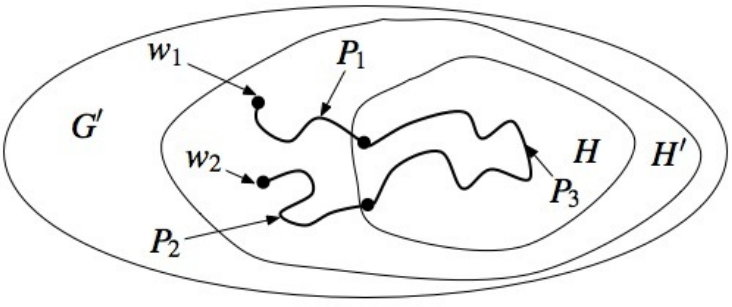

Fig. 1: The path $P = P_1 \cup P_2 \cup P_3$ from w_1 to w_2 in $\mathit{MWST}(H')$.

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 $\mathit{diam}(H) + 2\mathit{lp}(G'[V - V(H)]) + 2$. (See Figure 1) □

If $p < p'$ and $H_{n,p} \subseteq H_{n,p'}$, then Lemma 2 implies that $\mathit{diam}(\mathit{MWST}(H_{n,p'})) \leq \mathit{diam}(\mathit{MWST}(H_{n,p})) + 2\mathit{lp}(G_{n,p'}[V - V(H_{n,p})])$. 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. (This is similar to Łuczak's method of considering "moments" of the graph process (Łuczak, 1990).) For each p_i, we consider the largest component $H_i = H_{n,p_i}$ of G_{n,p_i}. We define dt_i to be the diameter of $\mathit{MWST}(K_n) \cap H_i$.

Amongst the p_i, $0 \leq i \leq t$, we consider a key probability p_{t^*}, which is the (random) time after which the random graph process exhibits "typical" behavior, i.e., for all $p \geq p_{t^*}$ statements akin to (A) and (B), above, hold. Bounds on the probability that such events fail have already been given in Łuczak (1990); the bulk of the work of this paper is in improving these bounds to the point where they yield the expectation bounds we seek.

Our approach is to apply Lemma 2 to figure out deterministic bounds on the differences $\mathit{dt}_{i+1} - \mathit{dt}_i$ for $t^* \leq i < t$. We choose $p_i = 1/n + (3/2)^i \delta_0$. We will show that the components of $G_{n,p_{i+1}}[V - V(H_i)]$ are very likely trees of size $O(n^{2/3}/(3/2)^{2i})$, and thus have expected diameter $O(n^{1/3}/(3/2)^i)$. We insist on a weaker condition as part of the definition of t^*: that for every $i \geq t^*$, the components of $G_{n,p_{i+1}}[V - V(H_i)]$ have diameter $O(n^{1/3}/(3/2)^{i/6})$. Lemma 2 then implies that for every $i \geq t^*$, $\mathit{dt}_{i+1} - \mathit{dt}_i = O(n^{1/3}/(3/2)^{i/6})$. With these differences geometrically this bound implies $\mathit{dt}_t - \mathit{dt}_{t^*} = O(n^{1/3})$.

We use separate arguments to bound dt_{t^*} and to bound the difference between dt_t and the diameter of $\mathit{MWST}(K_n)$. Turning to the latter bound first, we choose $p_t = 1/n + 1/(n \log n)$ and show that $\mathbf{E}\{\mathit{diam}(\mathit{MWST}(K_n))\} - \mathit{dt}_t = O(\log^6 n)$. To do so, we show that with high probability, the size of every component of $\mathit{MWST}(K_n)[V - V(H_t)]$ is $O(\log^6 n)$ and apply Lemma 2. For the moment, we condition on the event that $|H_t| > n/\log n$ and every other component of H_t has at most $\log^3 n$ vertices, and establish our assertion assuming this event holds.

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

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 $|H_t|/n > 1/\log n$, the second endpoint is in H_t. If the second endpoint is not in H_t, we can think of C joining another component to create C'. The component C' has size at most $2\log^3 n$.

Conditional upon this choice of e, the edge e' leaving C' which minimizes $w(e')$ is also in $\mathit{MWST}(K_n)$. Again, with probability at least $1/\log n$ the second endpoint lies in H_t. If not, C' joins another component to create C'' of size at most $3\log^3 n$. Continuing in this fashion, we see that the probability the component containing C has size more than $r\log^3 n$ when it joins to H_t is at most $(1 - 1/\log n)^r$. In particular, the probability that it has size more than $\log^6 n$ 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 reaches size more than $\log^6 n$ before joining H_t. It follows from Lemma 2 that with probability $1 - o(1/n)$, $\mathit{diam}(\mathit{MWST}(K_n)) - \mathit{dt}_t \leq 2\log^6 n + 2$. Since $\mathit{diam}(\mathit{MWST}(K_n))$ never exceeds n, it follows that

$$\mathbf{E}\{\mathit{diam}(\mathit{MWST}(K_n)) - \mathit{dt}_t\} = O(\log^6 n).$$

Our definition of t^* ensures that to establish bounds for $\mathbf{E}\mathit{dt}_{t^*}$, we need only bound $\mathbf{P}\{t^* \geq i\}$ for $1 \leq i \leq t$. Showing that $\mathbf{P}\{t^* \geq i\}$ decreases rapidly as i increases will form a substantial part of the paper.

The paper is roughly organized as follows. In Section 2 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 previous known results. In Section 3 we explain the bounds we need in order to show that for p' not too much larger than p, the diameter of $H_{n,p'}$ is not too much longer than that of $H_{n,p}$, and in Section 4 we bound the diameter of $H_{n,p}$ itself. Finally, Section 5 we explain how to put all the result together to prove Theorem 1, proceeding along the lines outlined above. The proofs of all auxiliary results will appear in the journal version of this paper, and will be provided upon request.

We conclude the introduction by mentioning some previous results about the size and structure of the components $G_{n,p}$ for p in the critical range $p - 1/n = o(1/n)$. This has been studied combinatorially (Bollobás, 1984; Łuczak, 1990, 1991), using generating functions (Janson et al., 1993), and from a weak limit perspective (Aldous, 1990, 1997). For a comprehensive overview of the literature and known results on this subject, the reader is referred to Janson et al. (2000).

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

We analyze the component structure of $G_{n,p}$ using a process similar to breadth-first search (BFS) (Cormen et al., 2001) and to a process used by Aldous (1990) 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 ($0 \leq i < n-1$): Let v be an arbitrary vertex of $\mathcal{O}_i$ and let N_i be the random set of neighbours of v. 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, let us 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 $\text{Bin}(n - i - |\mathcal{O}_i|, p)$. 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 $\max(|\mathcal{O}_i| + \text{Bin}(n - i - |\mathcal{O}_i|, p) - 1, 1)$.

We can thus analyze the growth of each component of $G_{n,p}$, created via the above coupling with the BFS-based process, by coupling the process to the following random walk. Let $S_0 = 1$. For $i \geq 0$, let $X_{i+1} = \text{Bin}(n - i - S_i, p) - 1$, assigned independently, and let

$$S_{i+1} = \max(S_i + X_{i+1}, 1).$$

With this definition, for all i, 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.
- $S^{u,r}$ is the walk with $S^{u,r}_0 = 1$ which estimates the children of a node by setting $S^{u,r}_{i+1} = S^{u,r}_i + \text{Bin}(n - r, p) - 1$, for $i \geq 0$.
- 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 emphasize that until the first visit of S to 0, S' agrees with S while S^{ind} strictly dominates it. On the other hand, $S^{u,r}$ underestimates it until the first time that $i + |\mathcal{O}_i| > r$. Finally, S dominates S^h until the first time that S exceeds $h + 1$. We will rely on the properties of these simpler walks when analyzing S.

A key element of our proof will be to establish the following facts for any p such that $p - 1/n = \Omega(1/n^{4/3})$ and $p - 1/n = O(1/n \log n)$: (1) the largest component of $G_{n,p}$ has size $O(n^2(p - 1/n))$, and (2) that any component of this size must arise early in the branching process. (For the remainder of the paper we presume p falls in this range unless explicitly stated otherwise.) The main goal of the rest of this section is to state and prove precise versions of these claims. To do so, we need to tie down the behavior of S. First, however, we analyze S^{ind},$S^{u,r}$,S^h and S', as they buck a little less wildly.

2.1 The height of the tamer walks

We will focus on how we can control the height of the walks S^{ind} and S'. We choose these two walks for expository purposes and because they show up in the proof of one of the key theorems - Theorem 9, below. Bounds very similar to those we derive for S^{ind} also hold for $S^{u,r}$ and S^h with essentially identical proofs.

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 $t \geq 1$, $S^{ind}_t + (t-1)$ is distributed like $\text{Bin}(nt - \binom{t+1}{2}, p)$, so by linearity of expectation, we have:

Fact 3 *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 $\mathrm{Bin}(n,p)$ random variable is $n(p-p^2)$, we also have that for $p = 1/n+\delta$ with $\delta = o(1/n)$ and $t = o(n)$, $\mathbf{Var}\left\{S_t^{ind}\right\} = (1+o(1))t$. 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 Chernoff (1952) 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.

As we noted above, $\mathbf{Var}\left\{S_t^{ind}\right\} = (1+o(1))t$ for this range of p, 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 2 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.

Theorem 4 (Chernoff, 1952) *If $Y = \mathrm{Bin}(m,q)$ and $\mathbf{E}Y = \lambda = mq$, then for any real number $r > 0$, $\mathbf{P}\{|Y - \mathbf{E}Y| > r\} \leq 2e^{-r^2/2(\lambda+r/3)}$.*

Lemma 5 *Fix $0 < \epsilon < 1$. Then there is $\xi > 0$ such that for all $t \geq 1$ and $t = o(1/n)$,*

$$\mathbf{P}\left\{\left|S_t^{ind} - \mathbf{E}S_t^{ind}\right| > \frac{\epsilon tf}{n^{1/3}}\right\} \leq 2e^{-\xi tf^2/n^{2/3}}.$$

Proof of Lemma 5: The tail bound on S_t^{ind} is obtained by applying Theorem 4 to $S_t^{ind} + (t-1)$, which is a binomial random variable. Before applying it, we observe that by Fact 3, $\lambda = \mathbf{E}S_t^{ind} + (t-1) \leq 2t$. Thus

$$\mathbf{P}\left\{\left|S_t^{ind} - \mathbf{E}S_t^{ind}\right| > \frac{\epsilon tf}{n^{1/3}}\right\} \leq 2e^{-(\epsilon tf/n^{1/3})^2/(4t+2\epsilon tf/3n^{1/3})} \leq 2e^{-\xi tf^2/n^{2/3}},$$

where $\xi = \epsilon^2/5$. In the second inequality we use that $4t + 2\epsilon tf/3n^{1/3} < 5t$, always true as $f < n^{1/3}$ and $\epsilon < 1$. □

In bounding the height of S, the fact that we set $S_t = 1$ when $S_{t-1} + X_t = 0$ complicates our lives considerably. To handle these complications 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 can obtain an upper bound on S_t by combining Lemma 5 with the following lemma, which provides bounds on Z_t.

Lemma 6 *Fix $0 < \epsilon < 1$. Then there exists $\xi > 0$ such that for any $0 < \alpha < 1$, for n large enough, for $f > (16/\epsilon)^2$ and $t \in [n^{2/3}/f, (2-\alpha)fn^{2/3}]$,*

$$\mathbf{P}\left\{\min_{1\leq i\leq t} S'_i \leq \frac{-\epsilon tf}{n^{1/3}}\right\} \quad \leq \quad \frac{1}{4n^2} + e^{-\xi tf^2/n^{2/3}}. \tag{1}$$

The proof of this lemma proceeds by comparing S' and S^h for $h = \epsilon t$ for some small $\epsilon > 0$. As $S'_i \geq S_i^h$ until the first time $|\mathcal{O}_i| > h+1$, if $\min_{1\leq i\leq t} S'_i$ is small then either $\min_{1\leq i\leq t} S_i^h$ is small or $S_i = |\mathcal{O}_i| > h+1$ for some $1 \leq i \leq t$. We bound the probability of the latter event by comparing S_i to $S_{u,t}$ and applying bounds analogous to those given in Lemma 5 for S^{ind}. To bound the probability of the former event, we end up needing a binomial analogue of the *ballot theorem*. The ballot theorem states that if $0 = T_0, T_1, \ldots$ is a symmetric simple random walk (so $T_{i+1} = T_i \pm 1$, each with probability $1/2$ and independently of all previous steps), then for any $k > 0$, the probability that T_i is positive for all $i = 1, 2, \ldots, n$ and $T_n = k$ is precisely k/n times the probability that $T_n = k$ (see Grimmett and Stirzaker (1992, page 77)). We need:

Lemma 7 *Fix $0 < q < 1$ and integers $\{m_i | i \geq 1\}$ satisfying $m_i q \leq 2$ for all i. Let $\{X_i | i \geq 1\}$ be independent random variables, X_i distributed as $\mathrm{Bin}(m_i, q)$. Let U be the simple random walk with $U_0 = 0$ and, for $i \geq 1$, $U_i = \sum_{j=1}^{i} X_i$. Then for any integers $r \geq 8$, $s > 0$,*

$$\mathbf{P}\left\{U_s = \mathbf{E}U_s - r \text{ and } U_i < \mathbf{E}U_i \; \forall 1 \leq i \leq s\right\} \leq \frac{64e^{-r^2/(64s+8r/3)}}{r^2}.$$

To justify calling this an analogue of the ballot theorem, note that if $r = \Theta(\sqrt{s})$ then $e^{-r^2/s} = \Theta(1)$. In this case $\mathbf{P}\{U_s = \mathbf{E}U_s - r\} = \Theta(1/r)$ by standard binomial estimates. Ignoring constants, we can write the bound of the previous lemma as $1/r^2 = 1/s = (r/s)(1/r) = (r/s)\mathbf{P}\{U_s = \mathbf{E}U_s - r\}$. We note that by applying a generalized ballot theorem of Takács (1967, Page 10), a lower bound of the same order follows. We include the proof of Lemma 7 here as we believe it is of some independent interest. We note that using a similar approach, Addario-Berry and Reed (2006) have proved a ballot-style theorem that holds for any mean zero random walk whose step size has bounded variance. In proving Lemma 7, we will have use of the following simple fact.

Fact 8 *Given any binomial random variable $B = \mathrm{Bin}(m, q)$ and integer $a \geq 1$, $\mathbf{P}\{B = \lceil \mathbf{E}B \rceil + a\} \leq \mathbf{P}\{B > \mathbf{E}B + a/2\} / \lceil a/2 \rceil$.*

This is clear as for any a' such that $a/2 < a' \leq a$, it is at least as likely that $B = \mathbf{E}B + a'$ as that $B = \mathbf{E}B + a$.

Proof of Lemma 7: We analyze a refinement of U in which the terms are Bernoulli random variables instead of binomials. We can write $X_i = \sum_{j=1}^{m_i} B_{i,j}$, where the $B_{i,j}$ are i.i.d. Bernoulli random variables with mean q. We assume for simplicity that $qz = 1$ for some natural number $z \geq 1$; in the full version of the paper we explain how to eliminate this technical restriction. Let Z be the random walk in which $Z_0 = 0$ and for $t \geq 1$, Z_t is the sum of all $B_{i,j}$ satisfying $j + \sum_{i'<i} m_{i'} \leq t$. We will also think of this walk as being indexed by pairs (i, j) via the bijection $(i, j) \leftrightarrow (\sum_{i'<i} m_{i'}) + j$. We use these two forms of indices for Z interchangeably. The random walk U is determined by Z; more strongly, for all i and all $j \leq m_i$, $U_i = Z_{i,m_i} \geq Z_{i,j}$. Let $t_s = \sum_{i=1}^{s} m_i$, so that $U_s = Z_{t_s}$.

We define a sequence of stopping times for Z. Let $T_0 = 0$, and for $j > 0$, let $T_j = (V_j, F_j)$ be the first time greater than T_{j-1} for which

$$|(Z_{T_j} - \mathbf{E}Z_{T_j}) - (Z_{T_{j-1}} - \mathbf{E}Z_{T_{j-1}})| = 2^j.$$

We will also write (V_j, F_j) in place of T_j, where the pair (V_j, F_j) is obtained via the bijection noted above. Let the difference above be called Y_j; then for all $j > 0$, $Y_j = \pm 2^j$.

Note that for all $k > 0$, $Z_{T_{k-1}} - \mathbf{E}Z_{T_{k-1}} \geq 2 - 2^k$, so if Y_k is positive then $Z_{T_k} - \mathbf{E}Z_{T_k} \geq 2$. Furthermore,

$$\begin{aligned}
\mathbf{E}U_{V_k} - \mathbf{E}Z_{T_k} &= \mathbf{E}Z_{V_k,m_{V_k}} - \mathbf{E}Z_{V_k,F_k} \\
&\leq \sup_{(v,f):0\leq f\leq m_v} \mathbf{E}\left\{Z_{V_k,m_{V_k}} - Z_{V_k,F_k} | (V_k, F_k) = (v, f)\right\} \\
&= \sup_{(v,f):0\leq f\leq m_v} \mathbf{E}\left\{Z_{v,m_v} - Z_{v,f}\right\} \leq 2,
\end{aligned}$$

as $(m_v - f)q \leq 2$ for all such pairs (v, f). Thus if ever $Z_{T_k} - \mathbf{E}Z_{T_k} \geq 2$ then

$$U_{V_j} - \mathbf{E}U_{V_j} \geq Z_{T_j} - \mathbf{E}U_{V_j} \geq Z_{T_j} - \mathbf{E}Z_{T_j} - 2 \geq 0.$$

Let E be the event that $U_s = \mathbf{E}\{U_s\} - r$ and $U_i < \mathbf{E}\{U_i\}$ for all $1 \leq i \leq s$, and set $b = \lfloor \log_2 r \rfloor - 2$; b is at least one as $r \geq 8$. It follows from the above comments that if E is to occur, the following three events must hold:

- $E_1 = \cap_{i=1}^{b} \{Y_i < 0\} = \{(Z_{T_b} - \mathbf{E}Z_{T_b}) = 2 - 2^{b+1}\}$,
- $E_2 = \{T_b < t_s\}$, and
- $E_3 = \{Z_{t_s} - \mathbf{E}\{Z_{t_s}\} - (Z_{T_b} - \mathbf{E}Z_{T_b}) = -(r + 2 - 2^{b+1})\}$.

We bound $\mathbf{P}\{E\}$ by writing

$$\mathbf{P}\{E\} = \mathbf{P}\{E_1 \cap E_2\}\,\mathbf{P}\{E_3|E_1 \cap E_2\} \leq \mathbf{P}\{E_1\}\,\mathbf{P}\{E_3|E_1 \cap E_2\}. \quad (2)$$

To bound E_1, write

$$\begin{aligned}\mathbf{E}\{Y_i\} &= \mathbf{E}\{Y_i|Y_i > 0\}\,\mathbf{P}\{Y_i > 0\} + \mathbf{E}\{Y_i|Y_i < 0\}\,\mathbf{P}\{Y_i < 0\} \\ &= 2^i\mathbf{P}\{Y_i > 0\} - 2^i\mathbf{P}\{Y_i < 0\}. \end{aligned} \quad (3)$$

Since T is a stopping time, $\mathbf{E}Y_i = 0$ by Wald's Identity (see, e.g., Grimmett and Stirzaker (1992), Page 493), so $\mathbf{P}\{Y_i > 0\} = \mathbf{P}\{Y_i < 0\} = 1/2$. For $j \neq i$, $\{Y_i > 0\}$ and $\{Y_j > 0\}$ are determined on disjoint intervals of the random walk: they are sums of disjoint sets of independent random variables, and hence are independent. Thus $\mathbf{P}\{E_1\} = \mathbf{P}\left\{\cap_{i=1}^{b}\{Y_i > 0\}\right\} = 1/2^b \leq 8/r$.

Next, note that $E_1 \cap E_2$ is determined by $Z_1, \ldots, Z_{T_b}$. As E_3 is determined by $Z_{T_b+1}, \ldots, Z_{t_s}$ and Z has independent increments, we have

$$\begin{aligned}\mathbf{P}\{E_3|E_1 \cap E_2\} &\leq \max_{1\leq t\leq t_s} \mathbf{P}\{E_3|E_1 \cap \{T_b = t_s - t\}\} \\ &= \max_{1\leq t\leq t_s} \mathbf{P}\Big\{Z_{t_s} - Z_{t_s-t} = \mathbf{E}\{Z_{t_s} - Z_{t_s-t}\} - (r + 2 - 2^{b+1})\Big| \\ &\qquad\qquad E_1 \cap \{T_b = t_s - t\}\Big\} \\ &= \max_{1\leq t\leq t_s} \mathbf{P}\left\{Z_{t_s} - Z_{t_s-t} = \mathbf{E}\{Z_{t_s} - Z_{t_s-t}\} - (r + 2 - 2^{b+1})\right\} \\ &\leq \max_{1\leq t\leq t_s} \mathbf{P}\{\mathrm{Bin}(t, q) = tq - r/2\}, \end{aligned}$$

where we use in the last step that $(r+2-2^{b+1}) > r/2$. For any such t, $\mathbf{E}\mathrm{Bin}(t, q) = tq \leq \sum_{i=1}^{s} m_i q \leq 2s$, so by Fact 8 and Theorem 4, we have

$$\begin{aligned}\mathbf{P}\{\mathrm{Bin}(t, q) = tq - r/2\} &\leq \mathbf{P}\{\mathrm{Bin}(t, q) \leq tq - r/4\} \cdot 4/r \\ &\leq \frac{8e^{-r^2/16(4s+r/6)}}{r} \leq \frac{8e^{-r^2/(64s+8r/3)}}{r}. \end{aligned}$$

Finally, substituting this bound and the bound on $\mathbf{P}\{E_1\}$ into (2) yields

$$\mathbf{P}\{E\} \leq \frac{8}{r}\frac{8e^{-r^2/(64s+8r/3)}}{r} = \frac{64e^{-r^2/(64s+8r/3)}}{r^2},$$

as claimed. □

2.2 The height of S

Using these bounds on the random variables S^{ind} and S', we are now able to derive bounds on the height of S. For any $0 < \alpha < 1$ let $Z = Z(\alpha)$ be the event that $S_t = 0$ for some $n^{2/3}/f \leq t \leq (2-\alpha)fn^{2/3}$, and let $N = N(\alpha)$ be the event that $S_t \neq 0$ for any $(2-\alpha)fn^{2/3} \leq t \leq (2+\alpha)fn^{2/3}$.

Theorem 9 *Fix $0 < \alpha < 1$. Then there exist constants $\xi = \xi > 0$, $C > 0$, and $F > 1$ such that for all $f = \omega(F)$ and for n large enough,*

$$\mathbf{P}\{Z(\alpha) \cup N(\alpha)\} \leq \frac{2}{n} + Ce^{-\xi f}.$$

We obtain the bound on $\mathbf{P}\{N\}$ via a straightforward comparison of S' and S^{ind}. The bound on $\mathbf{P}\{Z\}$ is an immediate consequence of the following lemma:

Lemma 10 *Fix $0 < \alpha < 1$. Then there are constants $\xi > 0$, $F_1 > 1$, and $1 < c < 2$ such that for all $f > F_1$ and for n large enough, for any $t \in [n^{2/3}/f, (2-\alpha)n^{2/3}f]$,*

$$\mathbf{P}\{S_i = 0 \textit{ for some } i \textit{ in } [t, ct]\} \leq \frac{3}{n^2} + 17e^{-\xi t f^2/n^{2/3}}. \quad (4)$$

The proof of this lemma has much the same flavour as the proof of Lemma 6 - it is proved by comparing S_t to S_t^h for a carefully chosen h to prove that S_t and S_{ct} are both very likely close to their expected value, then showing that if both S_t and S_{ct} are close to their expected value, it is very unlikely that S visited zero at any time between t and ct.

Proof of Theorem 9: For $i \geq 0$, let $t_i = c^i n^{2/3}/f$, where c is the constant from Lemma 10, and let k be minimal so that $c^k \geq (2-\alpha)fn^{2/3}$; clearly $k = O(\log f) = o(n)$. For $0 \leq i < k$ let Z_i be the event that there exists $t_i \leq t \leq t_{i+1}$ such that $S_t = 0$. Then

$$\begin{aligned}
\mathbf{P}\{Z\} &\leq \sum_{i=0}^{k-1} \mathbf{P}\{Z_i\} \\
&\leq \sum_{i=0}^{k-1} \left(\frac{3}{n^2} + 17e^{-\xi c^i t_0 f^2/n^{2/3}} \right) \\
&= \frac{3k}{n^2} + 17\sum_{i=0}^{k-1} e^{-c^i \xi f} \leq \frac{1}{n} + C_0 e^{-\xi f}, \qquad (5)
\end{aligned}$$

for some C_0 and for $f \geq F_1$ and n large enough, as $k = o(n)$ and as the terms of the last sum are super-geometrically decreasing.

Turning our attention to N, let $\underline{t} = (2-\alpha)fn^{2/3}$, $\bar{t} = (2+\alpha)fn^{2/3}$. Recall that the first visit of S to zero after time $\underline{t}$ occurs for the smallest $k > \underline{t}$ such that $S'_k = \min\{S'_i | 1 \leq i \leq \underline{t}\} - 1$. As $S^{ind} \geq S'$, it follows that

$$\mathbf{P}\{N\} \leq \mathbf{P}\left\{ S^{ind}_{\bar{t}} > \min_{1 \leq i \leq \underline{t}} S'_i - 1 \right\}. \qquad (6)$$

Call the latter event E. We now show that for E to occur, one of $S^{ind}_{\bar{t}}$ or $\min\{S'_i | 1 \leq i \leq \underline{t}\}$ must be far from its expected value. Let $\epsilon = \alpha/4$. We apply Lemma 6 to obtain that as long as $f > (16/\epsilon)^2$, it is very unlikely that $\min\{S'_i | 1 \leq i \leq \underline{t}\} \leq -\epsilon \underline{t} f/n^{1/3}$. By Fact 3, we also have that

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

By this bound and by Lemmas 6 and 5, for $f \geq (16/\epsilon)^2$ and for n large enough,

$$\begin{aligned}
\mathbf{P}\{E\} &\leq \mathbf{P}\left\{ S^{ind}_{\bar{t}} > \frac{-\epsilon \underline{t} f}{n^{1/3}} \right\} + \mathbf{P}\left\{ \min_{1 \leq i \leq t} S'_i \leq \frac{-\epsilon \underline{t} f}{n^{1/3}} \right\} \\
&\leq \mathbf{P}\left\{ S^{ind}_{\bar{t}} - \mathbf{E}S^{ind}_{\bar{t}} > \frac{-\epsilon \bar{t} f}{n^{1/3}} \right\} + \frac{1}{4n^2} + e^{-\xi \underline{t} f^2/n^{2/3}} \\
&\leq 2e^{-\xi \bar{t} f^2/n^{2/3}} + \frac{1}{4n^2} + e^{-\xi \underline{t} f^2/n^{2/3}} \\
&\leq \frac{1}{4n^2} + 3e^{-\xi \underline{t} f^2/n^{2/3}} \leq \frac{1}{4n^2} + 3e^{-\xi f^3}. \qquad (7)
\end{aligned}$$

The bounds on $\mathbf{P}\{Z(\alpha) \cup N(\alpha)\}$ follows immediately from (5), (6) and (7). The theorem follows by setting $F = \max(F_1, (16/\epsilon)^2)$. □

2.3 The final stages of the process

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 $\alpha < 1/2$, say, then

$$\begin{aligned}
p &= \frac{1}{n} + \frac{(2-\alpha)f}{n^{4/3}} \leq \frac{1}{n'}\left(1 - \frac{(2-\alpha)f}{n^{1/3}}\right) + \frac{f}{n^{4/3}} \\
&< \frac{1}{n'} - \frac{(f/2)}{(n')^{4/3}}. \qquad (8)
\end{aligned}$$

So, we can use the following result (Łuczak, 1990, Lemma 1) to bound the sizes of the components constructed by the rest of the procedure.

Theorem 11 *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 tree or unicyclic 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 $3e^{-k}$.*

Combining this with Theorem 9, we obtain:

Theorem 12 *There are $C > 0$, $\xi > 0$ and $F > 1$ such that for $f > F$, with probability at least $1 - (2/n) - Ce^{-\xi f}$, the random graph $G_{n,p}$ contains one component of size at least $(2 - 2\alpha)fn^{2/3}$ and every other component has size at most $n^{2/3}/f$.*

Furthermore, a similar analysis (implicit in Łuczak (1990)) also bounds the variance of the number of giant components.

Theorem 13 *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 $n^{2/3}$ is at most $1 + \epsilon$.*

3 Moving between the snapshots

As discussed in the introduction, we are not interested in the behavior of the components of $G_{n,p}[V - V(H_{n,p})]$, but rather in the components of $G_{n,p'}[V - V(H_{n,p})]$. We can get a handle on these by applying Theorem 9, together with existing knowledge about subcritical random graphs.

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 $(2 - 2\alpha)fn^{2/3}$ and $(2 + 2\alpha)fn^{2/3}$ vertices. For $H \in \mathcal{H}$, let C_H be the event that H is a connected component of $G_{n,p}$, and let B be the event that no element of $\mathcal{H}$ is a connected component of $G_{n,p}$. If B occurs then $G_{n,p}$ has *no* component of size between $(2-2\alpha)fn^{2/3}$ and $(2+2\alpha)fn^{2/3}$, so using Theorem 12 to bound $\mathbf{P}\{B\}$, for *any* event E we can therefore write

$$\begin{aligned}\mathbf{P}\{E\} &\leq \mathbf{P}\{B\} + \sum_{H\in\mathcal{H}} \mathbf{P}\{E|C_H\}\,\mathbf{P}\{C_H\} \\ &\leq \frac{2}{n} + Ce^{-\xi f} + (\max_{H\in\mathcal{H}} \mathbf{P}\{E|C_H\})\mathbf{E}\{|\{H \,:\, C_H \text{ holds }\}|\}.\end{aligned}$$

Now applying Theorem 13 to bound the above expectation, we have that for f large enough,

$$\mathbf{P}\{E\} \leq \frac{2}{n} + Ce^{-\xi f} + 2(\max_{H\in\mathcal{H}} \mathbf{P}\{E|C_H\}). \tag{9}$$

Given *any* component $H \in \mathcal{H}$, the graph $G_{n,p'}[V - V(H)]$ is $G_{n',p'}$ for some $n' \leq (2 - 2\alpha)fn^{2/3}$. Supposing that $p' - p = (1/2)f/n^{4/3}$ for some fixed, if α is chosen small enough then a calculation such as (8) shows that $p' \leq 1/n' - (\alpha/2)f/(n')^{4/3}$. Therefore, Theorem 11 will apply to this graph as in Section 2.3. We henceforth assume α has been chosen small enough so that this bound on p' indeed holds. The following reformulation of Łuczak (1998, Theorem 11) also applies to $G_{n,p'}[V - V(H)]$.

Theorem 14 *There exists $F'' > 1$ such that for all $f > F''$, for n large enough and $p = 1/n - f/n^{4/3}$, the probability there is a component of $G_{n,p}$ with excess at most 0, size at most $n^{2/3}/f$ and longest path at least $12n^{1/3}\log f/\sqrt{f}$ is at most $e^{-\sqrt{f}}$.*

We thus let $Long$ be the event that some component of $G_{n,p'}[V - V(H_{n,p})]$ has longest path of length at least $n^{1/3}/f^{1/4}$. It follows easily from Theorems 11 and 14 that for any $H \in \mathcal{H}$, $\mathbf{P}\{Long|C_h\} \leq 2e^{-\sqrt{f}}$. By (9), we thus have

Lemma 15 *There exists $F > 1$ such that for $f > F$, for n large enough, $\mathbf{P}\{Long\} \leq \frac{2}{n} + 5e^{-\sqrt{f}}$.*

4 Bounding the diameter of the giant

A connected component K of $G_{n,p}$ is a uniformly chosen labeled graph with $|V(K)|$ vertices and $|E(K)|$ edges. As we show below, if K is not the giant component then $|E(K)|$ is not much larger than $|V(K)|$. This is what allows us to bound the diameter.

The quantity $q(K) = |E(K)| - |V(K)|$ is called the *excess* of K. If $q(K) = -1$, then K is a tree, so is a uniformly random labeled tree. Rényi and Szekeres (1967) have calculated the moments of the height of such a tree, and it is a straightforward calculation using this information to show:

Lemma 16 *Let K be a tree component of $G_{n,p}$. Then*

$$\mathbf{P}\left\{\mathrm{lp}(K) > t\sqrt{8\pi|K|}\right\} \leq t^{-t/2}.$$

It is possible to extend this result to the setting when $q(K)$ is not too large, so the structure of K is still rather treelike. In this case the length of the longest path in K is likely not much longer the longest path in a tree component of the same size:

Theorem 17 *Let K be a connected component of $G_{n,p}$ with $|E(K)| - |V(K)| = q \geq -1$. Then for any $t \geq 5$,*

$$\mathbf{P}\left\{\mathrm{lp}(K) \geq t(5q+6)\sqrt{8\pi|V(K)|} + 6q\right\} \leq (5q+6)t^{-t/2}. \tag{10}$$

To bound the lengths of the longest path of $H_{n,p}$, we can thus bound its excess and apply the above theorem.

4.1 Bounding the excess

The net excess of the giant component of $G_{n,p}$ can be analyzed much as we have analyzed its size. In the process defined at the beginning of Section 2, 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, then the net excess of this component is precisely $\sum_{i=t_1}^{t_2-1} \mathrm{Bin}(|\mathcal{O}_i| - 1, p) = \sum_{i=t_1}^{t_2-1} \mathrm{Bin}(S_i - 1, p)$. Thus, upper bounds on S provide upper bounds on the net excess of components of $G_{n,p}$. Theorem 9, together with the other information on the sizes of the components of $G_{n,p}$ we derived above, tells us that the component H of $G_{n,p}$ alive at time $(2-\alpha)fn^{2/3}$ of the random walk S is very likely the giant component. We are then able to bound the net excess of H by analyzing the height of S in much the same way as we did to prove Theorem 9 to prove:

Lemma 18 *Let Net be the event that $H_{n,p}$ has excess at most $20f^3$. There exist constants $F_6 > 0$, $C > 0$, and $\xi > 0$ such that n large enough, for all $f > F_6$,*

$$\mathbf{P}\left\{\overline{Net}\right\} \leq \frac{4}{n} + Ce^{-\xi f}. \tag{11}$$

This strengthens the previously known bounds given in Łuczak (1990) on the excess of the largest component of $G_{n,p}$ for p in this range.

5 Putting it all together

To prove Theorem 1, we follow the strategy described in Section 1. Recall that $H_{n,p}$ is the largest component of $G_{n,p}$. We let $p_i = 1/n + f_r/n^{4/3}$, where $f_r = (3/2)^i f_0$ and f_0 is chosen large enough so that the relevant lemmas and theorems of the preceding sections apply to G_{n,p_0} Let k be the smallest integer for which $n^{1/3}/\log n \leq f_k$ — clearly $k < 2\log n$. Given $1 \leq r \leq k$, we are interested in the following "good events":

- $E_{1,r}$ is the event that H_{n,p_r} has size between $(2-2\alpha)f_r n^{2/3}$ and $(2+2\alpha)f_r n^{2/3}$, and the excess of H_{n,p_r} is at most $20f_r^3$.
- $E_{2,r}$ is the event that the longest path in H_{n,p_r} has length at most $f_r^4 n^{1/3}$.
- $E_{3,r}$ is the event that every component $G_{n,p_{r+1}}[V - V[H_{n,p_r}]]$ has size at most $n^{2/3}/f_r$, excess at most $f_r^{1/4}$, and longest path of length at most $n^{1/3}/f_r^{1/4}$.

Let r^* be the smallest value for which $E_{1,r}, E_{2,r}$, and $E_{3,r}$ occur for every $r^* \leq r \leq k$, that is, the time from which the process exhibits "good" behavior. By Lemma 2 and a geometric sum, it is *deterministically* the case that

$$\mathit{diam}(H_{n,p_k}) = \mathit{diam}(H_{n,p_{r^*}}) + 10n^{1/3}/f_{r^*}^{1/6} \leq 2f_{r^*}^4 n^{1/3}. \tag{12}$$

If $r^* = i+1$, then one of the events $E_{1,i}, E_{2,i}, E_{3,i}$ fails, or else r^* would be i, not $i+1$. We have already seen bounds on these events failing in the previous sections; combining them yields that there is some constant C such that:

$$\mathbf{P}\{r^* = i+1\} \leq \frac{12k}{n} + Ce^{-f_i^{1/2}}. \tag{13}$$

Combining this equation with with (12), and the fact that a path has length no longer than n yields that

$$\begin{aligned}
\mathbf{E}\left\{\mathit{diam}(\mathit{MWST}(H_{n,p_k}))\right\} &\leq \sum_{i=0}^{k-1} \min\left(2f_{i+1}^4 n^{1/3}, n\right) \mathbf{P}\left\{r^* = i+1\right\} \\
&\leq \sum_{i=0}^{k-1} \min\left(2f_{i+1}^4 n^{1/3}, n\right) \left(\frac{12k}{n} + Ce^{-f_i^{1/2}}\right) \\
&\leq 48\log^2 n + 2Cn^{1/3} \sum_{i=1}^{k} f_i^4 e^{-f_{i-1}^{1/2}} = O(n^{1/3}).
\end{aligned}$$

As we saw in the introduction, if $|H_{n,p_k}| > cn/\log n$ for some $c > 0$ and all other components have size $O(\log^3 n)$, then

$$\mathbf{E}\left\{\mathit{diam}(\mathit{MWST}(K_n)) - \mathit{diam}(\mathit{MWST}(H_{n,p_k}))\right\} = O(\log^6 n). \tag{14}$$

Our above bounds on component sizes from Sections 3, 4, and 5 show that this indeed holds with probability $1 - O(1/n)$, so this expectation bound is valid. This establishes that $\mathbf{E}\left\{\mathit{diam}(\mathit{MWST}(K_n))\right\} = O(n^{1/3})$, completing the proof of Theorem 1.

6 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 a 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 $\mathbf{E}\left\{\mathit{diam}(\mathit{MWST}(K_n))\right\}/n^{1/3}$ converge to a constant? What constant? 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 $1/n$ (Chung and Lu, 2001). A key difference between the analysis required for the two problems is captured by the fact that there is some (random) p^* such that for $p \geq p^*$, the 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 $|p^* - p^{**}|$? 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

L. Addario-Berry and B. Reed. A generalized ballot theorem. manuscript, 2006.

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

D. Aldous. Brownian excursions, critical random graphs and the multiplicative coalescent. *Annals of Probability*, 25 (2):812–854, 1997.

B. Bollobás. *Random Graphs*. Cambridge Studies in Advanced Mathematics. Cambridge University Press, second edition, 2001.

B. Bollobás. The evolution of random graphs. *Transactions of the American Mathematical Society*, 286(1): 257–274, 1984.

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.

F. Chung and L. Lu. The diameter of random sparse graphs. *Advances in Applied Mathematics*, 26:257–279, 2001.

T.H. Cormen, C.E. Leiserson, R.L. Rivest, and C. Stein. *Introduction to algorithms*. MIT Press, Cambridge, MA, second edition, 2001.

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

P. Flajolet and A. Odlyzko. The average height of binary trees and other simple trees. *Journal of Computer and System Sciences*, 25(171–213), 1982.

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

G.R. Grimmett and D.R. Stirzaker. *Probability and Random Processes*. Clarendon, Oxford, second edition, 1992.

S. Janson, D.E. Knuth, T. Luczak, and B. Pittel. The birth of the giant component. *Random Structures nd Algorithms*, 4(3):233–359, 1993.

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

J. B. Kruskal. On the shortest spanning subtree of a graph and the traveling salesman problem. *Proc. American Math. Society*, 2:48–50, 1956.

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

T. Łuczak. Cycles in a random graph near the critical point. *Random Structures and Algorithms*, 2:421–440, 1991.

T. Łuczak. Random trees and random graphs. *Random Structures and Algorithms*, 13(3-4):485–500, 1998.

A. Rényi and G. Szekeres. On the height of trees. *Journal of the Australian Mathematical Society*, 7: 497–507, 1967.

L. Takács. *Combinatorial Methods in the Theory of Stochastic Processes*. John Wiley & Sons, New York, NY, 1967.