The largest hole in sparse random graphs

Nemanja Draganić * Stefan Glock † Michael Krivelevich ‡

Abstract

We show that for any $d = d(n)$ with $d_0(\epsilon) \leq d = o(n)$, with high probability, the size of a largest induced cycle in the random graph $G(n, d/n)$ is $(2 \pm \epsilon)\frac{n}{d} \log d$. This settles a long-standing open problem in random graph theory.

1 Introduction

Let $G(n,p)$ denote the binomial random graph on n vertices, where each edge is included independently with probability p. In this paper, we are concerned with *induced* subgraphs of $G(n,p)$, specifically trees, forests, paths and cycles.

The study of induced trees in $G(n,p)$ was initiated by Erdős and Palka [11] in the 80s. Among other things, they showed that for constant p, with high probability (**whp**) the size of a largest induced tree in $G(n,p)$ is asymptotically equal to $2\log_q(np)$, where $q = \frac{1}{1-p}$. The obtained value coincides asymptotically with the *independence number* of $G(n,p)$, the study of which dates back even further to the work of Bollobás and Erdős [5], Grimmett and McDiarmid [18] and Matula [23].

As a natural continuation of their work, Erdős and Palka [11] posed the problem of determining the size of a largest induced tree in *sparse* random graphs, when $p = d/n$ for some fixed constant d. More precisely, they conjectured that for every $d > 1$ there exists $c(d) > 0$ such that **whp** $G(n,p)$ contains an induced tree of order at least $c(d) \cdot n$. This problem was settled independently in the late 80s by Fernandez de la Vega [12], Frieze and Jackson [16], Kučera and Rödl [20] as well as Łuczak and Palka [22]. In particular, Fernandez de la Vega [12] showed that one can take $c(d) \sim \frac{\log d}{d}$, and a simple first moment calculation reveals that this is tight within a factor of 2. Here and throughout, log denotes the natural logarithm if no base is specified.

Two natural questions arise from there. First, one might wonder whether it is possible to find not only some *arbitrary* induced tree, but a *specific* one, say a long induced path. Indeed, Frieze and Jackson [15] in a separate paper showed that **whp** there is an induced path of length $\tilde{c}(d) \cdot n$. Two weaknesses of this result were that their proof only worked for sufficiently large d, and that the value obtained for $\tilde{c}(d)$ was far away from the optimal one. Later, Łuczak [21] and Suen [27] independently remedied this situation twofold. They proved that an induced path of length linear in n exists for all $d > 1$, showing that the conjecture of Erdős and Palka holds even for induced paths. Moreover, they showed that one can take $\tilde{c}(d) \sim \frac{\log d}{d}$ as in the case of arbitrary trees.

A second obvious question is to determine the size of a largest induced tree (and path) more precisely. The aforementioned results were proved by analysing the behaviour of certain

2020 *Mathematics Subject Classification.* Primary 05C80, 05C38; Secondary 05C05, 05D40.

Key words and phrases. random graph, induced path, hole.

*Department of Mathematics, ETH, 8092 Zürich, Switzerland. Email: nemanja.draganic@math.ethz.ch.

†Institute for Theoretical Studies, ETH, 8092 Zürich, Switzerland. Email: dr.stefan.glock@gmail.com.
Research supported by Dr. Max Rössler, the Walter Haefner Foundation and the ETH Zürich Foundation.

‡School of Mathematical Sciences, Raymond and Beverly Sackler Faculty of Exact Sciences, Tel Aviv University, Tel Aviv, 6997801, Israel. Email: krivelev@tauex.tau.ac.il. Research supported in part by USA-Israel BSF grant 2018267, and by ISF grant 1261/17.

constructive algorithms which produce large induced trees and paths. The value $\frac{\log d}{d}$ seems to constitute a natural barrier for such approaches. On the other hand, recall that in the dense case, the size of a largest induced tree coincides asymptotically with the independence number. In 1990, Frieze [14] showed that the first moment bound $\sim 2\frac{n}{d}\log d$ is tight for the independence number, also in the sparse case. His proof is based on the profound observation that the second moment method can be used even in situations where it apparently does not work, if one can combine it with a strong concentration inequality. Finally, in 1996, Fernandez de la Vega [13] observed that the earlier achievements around induced trees can be combined with Frieze's breakthrough to prove that the size of a largest induced tree is indeed $\sim 2\frac{n}{d}\log d$. This complements the result of Erdős and Palka [11] in the dense case. (When $p = o_n(1)$, we have $2\log_q(np) \sim 2\frac{n}{d}\log d$.)

Fernandez de la Vega [13] also posed the natural problem of improving the Łuczak–Suen bound [21, 27] for induced paths, for which his approach was "apparently helpless". Despite the widely held belief (see [7, 10] for instance) that the upper bound $\sim 2\frac{n}{d}\log d$ obtained via the first moment method is tight, the implicit constant 1 has not been improved in the last 30 years.

1.1 Long induced paths and cycles

Our main result is the following, which settles this problem and gives an asymptotically optimal result for the size of a largest induced path in $G(n,p)$.

Theorem 1.1. *For any $\epsilon > 0$ there is d_0 such that **whp** $G(n,p)$ contains an induced path of length at least $(2-\epsilon)\frac{n}{d}\log d$ whenever $d_0 \le d = pn = o(n)$.*

For the sake of generality, we state our result for a wide range of functions $d = d(n)$. However, we remark that the most interesting case is when d is a sufficiently large constant. In fact, for dense graphs, when $d \ge n^{1/2}\log^2 n$, more precise results are already known (cf. [10, 25]).

Some of the earlier results [10, 15, 21] are phrased in terms of induced cycles (*holes*). This does not make the problem much harder (see Remark 4.2). We also note that our proof is self-contained, except for well-known facts from probability and graph theory.

We now briefly outline our strategy. Roughly speaking, the idea is to find a long induced path in two steps. First, we find many disjoint paths of some chosen length $L = L(d)$, such that the subgraph consisting of their union is induced. To achieve this, we generalize a recent result of Cooley, Draganić, Kang and Sudakov [7] who obtained large induced matchings. We will discuss this further in Section 1.2. Assuming now we can find such an induced linear forest F, the aim is to connect almost all of the small paths into one long induced path, using a few additional vertices.[1] As a "reservoir" for these connecting vertices, we find (actually, even before finding F) a large independent set I which is disjoint from F. To model the connecting step, we give each path in F a direction, and define an auxiliary digraph whose vertices are the paths, and two paths (P_1, P_2) form an edge if there exists a "connecting" vertex $a \in I$ that has some edge to one of the last ϵL vertices of P_1 and some edge to one of the first ϵL vertices of P_2, but no edge to the rest of F. Our goal is to find an almost spanning path in this auxiliary digraph. Observe that this will provide us with a path in $G(n,p)$ of length roughly $|F|$. The intuition is that the auxiliary digraph behaves quite randomly, which gives us hope that, even though it is very sparse, we can find an almost spanning path.

Crucially, we do not perform this connecting step in the whole random graph. This is because ensuring that the new connecting vertices are only connected to two vertices of F is too costly, making the auxiliary digraph so sparse that it is impossible to find an almost spanning path. Instead, we use a sprinkling argument, meaning that we view $G(n,p)$ as the union of two independent random graphs G_1 and G_2, where the edge probability of G_2 is much smaller than p. We then reveal the random choices in several stages. When finding F and I as above, we make

[1] Recall that a forest is called *linear* if its components are paths.

sure that there are no G_1-edges between F and I. Then, in the final connecting step, it remains to expose the G_2-edges between F and I, with the advantage that now the edge probability is much smaller, making it much more suitable for a desired "sparse" connection.

1.2 Induced forests with small components

As outlined above, in the first step of our argument, we seek an induced linear forest whose components are reasonably long paths. For this, we generalize a recent result of Cooley, Draganić, Kang and Sudakov [7]. They proved that **whp** $G(n,p)$ contains an induced matching with $\sim 2\log_q(np)$ vertices, which is asymptotically best possible. They also anticipated that using a similar approach one can probably obtain induced forests with larger, but bounded components. As a by-product, we confirm this. To state our result, we need the following definition. For a given graph T, a *T-matching* is a graph whose components are all isomorphic to T. Hence, a K_2-matching is simply a matching, and specifying the following statement for $T = K_2$ implies the main result of [7].

Theorem 1.2. *For any $\epsilon > 0$ and any fixed tree T, there exists $d_0 > 0$ such that **whp** the order of the largest induced T-matching in $G(n,p)$ is $(2 \pm \epsilon)\log_q(np)$, where $q = \frac{1}{1-p}$, whenever $\frac{d_0}{n} \leq p \leq 0.99$.*

We use the same approach as in [7], which goes back to the work of Frieze [14] (see also [4, 26]). The basic idea is as follows. Suppose we have a random variable X and want to show that **whp**, $X \geq b-t$, where b is some "target" value and t a small error. For many natural variables, we know that X is "concentrated", say $\mathbb{P}[|X - \mathbb{E}[X]| \geq t/2] < \rho$ for some small ρ. This is the case for instance when X is determined by many independent random choices, each of which has a small effect. However, it might be difficult to estimate $\mathbb{E}[X]$ well enough. But if we know in addition that $\mathbb{P}[X \geq b] \geq \rho$, then we can combine both estimates to $\mathbb{P}[X \geq b] > \mathbb{P}[X \geq \mathbb{E}[X] + t/2]$, which clearly implies that $b \leq \mathbb{E}[X] + t/2$. Now applying the other side of the concentration inequality, we infer $\mathbb{P}[X \leq b - t] \leq \mathbb{P}[X \leq \mathbb{E}[X] - t/2] < \rho$, as desired.

In our case, say X is the maximum order of an induced T-matching in $G(n,p)$. Since adding or deleting edges at any one vertex can create or destroy at most one component, we know that X is $|T|$-Lipschitz and hence concentrated (see Section 3). Using the above approach, it remains to complement this with a lower bound on the probability that $X \geq b$. Introduce a new random variable Y which is the *number* of induced T-matchings of order b (a multiple of $|T|$). Then we have $X \geq b$ if and only if $Y > 0$. The main technical work is to obtain a lower bound for the probability of the latter event using the second moment method. We note that by applying the second moment method to labelled copies (instead of unlabelled copies as in [7]) we obtain a shorter proof even in the case of matchings (see Section 2). More crucially, it turns out that one can even find induced forests where the component sizes can grow as a function of d, which we need in the proof of Theorem 1.1 (specifically, we need $L \gg \log d$). This is provided by the following auxiliary result. We note that the same holds for forests with arbitrary components of bounded degree, and one can also let the degree slowly grow with d, but we choose to keep the presentation simple.

Lemma 1.3. *For any $\epsilon > 0$, there exists $d_0 > 0$ such that **whp** $G(n,p)$ contains an induced linear forest of order at least $(2-\epsilon)p^{-1}\log(np)$ and component paths of order $d^{1/2}/\log^4 d$, whenever $d_0 \leq d = np \leq n^{1/2}\log^2 n$.*

1.3 Notation

We use standard graph theoretical notation. In particular, for a graph G and $U \subset V(G)$, we let $|G|$ denote the order of G, $e(G)$ the number of edges in G, $\Delta(G)$ the maximum degree and $G[U]$ the subgraph induced by U.

For functions $f(n), g(n)$, we write $f \sim g$ if $\lim_{n\to\infty} \frac{f(n)}{g(n)} = 1$. We also use the standard Landau symbols $o(\cdot), \Omega(\cdot), \Theta(\cdot), O(\cdot), \omega(\cdot)$, where subscripts disclose the variable that tends to infinity if this is not clear from the context. We use $\approx$ non-rigorously in informal discussions and ask the reader to interpret it correctly.

An event $\mathcal{E}_n$ holds *with high probability* (**whp**) if $\mathbb{P}[\mathcal{E}_n] = 1 - o_n(1)$. Also, $[n] = \{1, \dots, n\}$ and $(n)_k = n(n-1)\cdots(n-k+1)$. As customary, we tacitly treat large numbers like integers whenever this has no effect on the argument.

2 Second moment

In this section, we use the second moment method to derive a lower bound on the probability that $G(n, d/n)$ contains a given induced linear forest of size $\sim 2\frac{n}{d}\log d$. Here, it does not matter whether the components are small. More precisely, we prove that for fixed $\epsilon > 0$ and $d \geq d_0(\epsilon)$, *any* bounded degree forest F on $k \leq (2-\epsilon)\frac{n}{d}\log d$ vertices is an induced subgraph of $G(n, d/n)$ with probability at least $\exp(-O(\frac{n\log^2 d}{d^2}))$. Moreover, when $d = \omega(n^{1/2}\log n)$, the obtained probability bound tends to 1. In particular, in this regime, the lemma readily implies the existence of an induced path of the asymptotically optimal length $\sim 2\frac{n}{d}\log d$ **whp**.

Lemma 2.1. *For any $\epsilon > 0$, there exists d_0 such that the following holds for all $d_0 \leq d < n$, where $p = \frac{d}{n}$ and $q = \frac{1}{1-p}$. For any forest F on $k \leq (2-\epsilon)\log_q d$ vertices with maximum degree $\Delta \leq d^{\epsilon/6}$, the probability that $G(n,p)$ contains an induced copy of F is at least*

$$\exp\left(-10^4\Delta^2\frac{n\log^2 d}{d^2} - 2d^{-\epsilon/7}\right).$$

The proof of Lemma 2.1 is based on the second moment method and will be given below. We start off with some basic preparations which will also motivate the main counting tool.

Fix a forest F of order k. Let Y be the random variable which counts the number of *labelled* induced copies of F in $G(n,p)$. More formally, let $\mathcal{F}$ be the set of all injections $\sigma\colon V(F) \to [n]$, and for $\sigma \in \mathcal{F}$, let F_σ be the graph with vertex set $\{\sigma(x) : x \in V(F)\}$ and edge set $\{\sigma(x)\sigma(y) : xy \in E(F)\}$. Let A_σ be the event that F_σ is an induced subgraph of $G(n,p)$. Hence,

$$\mathbb{P}[A_\sigma] = p^{e(F)}(1-p)^{\binom{k}{2}-e(F)},$$

and setting $Y = \sum_{\sigma\in\mathcal{F}} \mathbb{1}(A_\sigma)$, we have

$$\mathbb{E}[Y] = (n)_k p^{e(F)}(1-p)^{\binom{k}{2}-e(F)}. \tag{2.1}$$

Ultimately, we want to obtain a lower bound for $\mathbb{P}[Y > 0]$. Fix some $\sigma_0 \in \mathcal{F}$. By symmetry, the second moment of Y can be written as

$$\mathbb{E}[Y^2] = \mathbb{E}[Y]\sum_{\sigma\in\mathcal{F}} \mathbb{P}[A_\sigma \mid A_{\sigma_0}].$$

Applying the Paley–Zygmund inequality, we thus have

$$\mathbb{P}[Y > 0] \geq \frac{\mathbb{E}[Y]^2}{\mathbb{E}[Y^2]} = \frac{\mathbb{E}[Y]}{\sum_{\sigma\in\mathcal{F}}\mathbb{P}[A_\sigma \mid A_{\sigma_0}]}. \tag{2.2}$$

The remaining difficulty is to control the terms $\mathbb{P}[A_\sigma \mid A_{\sigma_0}]$. We say that $\sigma \in \mathcal{F}$ is *compatible* (with σ_0) if $\mathbb{P}[A_\sigma \mid A_{\sigma_0}] > 0$. This means that, in the intersection $V(F_\sigma) \cap V(F_{\sigma_0})$, a pair uv which is an edge in F_σ cannot be a non-edge in F_{σ_0}, and vice versa, as otherwise F_σ and F_{σ_0} could not be induced subgraphs of $G(n,p)$ simultaneously. From now on, we can ignore all σ that are not compatible with σ_0.

If $\sigma \in \mathcal{F}$ is compatible with σ_0, we denote by $I_\sigma := F_\sigma \cap F_{\sigma_0}$ the graph on $S = V(F_\sigma) \cap V(F_{\sigma_0})$ with edge set $E(F_\sigma[S]) = E(F_{\sigma_0}[S])$. This "intersection graph" assumes a crucial role in the analysis. Suppose that I_σ has s vertices and c components. Since I_σ is a forest, we have $e(I_\sigma) = s - c$. These are the edges of F_σ that we already know to be there when conditioning on A_{σ_0}, and for F_σ, we need $e(F) - e(I_\sigma)$ "new" edges. Moreover, there are $\binom{k}{2} - \binom{s}{2} - e(F) + e(I_\sigma)$ additional non-edges. Therefore,

$$\mathbb{P}\left[A_\sigma \mid A_{\sigma_0}\right] = p^{e(F)-s+c}(1-p)^{\binom{k}{2}-\binom{s}{2}-e(F)+s-c}. \tag{2.3}$$

Note here that when the number of components c is large, then the exponent of p is large and hence we have a stronger upper bound on $\mathbb{P}\left[A_\sigma \mid A_{\sigma_0}\right]$. On the other hand, if c is small, then $\mathbb{P}\left[A_\sigma \mid A_{\sigma_0}\right]$ is larger, but this will be compensated by the fact that there are fewer such σ. In the following, we bound the number of compatible $\sigma \in \mathcal{F}$ for which I_σ has s vertices and c components. We remark that this kind of analysis was also carried out in [9] in the study of dense random graphs. We include the details for completeness, with an improved dependence on Δ. We make use of the following well-known counting result, see, e.g., [3, Lemma 2] for the proof.

Proposition 2.2. *For a graph H with $\Delta(H) \le \Delta$ and $v \in V(H)$, the number of (unlabelled) trees in H of order s which contain v is at most $(e\Delta)^{s-1}$.*

Proposition 2.3. *For all $0 \le c \le s$, the number of compatible $\sigma \in \mathcal{F}$ for which I_σ has s vertices and c components is at most*

$$\binom{k}{c} k^c (6\Delta^2)^s (n-k)_{k-s}.$$

Proof. Fix s and c. We can obviously assume that $s \ge c \ge 1$, as the only other non-void case is when $s = c = 0$, in which case the expression $(n-k)_k$ is exactly correct. The first claim is that the number of subgraphs of F_{σ_0} with s vertices and c components is at most $\binom{k}{c}(2e\Delta)^s$. To see this, we first choose root vertices $v_1, \dots, v_c$ for the components, for which there are at most $\binom{k}{c}$ choices. For $i \in [c]$, let T_i denote the component of I_σ which will contain v_i. Next, we fix the sizes of the components. Writing $s_i = |T_i|$, the number of possibilities is bounded by the number of positive integer solutions of $s_1 + \dots + s_c = s$, which is $\binom{s-1}{c-1} \le 2^s$ by a well-known formula. Now, having fixed the sizes, we can apply Proposition 2.2 for each $i \in [c]$, with F_{σ_0}, v_i playing the roles of H, v, to see that the number of choices for T_i is at most $(e\Delta)^{s_i - 1}$, which combined amounts to $(e\Delta)^{s-c}$. This implies the claim, and immediately yields an upper bound on the number of possibilities for the intersection graph I_σ.

Now, fix a choice of I_σ. Since I_σ is a forest with c components, its vertices can be ordered such that every vertex, except for the first c vertices in the ordering (which we can set to be the root vertices of the components), has exactly one neighbour preceding it. In order to count the number of possibilities for σ, we proceed as follows. First, choose the preimages under σ for the first c vertices, for which there are at most $(k)_c$ choices. Now, we choose the preimages of the remaining vertices of I_σ one-by-one in increasing order. In each step, there are at most Δ choices, since one neighbour of the current vertex has already chosen its preimage, and I_σ has to be an induced subgraph of F_σ. Hence, there are at most Δ^{s-c} choices for the preimages of the remaining vertices of I_σ. Finally, we have used s vertices of F as preimages for the vertices in I_σ. The remaining $k - s$ vertices of F must be mapped to $[n] \setminus V(F_{\sigma_0})$, so there are at most $(n-k)_{k-s}$ possibilities. □

With the preparations done, the proof of the lemma reduces to a chain of estimates.

Proof of Lemma 2.1. By (2.2), it suffices to show that

$$\frac{\sum_{\sigma \in \mathcal{F}} \mathbb{P}\left[A_\sigma \mid A_{\sigma_0}\right]}{\mathbb{E}\left[Y\right]} \le \exp\left(10^4 \Delta^2 \frac{n \log^2 d}{d^2} + 2d^{-\epsilon/7}\right).$$

We split the sum over compatible $\sigma \in \mathcal{F}$ according to the number of vertices and components of I_σ. Applying (2.1), (2.3) and Proposition 2.3, we obtain

$$\begin{aligned}\frac{\sum_{\sigma\in\mathcal{F}} \mathbb{P}\left[A_\sigma \mid A_{\sigma_0}\right]}{\mathbb{E}\left[Y\right]} &\leq \sum_{s=0}^{k}\sum_{c=0}^{s} \frac{\binom{k}{c} k^c (6\Delta^2)^s (n-k)_{k-s} p^{e(F)-s+c}(1-p)^{\binom{k}{2}-\binom{s}{2}-e(F)}}{(n)_k p^{e(F)}(1-p)^{\binom{k}{2}-e(F)}} \\ &= \sum_{s=0}^{k} \frac{(n-k)_{k-s}}{(n)_k} p^{-s} q^{\binom{s}{2}} (6\Delta^2)^s \sum_{c=0}^{s} \binom{k}{c} (kp)^c\end{aligned}$$

In order to bound this expression, we make some observations. Note that we always have

$$k \leq 2\log_q(np) \leq 2\frac{n}{d}\log d \tag{2.4}$$

since $\log q \geq p$. Hence, $kp \leq 2\log d$. Moreover, we have $\frac{(n-k)_{k-s}}{(n)_k} \leq \frac{1}{(n)_s} \leq (4/n)^s$ since $s \leq k \leq 3n/4$, say, using (2.4). Observe further that $\binom{k}{c} \leq \binom{2k}{c} \leq \binom{2k}{s} \leq \frac{(2k)^s}{s!}$. Finally, c takes only $s+1 \leq 2^s$ values. Hence, the last sum displayed above is at most

$$\sum_{s=0}^{k} (4/n)^s p^{-s} q^{s^2/2} (6\Delta^2)^s \frac{(16k\log d)^s}{s!} = \sum_{s=0}^{k} \frac{\left(\frac{384\Delta^2 k\log d}{d} q^{s/2}\right)^s}{s!}.$$

We split the final sum into two terms. First, consider the range $s \leq k/\log d$. Then $q^{s/2} \leq q^{1/\log q} = e$. Hence, recalling that $e^x = \sum_{s\geq 0} \frac{x^s}{s!}$, we obtain the bound

$$\sum_{s=0}^{\lfloor k/\log d\rfloor} \frac{\left(\frac{384\Delta^2 k\log d}{d} q^{s/2}\right)^s}{s!} \leq \exp\left(\frac{384e\Delta^2 k\log d}{d}\right) \leq \exp\left(\frac{10^4\Delta^2 n\log^2 d}{d^2}\right).$$

Finally, for $s \geq k/\log d$, we use $s! \geq (s/e)^s$ to bound each summand as

$$\frac{\left(\frac{384\Delta^2 k\log d}{d} q^{s/2}\right)^s}{s!} \leq \left(\frac{384e\Delta^2 k\log d}{ds} q^{s/2}\right)^s \leq \left(\frac{384e\Delta^2\log^2 d}{d} q^{s/2}\right)^s.$$

Crucially, since $s \leq k \leq (2-\epsilon)\log_q d$, we have $q^{s/2} \leq q^{(1-\epsilon/2)\log_q d} = d^{1-\epsilon/2}$. Now, for sufficiently large $d \geq d_0$ the exponent base above is bounded by $\frac{384e\Delta^2\log^2 d}{d^{\epsilon/2}} \leq d^{-\epsilon/7} < 1$. Therefore the geometric series tells us that

$$\sum_{s=\lceil k/\log d\rceil}^{k} \frac{\left(\frac{384\Delta^2 k\log d}{d} q^{s/2}\right)^s}{s!} \leq \frac{1}{1-d^{-\epsilon/7}} - 1 \leq 2d^{-\epsilon/7}.$$

Altogether, we conclude that

$$\frac{\sum_{\sigma\in\mathcal{F}} \mathbb{P}\left[A_\sigma \mid A_{\sigma_0}\right]}{\mathbb{E}\left[Y\right]} \leq \exp\left(\frac{10^4\Delta^2 n\log^2 d}{d^2}\right) + 2d^{-\epsilon/7} \leq \exp\left(\frac{10^4\Delta^2 n\log^2 d}{d^2} + 2d^{-\epsilon/7}\right),$$

completing the proof. □

3 Concentration

In this section, we deduce Theorem 1.2 and Lemma 1.3 from Lemma 2.1. We will use Talagrand's inequality. To state it, we need the following definitions. Given a product probability space $\Omega = \prod_{i=1}^{n} \Omega_i$ (endowed with the product measure) and a random variable $X\colon \Omega \to \mathbb{R}$, we say that X is

- *L-Lipschitz* (for some $L > 0$) if for any $\omega, \omega' \in \Omega$ which differ only in one coordinate, we have $|X(\omega) - X(\omega')| \le L$;
- *f-certifiable* (for a function $f\colon \mathbb{N} \to \mathbb{N}$) if for every s and $\omega \in \Omega$ such that $X(\omega) \ge s$, there exists a set $I \subset [n]$ of size $\le f(s)$ such that $X(\omega') \ge s$ for every ω' that agrees with ω on the coordinates indexed by I.

Theorem 3.1 (Talagrand's inequality, see [1])**.** *Suppose that X is L-Lipschitz and f-certifiable. Then, for all $b, t \ge 0$,*

$$\mathbb{P}\left[X \le b - tL\sqrt{f(b)}\right] \mathbb{P}\left[X \ge b\right] \le \exp\left(-t^2/4\right).$$

Our probability space is of course $G(n,p)$. Although this comes naturally as a product of $\binom{n}{2}$ elementary probability spaces Ω_{ij}, one for each potential edge ij, it can be more effective, depending on the problem, to consider a description that is vertex-oriented, where the edges incident to a vertex are combined into one probability space ("vertex exposure"). Concretely, for $i \in [n-1]$, let $\Omega_i = \prod_{j>i} \Omega_{ij}$ represent all edges from vertex i to vertices $j > i$. Then $G(n,p) = \prod_{i=1}^{n-1} \Omega_i$. Note here that the vertices are ordered to describe the product space in a way that every edge appears exactly once. Apart from that, this ordering plays no role.

Proof of Theorem 1.2. Fix $\epsilon > 0$, a tree T, and assume d_0 is sufficiently large. Let $L = |V(T)|$ and $d = np$. We first show the upper bound, by arguing that **whp** every set of size at least $t = (2+\epsilon)\log_q(np)$ spans at least t edges, which prevents us from finding any induced forest of order t. Indeed, the probability that a fixed t-set spans at most t edges is at most

$$(t+1)\binom{\binom{t}{2}}{t} p^t (1-p)^{\binom{t}{2}-t},$$

as the number of edges in such a set follows a binomial distribution $\mathrm{Bin}\left(\binom{t}{2}, p\right)$, with the mean being larger than t (where we used that $p < 0.99$). Hence we have that the expected number of t-sets which span at most t edges is at most

$$\begin{aligned}(t+1)\binom{n}{t}\binom{\binom{t}{2}}{t} p^t (1-p)^{\binom{t}{2}-t} &\le (t+1)\left(\frac{en}{t}\right)^t \left(\frac{e\binom{t}{2}}{t}\right)^t \left(\frac{d}{n}\right)^t d^{-(1+\epsilon/3)t} \\ &< (t+1)(e^2/2)^t d^{-\epsilon t/3} = o(1),\end{aligned}$$

where we used standard estimates and the fact that $(1-p)^t = q^{-t} = d^{-(2+\epsilon)}$.

We now turn to the lower bound. Let X be the maximum order of an induced T-matching in $G(n,p)$. Our goal is to show that $X \ge (2-\epsilon)\log_q d$ **whp**. Set $b = (2-\epsilon/2)\log_q d$.

First, by Lemma 2.1, we have

$$\mathbb{P}\left[X \ge b\right] \ge \exp\left(-10^4 L^2 \frac{n \log^2 d}{d^2} - 2d^{-\Omega(\epsilon)}\right).$$

This means that in the case $d \ge n^{1/2}\log^2 n$, we are already done. Assume now that $d \le n^{1/2}\log^2 n$. Then the above bound simplifies to

$$\mathbb{P}\left[X \ge b\right] \ge \exp\left(-\frac{n \log^3 d}{d^2}\right). \tag{3.1}$$

Recall also that in the regime $d = o(n)$ we have $\log_q d \sim \frac{n}{d} \log d$.

It is easy to check that X is L-Lipschitz and f-certifiable, where $f(s) = s + L$. Indeed, adding or deleting edges arbitrarily at one vertex can change the value of X by at most L, hence X is L-Lipschitz. Moreover, if $X \geq s$, this means there is a set $I \subset [n]$ of size $s \leq |I| < s + L$ which induces a T-matching. If we leave the coordinates indexed by I unchanged, this means in particular that I still induces a T-matching, hence we still have $X \geq s$.

Hence, Talagrand's inequality applied with $t = \frac{\sqrt{n} \log^3 d}{d}$ yields

$$\mathbb{P}\left[X \leq b - tL\sqrt{b+L}\right] \mathbb{P}\left[X \geq b\right] \leq \exp\left(-\frac{n \log^6 d}{4d^2}\right).$$

Together with (3.1) and since

$$tL\sqrt{b+L} \leq \frac{\sqrt{n} \log^3 d}{d} L \sqrt{2\frac{n}{d} \log d} \leq \frac{n}{d}, \tag{3.2}$$

we infer that the probability of $X \leq b - \frac{n}{d}$ is at most $\exp\left(-\frac{n \log^6 d}{5d^2}\right) = o_n(1)$. This completes the proof since $b - \frac{n}{d} \geq (2 - \epsilon) \log_q d$. □

In the above proof, we had some room to spare in (3.2). We will now exploit this to allow the component sizes to grow with d. The proof is almost verbatim the same, so we only point out the differences.

Proof of Lemma 1.3. Note that we are only interested in the case $d \leq n^{1/2} \log^2 n$ and when T is a path of order L. Since $\Delta(T) \leq 2$, Lemma 2.1 still provides the lower bound in (3.1). All we have to ensure now is that (3.2) still holds, and this is easily seen to be the case as long as $L \leq d^{1/2} / \log^4 d$. □

4 Connecting

In this section, we use Lemma 1.3 to prove Theorem 1.1 as outlined in Section 1.1. Recall that we intend to define an auxiliary digraph on the components of a linear forest, where an edge corresponds to a suitable connection between two paths. Our goal is to find an almost spanning path in this random digraph. The tool which enables us to achieve this, Lemma 4.1 below, is based on the well-known graph exploration process *depth-first-search* (DFS). The usefulness of DFS to find long paths in random graphs was demonstrated by Krivelevich and Sudakov [19] in a paper where they give rather short and simple proofs of classical results in random graph theory. In our proof, we use the following straightforward consequence of DFS.

Lemma 4.1 ([2, Lemma 4.4])**.** *Let D be a digraph on N vertices and suppose that for any two disjoint sets $S, T \subset V(D)$ of size k, there exists an edge directed from S to T. Then D contains a path of length $N - 2k + 1$.*

Proof of Theorem 1.1. Fix $\epsilon > 0$ and assume that $d \geq d_0$ is sufficiently large. We will assume that $d = o(n^{1/2} \log^2 n)$. For the case $d = \omega(n^{1/2} \log n)$, Lemma 2.1 implies that **whp** there exists an induced path of length $(2 - \epsilon)\frac{n}{d} \log d$.

We expose the random graph in several stages, and will after each step fix an outcome that holds with high probability. First, we consider $G \sim G(n,p)$ as the union of two independent random graphs G_1 and G_2, where $G_2 \sim G(n, p_2)$ with $p_2 = \frac{d}{n \log d}$, and $G_1 \sim G(n, p_1)$ with p_1 such that $1 - p = (1 - p_1)(1 - p_2)$. (Hence, $G_1 \cup G_2$ has the same distribution as G.) Note that clearly $p_1 \leq p = d/n$.

Fix a subset $V_0 \subset [n]$ of size $\frac{n}{2d}$. Now, in the first exposure round, we reveal all random edges from $G = G_1 \cup G_2$ inside V_0. The expected number of edges is at most $\frac{n}{8d}$, and using a

standard Chernoff bound, it is easily seen that **whp**, the number of edges inside V_0 is at most $\frac{n}{6d}$, say. From now on, fix such an outcome. By deleting a vertex from each edge, we can find an independent set $I \subset V_0$ in $G[V_0]$ of size $\frac{n}{3d}$.

In the second round, we expose all edges with one vertex in I and the other in $[n] \setminus V_0$, but only those which belong to G_1. The expected number of such edges is at most $n/3$. Again using a Chernoff bound, the number of G_1-edges leaving I is at most $2n/5$ **whp**. From now on, fix such an outcome. Let $V_1 = [n] \setminus (I \cup N_{G_1}(I))$. Since $|N_{G_1}(I)| + |I| \leq n/2$, we have $|V_1| \geq n/2$.

In the third step, we reveal the random edges of $G = G_1 \cup G_2$ inside V_1. Now we apply Lemma 1.3 to $G[V_1]$; we get that **whp** there exists an induced forest F of order

$$(2-\epsilon)\, p^{-1} \log(np/2) \geq (2-2\epsilon)\, \frac{n}{d} \log d$$

whose components are paths of order $L = \Theta\left(\frac{\sqrt{d}}{\log^4 d}\right)$. Again, we fix such F. Note that F is induced in $G = G_1 \cup G_2$ and that I is independent in $G = G_1 \cup G_2$. Moreover, by definition of V_1, we know that there are no edges in G_1 between F and I.

In the fourth and final step, we reveal all the remaining random edges, which in particular includes the G_2-edges between F and I. Our goal is to use some vertices from I to connect most of the paths of F into one long induced path. To achieve this, we define the following auxiliary digraph D. Give each of the component paths P of F an arbitrary direction, and let P^- denote the initial ϵL vertices and P^+ the last ϵL vertices on P. Now, the vertex set of D is simply the set of components of F. For two paths P_1, P_2, we include (P_1, P_2) as an edge in D if there exists a vertex $a \in I$ such that a has exactly one edge (of G_2) to both P_1^+ and P_2^-, but no other edge (of G_2) to any vertex of F. Note that D is a random digraph. Our claim is that, with high probability, it contains an almost spanning path. Let $N = |V(D)| = \Theta\left(\frac{n}{dL} \log d\right)$. Consider any two disjoint sets $S, T \subset V(D)$ of size ϵN. For each $a \in I$ and all $P_1 \in S, P_2 \in T$, the probability that a is a suitable connection from P_1 to P_2 is the sum of probabilities over the $(\epsilon L)^2$ possible pairs to form a suitable connection (since those events are disjoint) and equals to

$$(\epsilon L p_2)^2 (1-p_2)^{|F|-2} \geq (\epsilon L p_2)^2 e^{-2p_2|F|} = \Theta\left(\epsilon^2 L^2 p_2^2\right)$$

since $p_2|F| = \Theta(1)$ by our choice of p_2. Moreover, for distinct pairs (P_1, P_2) and fixed a, these events are disjoint. Hence, the probability of a giving some good connection from S to T is $\Theta(\epsilon^4 N^2 L^2 p_2^2) = \Theta(\epsilon^4)$. Finally, for distinct a, these events are determined by disjoint sets of edges, and hence independent, so the probability that there is no edge from S to T in D is at most

$$\left(1 - \Theta(\epsilon^4)\right)^{|I|} \leq \exp\left(-\Omega(\epsilon^4 n/d)\right).$$

The total number of choices for S and T is at most $4^N = \exp(\Theta(\frac{n}{dL} \log d))$. Thus, since we have that $L \gg \log d/\epsilon^4$, a union bound yields that **whp**, we can apply Lemma 4.1 to get a path $P_1 P_2 \dots P_t$ in D of length $(1-2\epsilon)N$. This translates to an induced path of G as follows: for each $i \in [t-1]$, since $P_i P_{i+1} \in E(D)$, there exists a vertex $a_i \in I$ which has exactly one edge (of G_2) to both P_i^+ and P_{i+1}^-, but no other edge (of G_2) to any vertex of F. Clearly, the a_i's are distinct, hence we obtain a path in G in the obvious way (start with P_1, then from the appropriate vertex in P_1^+, go via a_1 to P_2^-, follow P_2, and then go from P_2^+ via a_2 to P_3^-, etc.). As remarked earlier, there are no G_1-edges between F and I, hence by the definition of $E(D)$, the path will be induced. Finally, from each P_i, we only lose at most $2\epsilon L$ vertices, hence the length of the path will be at least

$$(L - 2\epsilon L)(1-2\epsilon)N = (1-2\epsilon)^2 |F| \geq (2 - O(\epsilon))\, \frac{n}{d} \log d,$$

completing the proof. □

Remark 4.2. The condition in Lemma 4.1 actually also implies the existence of a cycle of length at least $N - 4k + 4$ (since there is an edge from the last k vertices on the obtained path to the first k vertices). Using this, in the above proof, we can connect the paths of F into an induced cycle of length $(2 - O(\epsilon))\frac{n}{d}\log d$.

5 Concluding remarks

- Our proof is not constructive, since the first part of the argument uses the second moment method. The previously best bound $\sim \frac{n}{d}\log d$ due to Łuczak [21] and Suen [27] was obtained via certain natural algorithms. It seems that this could be a barrier for such approaches. A (rather unsophisticated) heuristic giving evidence is that when we have grown an induced tree of this size, and assume the edges outside are still random, then the expected number of vertices which could be attached to a given vertex of the tree is less than one for $|V(T)| \geq (1+\epsilon)\frac{n\log d}{d}$. Moreover, such an "algorithmic gap" has been discovered for many other natural problems. In particular, despite decades of research, no polynomial-time algorithm is known which finds an independent set of size $(1+\epsilon)\frac{n}{d}\log d$ for any fixed $\epsilon > 0$, and evidence has emerged that in fact such an algorithm might not exist (see e.g. [6, 17, 24]).

- In [7] it is conjectured that one should not only be able to find an induced path of size $\sim 2\frac{n}{d}\log d$, but an induced copy of any given bounded degree tree of this order. For dense graphs, when $d = \omega(n^{1/2}\log n)$, this follows from the second moment method (see [9]). In fact, Lemma 2.1 shows that the maximum degree can even be a small polynomial. On the other hand, the sparse case seems to be more difficult, mainly because the vanilla second moment method does not work. However, Dani and Moore [8] demonstrated that one can actually make the second moment method work, at least for independent sets, by considering a *weighted* version. This even gives a more precise result than the classical one due to Frieze [14]. It would be interesting to find out whether this method can be adapted to induced trees.

Acknowledgement

The first two authors are grateful to Benny Sudakov for very useful discussions.

References

[1] N. Alon and J. H. Spencer, **The probabilistic method**, 4th ed., Wiley-Intersci. Ser. Discrete Math. Optim., Wiley, 2016.

[2] I. Ben-Eliezer, M. Krivelevich, and B. Sudakov, *The size Ramsey number of a directed path*, J. Combin. Theory Ser. B **102** (2012), 743–755.

[3] A. Beveridge, A. Frieze, and C. McDiarmid, *Random minimum length spanning trees in regular graphs*, Combinatorica **18** (1998), 311–333.

[4] B. Bollobás, *The chromatic number of random graphs*, Combinatorica **8** (1988), 49–55.

[5] B. Bollobás and P. Erdős, *Cliques in random graphs*, Math. Proc. Cambridge Philos. Soc. **80** (1976), 419–427.

[6] A. Coja-Oghlan and C. Efthymiou, *On independent sets in random graphs*, Random Structures Algorithms **47** (2015), 436–486.

[7] O. Cooley, N. Draganić, M. Kang, and B. Sudakov, *Large induced matchings in random graphs*, SIAM J. Discrete Math. **35** (2021), 267–280.

[8] V. Dani and C. Moore, *Independent sets in random graphs from the weighted second moment method*, Approximation, randomization, and combinatorial optimization, Lecture Notes in Comput. Sci. 6845, Springer, 2011, pp. 472–482.

[9] N. Draganić, *Large induced trees in dense random graphs*, arXiv:2004.02800 (2020).

[10] K. Dutta and C. R. Subramanian, *On induced paths, holes and trees in random graphs*, 2018 Proceedings of the Fifteenth Workshop on Analytic Algorithmics and Combinatorics (ANALCO), SIAM, Philadelphia, PA, 2018, pp. 168–177.

[11] P. Erdős and Z. Palka, *Trees in random graphs*, Discrete Math. **46** (1983), 145–150.

[12] W. Fernandez de la Vega, *Induced trees in sparse random graphs*, Graphs Combin. **2** (1986), 227–231.

[13] ———, *The largest induced tree in a sparse random graph*, Random Structures Algorithms **9** (1996), 93–97.

[14] A. M. Frieze, *On the independence number of random graphs*, Discrete Math. **81** (1990), 171–175.

[15] A. M. Frieze and B. Jackson, *Large holes in sparse random graphs*, Combinatorica **7** (1987), 265–274.

[16] ———, *Large induced trees in sparse random graphs*, J. Combin. Theory Ser. B **42** (1987), 181–195.

[17] D. Gamarnik and M. Sudan, *Limits of local algorithms over sparse random graphs*, Ann. Probab. **45** (2017), 2353–2376.

[18] G. R. Grimmett and C. J. H. McDiarmid, *On colouring random graphs*, Math. Proc. Cambridge Philos. Soc. **77** (1975), 313–324.

[19] M. Krivelevich and B. Sudakov, *The phase transition in random graphs: a simple proof*, Random Structures Algorithms **43** (2013), 131–138.

[20] L. Kučera and V. Rödl, *Large trees in random graphs*, Comment. Math. Univ. Carolin. **28** (1987), 7–14.

[21] T. Łuczak, *The size of the largest hole in a random graph*, Discrete Math. **112** (1993), 151–163.

[22] T. Łuczak and Z. Palka, *Maximal induced trees in sparse random graphs*, Discrete Math. **72** (1988), 257–265.

[23] D. W. Matula, *The largest clique size in a random graph*, Department of Computer Science, Southern Methodist University, Tech. Report CS 7608, 1976.

[24] M. Rahman and B. Virág, *Local algorithms for independent sets are half-optimal*, Ann. Probab. **45** (2017), 1543–1577.

[25] A. Ruciński, *Induced subgraphs in a random graph*, Ann. Discrete Math. **33** (1987), 275–296.

[26] E. Shamir and J. Spencer, *Sharp concentration of the chromatic number on random graphs* $G_{n,p}$, Combinatorica **7** (1987), 121–129.

[27] W. C. S. Suen, *On large induced trees and long induced paths in sparse random graphs*, J. Combin. Theory Ser. B **56** (1992), 250–262.