Finding paths in sparse random graphs requires many queries

Asaf Ferber [*] Michael Krivelevich[†] Benny Sudakov [‡] Pedro Vieira [§]

Abstract

We discuss a new algorithmic type of problem in random graphs studying the minimum number of queries one has to ask about adjacency between pairs of vertices of a random graph $G \sim \mathcal{G}(n,p)$ in order to find a subgraph which possesses some target property with high probability. In this paper we focus on finding long paths in $G \sim \mathcal{G}(n,p)$ when $p = \frac{1+\varepsilon}{n}$ for some fixed constant $\varepsilon > 0$. This random graph is known to have typically linearly long paths.

To have ℓ edges with high probability in $G \sim \mathcal{G}(n,p)$ one clearly needs to query at least $\Omega\left(\frac{\ell}{p}\right)$ pairs of vertices. Can we find a path of length ℓ economically, i.e., by querying roughly that many pairs? We argue that this is not possible and one needs to query significantly more pairs. We prove that any randomised algorithm which finds a path of length $\ell = \Omega\left(\frac{\log\left(\frac{1}{\varepsilon}\right)}{\varepsilon}\right)$ with at least constant probability in $G \sim \mathcal{G}(n,p)$ with $p = \frac{1+\varepsilon}{n}$ must query at least $\Omega\left(\frac{\ell}{p\varepsilon\log\left(\frac{1}{\varepsilon}\right)}\right)$ pairs of vertices. This is tight up to the $\log\left(\frac{1}{\varepsilon}\right)$ factor.

1 Introduction

Let $\mathcal{P}$ be a monotone increasing graph property (that is, a property of graphs that cannot be violated by adding edges). Suppose that the edge probability $p = p(n)$ is chosen so that a random graph G drawn from the probability space $\mathcal{G}(n,p)$ has $\mathcal{P}$ with high probability (whp). How many queries of the type "is $(i,j) \in E(G)$?" are needed for an adaptive algorithm interacting with the probability space $\mathcal{G}(n,p)$ in order to reveal whp a subgraph $G' \subseteq G$ possessing $\mathcal{P}$?

This fairly natural algorithmic setting (see the excellent survey of Frieze and McDiarmid [10] for an extensive coverage of a variety of problems and results in Algorithmic Theory of Random Graphs) has been considered implicitly in several papers on random graphs (e.g. [14], [5]), but apparently has been stated explicitly only in the companion paper [9] of the authors. Notice that in this framework the issue of concern is not the amount of computation required for the algorithm to find a target structure, but rather the amount of its interaction with the underlying probability space.

In the discussion below, we assume some basic familiarity with results about the probability space $\mathcal{G}(n,p)$; the reader is advised to consult monographs [11] and [6] for background on the subject.

In general, given a monotone property $\mathcal{P}$, what can we expect? If all n-vertex graphs belonging to $\mathcal{P}$ have at least m edges, then the algorithm should get at least m positive answers to hit the

[*]Department of Mathematics, Yale University, and Department of Mathematics, MIT, USA. Emails: asaf.ferber@yale.edu, ferbera@mit.edu.

[†]School of Mathematical Sciences, Raymond and Beverly Sackler Faculty of Exact Sciences, Tel Aviv University, Tel Aviv, 6997801, Israel. Email: krivelev@post.tau.ac.il. Research supported in part by USA-Israel BSF Grant 2010115 and by grant 912/12 from the Israel Science Foundation.

[‡]Department of Mathematics, ETH, 8092 Zurich, Switzerland. Email: benjamin.sudakov@math.ethz.ch. Research supported in part by SNSF grant 200021-149111.

[§]Department of Mathematics, ETH, 8092 Zurich, Switzerland. Email: pedro.vieira@math.ethz.ch.

target property with the required absolute certainty. This means that the obvious lower bound in this case is at least $(1+o(1))m/p$ queries. Perhaps one of the simplest graph properties to consider in this respect is connectedness: for any connected graph G on n vertices a spanning tree can be found after $n-1$ queries with positive answers – the algorithm starts with an arbitrary vertex $v \in V(G)$, and each time queries the pairs leaving the current tree until the first edge is found, the tree is then updated by appending this edge. Thus for the regime where $\mathcal{G}(n,p)$ is whp connected (which is when $p(n) \geq \frac{\ln n + \omega(n)}{n}$ with $\lim_{n\to\infty} \omega(n) = 1$), we get an algorithm whp discovering a spanning tree after querying $(1+o(1))n/p$ pairs of vertices.

A much more challenging problem is that of Hamiltonicity, i.e., of finding a Hamilton cycle. In this case the trivial lower bound translates to n positive answers. In [9] we show that this lower bound is tight by providing an adaptive algorithm interacting with the probability space $\mathcal{G}(n,p)$, which whp finds a Hamilton cycle in $G \sim \mathcal{G}(n,p)$ after obtaining only $(1+o(1))n$ positive answers (provided p is above the sharp threshold for Hamiltonicity in $\mathcal{G}(n,p)$).

Yet another positive example is that of uncovering a giant component in the supercritical regime $p = \frac{1+\varepsilon}{n}$. Though this was not the main concern in [14], the second and the third author presented there a very natural adaptive algorithm (essentially performing the Depth First Search (DFS) on a random input $G \sim \mathcal{G}(n,p)$), typically discovering a connected component of size at least $\epsilon n/2$ after querying $\epsilon n^2/2$ vertex pairs.

Upon reviewing these results, the reader may arrive at a conclusion that the above stated trivial lower bound for this type of problems is nearly tight for almost every natural graph property. However, this happens **not** to be the case, and the main qualitative goal of the present paper is to provide such a negative example, including its analysis. Here we focus on the property of containing a path of length ℓ in the supercritical regime in $G \sim \mathcal{G}(n,p)$, that is, when $p = \frac{1+\varepsilon}{n}$ for some fixed constant $\varepsilon > 0$. For this regime, $G \sim \mathcal{G}(n,p)$ is known to contain whp a path of length linear in n, due to the classical result of Ajtai, Komlós and Szemerédi [3] (see [14] for a recent simple proof of this fact.) Note that in order to have ℓ edges with high probability in $G \sim \mathcal{G}(n,p)$ one needs to query at least $\Omega\left(\frac{\ell}{p}\right)$ pairs of vertices. Can we find a path of length ℓ by asking roughly that many queries, as in the case of Hamiltonicity mentioned above? We show that in this case one actually needs to query significantly more pairs of vertices:

Theorem 1. *There exists an absolute constant $C > 0$ such that the following holds. For every constant $q \in (0,1)$ there exist $n_0, \varepsilon_0 > 0$ such that for every fixed $\varepsilon \in (0, \varepsilon_0)$ and any $n \geq n_0$ there is no adaptive algorithm which reveals a path of length $\ell \geq \frac{3C}{\varepsilon} \ln\left(\frac{1}{\varepsilon}\right)$ with probability at least q in $G \sim \mathcal{G}(n,p)$, where $p = \frac{1+\varepsilon}{n}$, by querying at most $\frac{q\ell}{8640 C p \varepsilon \ln\left(\frac{1}{\varepsilon}\right)}$ pairs of vertices.*

Notice that [14] presents a simple adaptive DFS algorithm, finding a path of length $\frac{1}{5}\varepsilon^2 n$ with probability at least $1 - \exp(\Omega(\varepsilon n))$ in $G \sim \mathcal{G}(n,p)$ after querying only $O\left(\varepsilon n^2\right)$ pairs of vertices. In fact, if one uses the same algorithm to find a path of length $\ell \leq \frac{1}{5}\varepsilon^2 n$ in $G \sim \mathcal{G}(n,p)$ then the same argument shows that such a path is found with probability at least $1 - \exp\left(\Omega\left(\frac{\ell}{\varepsilon}\right)\right)$ after querying at most $O\left(\frac{\ell}{p\varepsilon}\right)$ pairs of vertices. This shows that up to the $\Theta\left(\log\left(\frac{1}{\varepsilon}\right)\right)$ factor, Theorem 1 is tight.

The key ingredient of the proof of Theorem 1 is the following result of independent interest.

Theorem 2. *There exist constants $C, \varepsilon_0 > 0$ such that for every fixed $\varepsilon \in (0, \varepsilon_0)$ and $p = \frac{1+\varepsilon}{n}$ we have whp that a graph $G \sim \mathcal{G}(n,p)$ does not contain a set of vertex disjoint paths of lengths at least $\frac{C}{\varepsilon} \ln\left(\frac{1}{\varepsilon}\right)$ whose union covers at least $13\varepsilon^2 n$ vertices.*

The rest of this paper is organised as follows. In Section 2 we provide auxiliary lemmas needed for the proofs of Theorem 1 and 2. In Section 3 we prove Theorem 1 assuming Theorem 2. In Section 4 we prove Theorem 2. Finally, in Section 5 we discuss some concluding remarks.

Notation. Our notation is fairly standard. Given a natural number n we use $[n]$ to denote the set $\{1, 2, \ldots, n\}$. Moreover, given a set V we use S_V to denote the permutation group of V and $\binom{V}{2}$ to denote the set of all (unordered) pairs of elements in V.

Given a subset S of the vertex set of a graph G, $G[S]$ denotes the subgraph of G induced by the vertices in S, i.e. the graph with vertex set S whose edges are the ones of G between vertices in S.

A subgraph P of the graph G is called a *path* if $V(P) = \{v_1, \ldots, v_\ell\}$ and the edges of P are v_1v_2, v_2v_3, $\ldots$, $v_{\ell-1}v_\ell$. We shall oftentimes refer to P simply by $v_1v_2 \ldots v_\ell$. We say that such a path P has *length* $\ell - 1$ (number of edges) and *size* ℓ (number of vertices).

If G is a graph then the 2-*core* of G is the maximal induced subgraph of G of minimum degree at least 2. If no such subgraph exists then the 2-core of G is the empty graph.

Given an ordered set V and a real number $p \in [0, 1]$, the binomial random graph model $\mathcal{G}(V, p)$ is a probability space whose ground set consists of all labeled graphs on the vertex set V. We can describe the probability distribution of $G \sim \mathcal{G}(V, p)$ by saying that each pair of elements of V forms an edge in G independently with probability p. If $V = [n]$ then we will abuse notation slightly and use $\mathcal{G}(n, p)$ to refer to $\mathcal{G}([n], p)$. Given a property $\mathcal{P}$ (that is, a collection of graphs) and a function $p = p(n) \in [0, 1]$, we say that $G \sim \mathcal{G}(n, p)$ has $\mathcal{P}$ *with high probability* (or whp for brevity) if the probability that $G \in \mathcal{P}$ tends to 1 as n tends to infinity.

2 Auxiliary Lemmas

2.1 Concentration inequalities

We need to employ standard bounds on large deviations of random variables. The following well-known lemma due to Chernoff (commonly known as the "Chernoff bound") provides a bound on the upper tail of the Binomial distribution (see e.g. [4], [11]).

Lemma 1. *Let* $X \sim \mathrm{Bin}(n, p)$ *and let* $\mu = \mathbb{E}[X]$. *Then* $\Pr[X \geq (1 + a)\mu] < e^{-\frac{a^2\mu}{3}}$ *for any* $0 < a < \frac{3}{2}$.

The next lemma is a concentration inequality for the edge exposure martingale in $\mathcal{G}(n, p)$ which follows easily from Theorem 7.4.3 of [4].

Lemma 2. *Suppose* X *is a random variable in the probability space* $\mathcal{G}(n, p)$ *such that* $|X(G) - X(H)| \leq C$ *if* G *and* H *differ in one edge. Then*

$$\Pr\left[|X - \mathbb{E}[X]| > C\alpha\sqrt{n^2p}\right] \leq 2e^{-\frac{\alpha^2}{4}}$$

for any positive $\alpha < 2\sqrt{n^2p}$.

2.2 Galton-Watson trees and paths

A Galton-Watson tree is a random rooted tree, constructed recursively from the root where each node has a random number of children and these random numbers are independent copies of some random variable ξ taking values in $\{0, 1, 2, \ldots\}$. We let $\mathcal{T}$ denote a (random) Galton-Watson tree. We view the children of each node as arriving in some random order, so that $\mathcal{T}$ is an ordered, or plane tree.

We consider the *conditioned Galton-Watson tree* $\mathcal{T}_t$, which is the random tree $\mathcal{T}$ conditioned on having exactly t vertices. In symbols, $\mathcal{T}_t := (\mathcal{T} \mid |\mathcal{T}| = t)$, where, for any tree T, $|T|$ denotes its number of vertices.

For a rooted tree T, the *depth* $h(v)$ of a vertex v is its distance to the root (in particular the root has depth 0). We define as usual the *height* of the rooted tree T by $H(T) := \max\{h(v) : v \in T\}$. The following lemma which appears in [1] provides essentially optimal uniform sub-Gaussian upper tail bounds on $\frac{H(\mathcal{T}_t)}{\sqrt{t}}$ for every offspring distribution ξ with finite variance.

Lemma 3. *Suppose that* $\mathbb{E}[\xi] = 1$ *and* $0 < Var[\xi] < \infty$. *Then there exist constants* $C, c > 0$ *(which may depend on* ξ*) such that*

$$\Pr[H(\mathcal{T}_t) \geq h] \leq C \exp\left(-\frac{ch^2}{t}\right)$$

for all $h \geq 0$ *and* $t \geq 1$.

As is well known, the distribution of the tree $\mathcal{T}_t$ is not changed if ξ is replaced by another random variable ξ' whose distribution is created from that of ξ by *tilting* or *conjugation* (see e.g. [13]): if for every $k \geq 0$ we have $\Pr[\xi' = k] = c'\mu^k \Pr[\xi = k]$ for some $\mu > 0$ and normalizing constant c'. Thus, we see that Lemma 3 remains true for $\xi \sim \text{Poisson}(\mu)$, with $\mu > 0$, in which case the parameters $C, c > 0$ are universal constants which do not depend on the parameter μ. It is also well known (see e.g. Section 6.4 of [7]) that if $\xi \sim \text{Poisson}(\mu)$ then $\mathcal{T}_t$ is distributed as a random rooted labelled tree, that is, a tree picked uniformly from the t^{t-1} trees on vertices $\{1, 2, \dots, t\}$ in which one vertex is declared to be the root. From this we obtain an estimate to be used by us later.

Lemma 4. *Given* $0 \leq \ell \leq t$ *let* $p_{t,\ell}$ *denote the proportion of (rooted) labeled trees on* t *vertices which contain a path of length at least* ℓ. *There exist constants* $C, \varepsilon_0 > 0$ *such that for any* $\varepsilon \in (0, \varepsilon_0)$ *if* $\ell = \frac{C}{\varepsilon} \ln\left(\frac{1}{\varepsilon}\right)$ *and* $t_0 = \frac{15}{\varepsilon^2} \ln\left(\frac{1}{\varepsilon}\right)$ *then*

$$\sum_{\ell \leq t \leq t_0} p_{t,\ell} \leq \varepsilon^3$$

Proof of Lemma 4. It follows from Lemma 3 and the considerations above that there exist constants $C', c' > 0$ such that for every $t \leq t_0$:

$$p_{t,\ell} \leq C' \exp\left(-\frac{c'\ell^2}{t}\right) \leq C' \exp\left(-\frac{c'\left(\frac{C}{\varepsilon}\ln\left(\frac{1}{\varepsilon}\right)\right)^2}{\frac{15}{\varepsilon^2}\ln\left(\frac{1}{\varepsilon}\right)}\right) = C'\varepsilon^{\frac{c'C^2}{15}}.$$

Thus, if $C > \sqrt{\frac{90}{c'}}$ and if ε_0 is sufficiently small then we see that for any $\varepsilon \in (0, \varepsilon_0)$ and for $t \leq t_0$ we have $p_{t,\ell} \leq \varepsilon^6$. Using this we conclude that

$$\sum_{\ell \leq t \leq t_0} p_{t,\ell} \leq \varepsilon^6 \cdot t_0 = 15\varepsilon^4 \ln\left(\frac{1}{\varepsilon}\right) \leq \varepsilon^3\,,$$

provided ε_0 is sufficiently small, as claimed. □

The next lemma concerns the sizes of Poisson Galton-Watson trees which contain long paths.

Lemma 5. *For $\varepsilon > 0$ let $0 < \mu < 1$ be such that $\mu e^{-\mu} = (1+\varepsilon)e^{-(1+\varepsilon)}$. Given $\ell \geq 1$ consider a Poisson(μ)-Galton-Watson tree $\mathcal{T}$ and the random variable*

$$T_\ell := \begin{cases} |\mathcal{T}| & \text{if } \mathcal{T} \text{ contains a path of length at least } \frac{\ell}{3} \\ 0 & \text{otherwise}, \end{cases}$$

where $|\mathcal{T}|$ denotes the number of vertices of $\mathcal{T}$. Then there exist constants $C, \varepsilon_0 > 0$ such that for every $\varepsilon \in (0, \varepsilon_0)$ and for $\ell = \frac{C}{\varepsilon}\ln\left(\frac{1}{\varepsilon}\right)$ we have $\mathbb{E}\left[T_\ell\right] \leq 14\varepsilon^3$ and $Var\left[T_\ell\right] \leq \frac{8}{\varepsilon^3}$.

Proof. We have

$$\mathbb{E}\left[T_\ell\right] = \mathbb{E}\left[\mathbb{E}\left[T_\ell \mid |\mathcal{T}|\right]\right] = \sum_{t \geq 1} \Pr\left[|\mathcal{T}| = t\right] \cdot \mathbb{E}\left[T_\ell \mid |\mathcal{T}| = t\right]. \tag{1}$$

It is well-known (see, e.g., Section 6.6 of [7]) that the size of the Poisson(μ)-Galton-Watson tree $\mathcal{T}$ follows a Borel(μ) distribution, namely,

$$\Pr\left[|\mathcal{T}| = t\right] = \frac{t^{t-1}\left(\mu e^{-\mu}\right)^t}{\mu \cdot t!}.$$

Moreover, as discussed in the remarks that follow Lemma 3, if we condition a Poisson(μ)-Galton-Watson tree on it having exactly t vertices then it is identically distributed to a random rooted labelled tree on t vertices. Thus, it follows that $\mathbb{E}\left[T_\ell \mid |\mathcal{T}| = t\right]$ is equal to $t \cdot p_{t,\frac{\ell}{3}}$, where $p_{t,\frac{\ell}{3}}$ denotes the proportion of rooted labeled trees on t vertices which contain a path of length at least $\frac{\ell}{3}$. Hence, setting $t_0 := \frac{15}{\varepsilon^2}\ln\left(\frac{1}{\varepsilon}\right)$ with foresight, it follows from (1) that

$$\begin{aligned} \mathbb{E}\left[T_\ell\right] &= \sum_{t \geq 1} \frac{t^{t-1}\left(\mu e^{-\mu}\right)^t}{\mu \cdot t!} \cdot t \cdot p_{t,\frac{\ell}{3}} \\ &\leq \frac{1}{\mu} \sum_{t \geq \frac{\ell}{3}} \frac{t^t}{t!} \cdot (1+\varepsilon)^t \cdot e^{-(1+\varepsilon)t} \cdot p_{t,\frac{\ell}{3}} \\ &\leq 2 \sum_{t \geq \frac{\ell}{3}} e^{-\frac{\varepsilon^2}{3}t} \cdot p_{t,\frac{\ell}{3}} \\ &\leq 2 \cdot \left(\sum_{\frac{\ell}{3} \leq t \leq t_0} p_{t,\frac{\ell}{3}} + \sum_{t \geq t_0} e^{-\frac{\varepsilon^2}{3}t} \right), \end{aligned} \tag{2}$$

where in the second inequality we used the facts that $\frac{t^t}{t!} \leq e^t$, that $1+\varepsilon \leq e^{\varepsilon - \frac{\varepsilon^2}{3}}$ (which holds since the first terms of the Taylor series expansion of $\ln(1+\varepsilon)$ are $\varepsilon - \frac{\varepsilon^2}{2}$) and that $\frac{1}{\mu} \leq 2$ provided ε_0 is chosen sufficiently small. By Lemma 4 there exist constants $C, \varepsilon_0 > 0$ such that the first sum in (2) is at most ε^3. Moreover, the second sum in (2) is

$$\sum_{t \geq t_0} e^{-\frac{\varepsilon^2}{3}t} = e^{-\frac{\varepsilon^2}{3}t_0} \cdot \frac{1}{1 - e^{-\frac{\varepsilon^2}{3}}} \leq \varepsilon^5 \cdot \frac{6}{\varepsilon^2} = 6\varepsilon^3, \tag{3}$$

where we used the fact that $\frac{1}{1-e^{-x}} \leq \frac{2}{x}$ for $x > 0$ sufficiently small (which holds since the first terms of the Taylor series expansion of e^{-x} are $1 - x$). Thus, all in all, we conclude that there exist constants $C, \varepsilon_0 > 0$ such that

$$\mathbb{E}\left[T_\ell\right] \leq 2 \cdot \left(\varepsilon^3 + 6\varepsilon^3\right) = 14\varepsilon^3$$

as claimed. Since $|\mathcal{T}| \sim \mathrm{Borel}(\mu)$ it follows that

$$\mathrm{Var}\left[T_\ell\right] \le \mathbb{E}\left[T_\ell^2\right] \le \mathbb{E}\left[|\mathcal{T}|^2\right] = \frac{1}{(1-\mu)^3}.$$

If $\mu \le 1 - \frac{\varepsilon}{2}$ then we can conclude that $\mathrm{Var}\left[T_\ell\right] \le \frac{8}{\varepsilon^3}$, finishing the proof.

It suffices then to show that $\mu \le 1 - \frac{\varepsilon}{2}$ provided ε_0 is chosen small enough. This is an immediate consequence of the fairly standard estimate in the theory of random graphs that $\mu = 1 - \varepsilon + O(\varepsilon^2)$ as $\varepsilon \to 0$ (see, e.g. p. 140 of [6]). For the sake of completeness we provide a brief sketch here. Recall that $\mu \in (0,1)$ is defined as being the solution to $\mu e^{-\mu} = (1+\varepsilon)e^{-(1+\varepsilon)}$. Let $f : \mathbb{R} \to \mathbb{R}$ denote the function $f(x) = xe^{-x}$. Note that $f'(x) = (1-x)e^{-x}$, which is strictly positive for $x \in (0,1)$. This implies that f is strictly increasing in $(0,1)$ and so, in order to show that $\mu \le 1 - \frac{\varepsilon}{2}$, it is enough to show that $f(1-\frac{\varepsilon}{2}) \ge (1+\varepsilon)e^{-(1+\varepsilon)} = f(\mu)$, provided $\varepsilon > 0$ is small enough. Note that:

$$f\left(1-\frac{\varepsilon}{2}\right) = \left(1-\frac{\varepsilon}{2}\right)e^{-\left(1-\frac{\varepsilon}{2}\right)} \ge (1+\varepsilon)e^{-(1+\varepsilon)} \Leftrightarrow \left(1-\frac{\varepsilon}{2}\right)e^{\frac{3\varepsilon}{2}} \ge 1+\varepsilon$$

Since $e^x = \sum_{n=0}^{\infty} \frac{x^n}{n!} \ge 1 + x + \frac{x^2}{2}$ for $x \ge 0$, it follows that:

$$\left(1-\frac{\varepsilon}{2}\right)e^{\frac{3\varepsilon}{2}} \ge \left(1-\frac{\varepsilon}{2}\right)\left(1+\frac{3\varepsilon}{2}+\frac{\left(\frac{3\varepsilon}{2}\right)^2}{2}\right) = 1+\varepsilon+\frac{3\varepsilon^2}{8}-\frac{9\varepsilon^3}{16}.$$

The latter is at least $1+\varepsilon$, if $\varepsilon > 0$ is small enough. Thus, we conclude that $\mu \le 1 - \frac{\varepsilon}{2}$, as claimed. □

Lemma 6. *Let $P = (V,E)$ be a path of length ℓ and $B \subseteq E$ a set of size $|B| \le \alpha\ell$, where $\alpha \ge \frac{1}{\ell}$. Let Q denote the graph obtained from P by deleting all the edges in B. Then there exist vertex disjoint subpaths $\{Q^j\}_{j\in J}$ of Q such that each Q^j has length at least $\frac{1}{3\alpha}$ and the subpaths $\{Q^j\}_{i\in J}$ cover at least $\left(\frac{1}{3}-\alpha\right)\ell$ vertices of V.*

Proof of Lemma 6. Since P is a path, Q consists of a union of vertex disjoint paths $\{Q^j\}_{j\in[k]}$ for some $k \le |B|+1 \le \alpha\ell+1$. Denoting by ℓ_j the length of the path Q^j for $j \in [k]$, note that

$$\sum_{j\in[k]} \ell_j = \ell - |B| \ge (1-\alpha)\ell. \tag{4}$$

Moreover, setting $J := \{j \in [k] : \ell_j \ge \frac{1}{3\alpha}\}$ we see that

$$\sum_{j\notin J} \ell_j \le k \cdot \frac{1}{3\alpha} \le \frac{1}{3}\ell + \frac{1}{3\alpha} \le \frac{2}{3}\ell. \tag{5}$$

Putting (4) and (5) together we get that

$$\sum_{j\in J} \ell_j \ge \left(\frac{1}{3}-\alpha\right)\ell.$$

Thus, it follows that the paths $\{Q^j\}_{j\in J}$ satisfy the desired conditions. □

2.3 Properties of random graphs

The next lemma provides bounds on the sizes of the largest and second largest connected components of $G \sim \mathcal{G}(n,p)$ as well as the size of its 2-core when $p = \frac{1+\varepsilon}{n}$, where $\varepsilon > 0$ is a small constant. This lemma is a simple consequence of Theorem 5.4 of [11] and Theorem 3 of [15].

Lemma 7. *Let $p = \frac{1+\varepsilon}{n}$ where $\varepsilon > 0$ is a constant. Then there exists a constant $\varepsilon_0 > 0$ such that for every $\varepsilon \in (0, \varepsilon_0)$ the following holds whp for $G \sim \mathcal{G}(n,p)$:*

(a) the largest connected component of G has between εn and $3\varepsilon n$ vertices.

(b) the second largest connected component of G has at most $\frac{20}{\varepsilon^2}\ln n$ vertices.

(c) the 2-core of the largest connected component of G has at most $2\varepsilon^2 n$ vertices.

In [8], Ding, Lubetzky and Peres established a complete characterization of the structure of the giant component $\mathcal{C}_1$ of $G \sim \mathcal{G}(n,p)$ in the strictly supercritical regime ($p = \frac{1+\varepsilon}{n}$ with $\varepsilon > 0$ constant). This was achieved by offering a tractable contiguous model $\tilde{\mathcal{C}}_1$, i.e. a model such that every graph property that is satisfied by $\tilde{\mathcal{C}}_1$ whp is also satisfied by $\mathcal{C}_1$ whp. In their model, $\tilde{\mathcal{C}}_1$ consists of a 2-core $\tilde{\mathcal{C}}_1^{(2)}$ where one attaches to each vertex of $\tilde{\mathcal{C}}_1^{(2)}$ one independent Poisson(μ)-Galton-Watson tree (where $0 < \mu < 1$ is such that $\mu e^{-\mu} = (1+\varepsilon)e^{-(1+\varepsilon)}$). In light of this, any graph property that is satisfied whp by the disjoint union of $|\tilde{\mathcal{C}}_1^{(2)}|$ independent Poisson(μ)-Galton-Watson trees must also be satisfied whp by $\mathcal{C}_1 \setminus \mathcal{C}_1^{(2)}$, the graph obtained from the giant component $\mathcal{C}_1$ by removing the edges of its 2-core $\mathcal{C}_1^{(2)}$. As one would expect, the random variable $|\tilde{\mathcal{C}}_1^{(2)}|$ is tightly concentrated around its expectation, which agrees with the expected size of the 2-core $\mathcal{C}_1^{(2)}$ of $\mathcal{C}_1$. By (c) of Lemma 7 this at most $2\varepsilon^2 n$. The next technical lemma which will be useful in the proof of Theorem 2 follows from the considerations above.

Lemma 8. *Let $\mathcal{C}_1$ denote the largest connected component of $G \sim \mathcal{G}(n,p)$ for $p = \frac{1+\varepsilon}{n}$, where $\varepsilon > 0$ is fixed, let $\mathcal{C}_1^{(2)}$ denote its 2-core and let $\mathcal{C}_1 \setminus \mathcal{C}_1^{(2)}$ denote the graph obtained from $\mathcal{C}_1$ by removing the edges in $\mathcal{C}_1^{(2)}$. Let $0 < \mu < 1$ be such that $\mu e^{-\mu} = (1+\varepsilon)e^{-(1+\varepsilon)}$ and consider $2\varepsilon^2 n$ independent Poisson(μ)-Galton-Watson trees $\mathcal{T}_1, \ldots, \mathcal{T}_{2\varepsilon^2 n}$. Then, for every ℓ and m (which might depend on n) if whp the disjoint union of $\mathcal{T}_1, \ldots, \mathcal{T}_{2\varepsilon^2 n}$ does not contain a set of vertex disjoint paths of length at least ℓ covering at least m vertices then the same holds whp for $\mathcal{C}_1 \setminus \mathcal{C}_1^{(2)}$.*

3 Proof of Theorem 1

We start this section by repeating the statement of Theorem 1 for the reader's convenience.

Theorem 1. *There exists an absolute constant $C > 0$ such that the following holds. For every constant $q \in (0,1)$ there exist $n_0, \varepsilon_0 > 0$ such that for every fixed $\varepsilon \in (0, \varepsilon_0)$ and any $n \geq n_0$ there is no adaptive algorithm which reveals a path of length $\ell \geq \frac{3C}{\varepsilon} \ln\left(\frac{1}{\varepsilon}\right)$ with probability at least q in $G \sim \mathcal{G}(n,p)$, where $p = \frac{1+\varepsilon}{n}$, by querying at most $\frac{q\ell}{8640 Cp\varepsilon \ln\left(\frac{1}{\varepsilon}\right)}$ pairs of vertices.*

Proof of Theorem 1. Suppose Alg is an adaptive algorithm which with probability at least q finds a path of length ℓ in $G \sim \mathcal{G}(n,p)$, where $p = \frac{1+\varepsilon}{n}$, after querying at most $\frac{q\ell}{8640 Cp\varepsilon \ln\left(\frac{1}{\varepsilon}\right)}$ pairs of vertices. We consider implicitly that Alg takes an *ordered* vertex set as part of its input. We shall assume henceforth that $n, C > 0$ are sufficiently large and $\varepsilon > 0$ is sufficiently small in order to obtain a contradiction. Note that, by restricting Alg to a set of n vertices, we get an algorithm which for any $n' \geq n$ with probability at least q finds in $G' \sim \mathcal{G}(n', p)$ a path of length ℓ after querying at most $\frac{q\ell}{8640Cp\varepsilon \ln\left(\frac{1}{\varepsilon}\right)}$ pairs of vertices. We shall abuse notation slightly and call Alg to all these algorithms.

Define $n' := \left(1 + \frac{720\varepsilon^2}{q}\right) n$, $V_0 := [n']$, $I_0 := \emptyset$ and $s := \frac{720\varepsilon^2 n}{q(\ell+1)}$. For $i = 1, \ldots, s$ do the following:

- Apply Alg to $G_{i-1} \sim \mathcal{G}(V_{i-1}, p)$, where the vertices in V_{i-1} are permuted according to a permutation $\pi_i \in S_{V_{i-1}}$ chosen uniformly at random. Let L_i be the graph of all pairs of vertices queried and let $K_i \subseteq L_i$ be the graph of edges present. By the algorithm we know that L_i has at most $\frac{q\ell}{8640Cp\varepsilon \ln\left(\frac{1}{\varepsilon}\right)}$ edges. If K_i contains a path of length ℓ then let P_i be one such path, define $V_i := V_{i-1} \setminus V(P_i)$ and set $I_i := I_{i-1} \cup \{i\}$. Otherwise, set $V_i := V_{i-1}$ and $I_i := I_{i-1}$.

Observe that $|V_s| \geq n' - (\ell+1)s = \left(1 + \frac{720\varepsilon^2}{q}\right) n - \frac{720\varepsilon^2}{q} n = n$ and so we can indeed apply Alg to V_{i-1} for any $i \in [s]$. We define a random graph H with vertex set V_0 in the following way. For every pair of vertices $\{u, v\} \subseteq V_0$ if $\{u, v\} \in E(L_i)$ for some $i \in [s]$ then let i_0 be the smallest such index and set $\{u, v\}$ as an edge of H if and only if $\{u, v\} \in E(K_{i_0})$. Consider all the other pairs $\{u, v\} \subseteq V_0$ as non-edges of H. From the procedure above it follows that for every $\{u, v\} \subseteq V_0$ we have independently that

$$\Pr\left[\{u, v\} \in E(H)\right] \leq p = \frac{1+\varepsilon}{n} = \frac{1+\varepsilon}{n'} \cdot \frac{n'}{n} = \frac{(1+\varepsilon)\left(1 + \frac{720\varepsilon^2}{q}\right)}{n'} \leq \frac{1+2\varepsilon}{n'},$$

provided $\varepsilon \leq \frac{q}{1440}$. Thus, the graph H can be viewed as a subgraph of a graph sampled from $\mathcal{G}\left(n', \frac{1+2\varepsilon}{n'}\right)$. In particular, if with probability at least $\frac{q^2}{4}$ the graph H contains a set of vertex disjoint paths of length at least $\frac{C}{\varepsilon} \ln\left(\frac{1}{\varepsilon}\right)$ which cover at least $52\varepsilon^2 n'$ vertices then the same must also hold with probability at least $\frac{q^2}{4}$ in $\mathcal{G}\left(n', \frac{1+2\varepsilon}{n'}\right)$. However, this would contradict Theorem 2 and so it suffices to prove the following claim:

Claim. *With probability at least $\frac{q^2}{4}$ the graph H contains a set of vertex disjoint paths of length at least $\frac{C}{\varepsilon} \ln\left(\frac{1}{\varepsilon}\right)$ which cover at least $52\varepsilon^2 n'$ vertices of V_0.*

Define for each $i \in I_s$ the graph H_i with vertex set V_{i-1} and edge set $\left(\bigcup_{j=1}^{i-1} E(L_j)\right) \cap \binom{V_{i-1}}{2}$ and note that

$$|E(H_i)| \leq s \cdot \frac{q\ell}{8640Cp\varepsilon \ln\left(\frac{1}{\varepsilon}\right)} \leq \frac{\varepsilon n^2}{12C \ln\left(\frac{1}{\varepsilon}\right)(1+\varepsilon)} \leq \frac{\varepsilon}{6C \ln\left(\frac{1}{\varepsilon}\right)} \cdot \binom{|V_{i-1}|}{2}. \tag{6}$$

Observe that for each $i \in I_s$ the set $V_{i-1} \setminus V_i$ consists of the vertex set of a path P_i in the graph K_i. For each such i set $B_i := E(P_i) \cap E(H_i)$ and let Q_i denote the graph obtained from P_i by deleting all the edges in B_i. Note crucially that $E(Q_i) \subseteq E(H)$ and that the graphs $\{Q_i\}_{i \in I_s}$ are vertex disjoint.

Consider now the set $I := \left\{i \in I_s : |B_i| \leq \frac{\varepsilon}{3C \ln\left(\frac{1}{\varepsilon}\right)} \ell\right\}$. By Lemma 6 it follows that for any $i \in I$ there exist vertex disjoint subpaths $\{Q_i^j\}_{j \in J_i}$ of Q_i each of length at least $\frac{C}{\varepsilon} \ln\left(\frac{1}{\varepsilon}\right)$ which cover at least $\left(\frac{1}{3} - \frac{\varepsilon}{3C \ln\left(\frac{1}{\varepsilon}\right)}\right) \ell \geq \frac{1}{4}(\ell+1)$ vertices of $V(Q_i)$. Thus, if $|I| \geq \frac{1}{3} sq$ then $\{Q_i^j\}_{i \in I, j \in J_i}$ forms a collection of vertex disjoint paths in H of length at least $\frac{C}{\varepsilon} \ln\left(\frac{1}{\varepsilon}\right)$ which cover at least $\frac{1}{4}(\ell+1) \cdot \frac{1}{3} sq = 60\varepsilon^2 n \geq 52\varepsilon^2 n'$ vertices of V_0. It suffices to show then that with probability at least $\frac{q^2}{4}$ we have $|I| \geq \frac{1}{3} sq$.

Let $I' := [s] \setminus I$ and note that for every $i \in [s]$ we have

$$\Pr\left[i \in I'\right] = \Pr\left[i \notin I_s\right] + \Pr\left[i \in I' \mid i \in I_s\right] \cdot \Pr\left[i \in I_s\right]. \tag{7}$$

It is clear from the procedure above that for each $i \in [s]$ we have $\Pr\left[i \in I_s\right] \geq q$. Note also crucially that, provided $i \in I_s$, the path P_i is a randomly mapped path of length ℓ on the vertex set V_{i-1}. Indeed, this happens because before the i-th application of Alg we permuted the vertices of V_{i-1} according to a permutation $\pi_i \in S_{V_{i-1}}$ chosen uniformly at random. Thus, by conditioning on the event that $i \in I_s$, on any possible graph H_i satisfying (6) and on the path $\pi_i^{-1}(P_i)$, we have for any $e \in E\left(\pi_i^{-1}(P_i)\right)$:

$$\Pr\left[\pi_i(e) \in E(H_i)\right] \leq \frac{\varepsilon}{6C \ln\left(\frac{1}{\varepsilon}\right)},$$

and so, by linearity of expectation it follows that:

$$\mathbb{E}\left[|E(P_i) \cap E(H_i)|\right] \leq \frac{\varepsilon}{6C \ln\left(\frac{1}{\varepsilon}\right)} \ell.$$

Thus, by Markov's inequality (see, e.g., [4]) we get that

$$\Pr\left[i \in I' \mid i \in I_s\right] \leq \frac{1}{2},$$

and so by equation (7) we see that for any $i \in [s]$ we have $\Pr\left[i \in I'\right] \leq 1 - \frac{1}{2}\Pr[i \in I_s] \leq 1 - \frac{q}{2}$. It follows then by linearity of expectation that $\mathbb{E}\left[|I'|\right] \leq s\left(1 - \frac{q}{2}\right)$. Hence, again by Markov's inequality we conclude that

$$\Pr\left[|I'| \geq \frac{s}{1 + \frac{q}{2}}\right] \leq 1 - \frac{q^2}{4}\ \text{, which implies}\ \ \frac{q^2}{4} \leq \Pr\left[|I| \geq \frac{sq}{2+q}\right] \leq \Pr\left[|I| \geq \frac{sq}{3}\right].$$

This completes the proof. □

4 Proof of Theorem 2

Theorem 2. *There exist constants $C, \varepsilon_0 > 0$ such that for every fixed $\varepsilon \in (0, \varepsilon_0)$ we have whp that $G \sim \mathcal{G}\left(n, \frac{1+\varepsilon}{n}\right)$ does not contain a set of vertex disjoint paths of lengths at least $\frac{C}{\varepsilon}\ln\left(\frac{1}{\varepsilon}\right)$ whose union covers at least $13\varepsilon^2 n$ vertices.*

Proof of Theorem 2. Let $G \sim \mathcal{G}(n,p)$ where $p = \frac{1+\varepsilon}{n}$. Let $\mathcal{C}_1$ denote the largest connected component of G, let $\mathcal{C}_1^{(2)}$ denote the 2-core of $\mathcal{C}_1$ and let $\mathcal{C}_1 \setminus \mathcal{C}_1^{(2)}$ denote the graph obtained from $\mathcal{C}_1$ by deleting the edges in $\mathcal{C}_1^{(2)}$. For $\ell \geq 1$ consider the following random variables:

- $X_\ell =$ number of vertices which belong to connected components of G of size at most $\frac{20}{\varepsilon^2}\ln n$ containing a path of length at least ℓ.
- $Y_\ell =$ maximum number of vertices covered by vertex disjoint paths of length at least ℓ in $\mathcal{C}_1$.
- $Z_\ell =$ maximum number of vertices covered by vertex disjoint paths of length at least $\frac{\ell}{3}$ in $\mathcal{C}_1 \setminus \mathcal{C}_1^{(2)}$.

By (b) of Lemma 7 it follows that whp $X_\ell + Y_\ell$ is an upper bound on the maximum number of vertices of G covered by vertex disjoint paths of length at least ℓ. Note that we may assume that all the paths considered have size at most 2ℓ by splitting larger paths into several paths of length at least ℓ. Moreover, if P is a path of length at least ℓ in $\mathcal{C}_1$ then, since $\mathcal{C}_1 \setminus \mathcal{C}_1^{(2)}$ consists of a disjoint union of trees, there must exist a subpath P' of the path P with at least $\frac{|P|}{3} \geq \frac{\ell}{3}$ vertices which lies in $\mathcal{C}_1^{(2)}$ or in $\mathcal{C}_1 \setminus \mathcal{C}_1^{(2)}$. Since $|P| \leq 6|P'|$ it follows that $Y_\ell \leq 6|\mathcal{C}_1^{(2)}| + 6Z_\ell$.

By (c) of Lemma 7 we know that whp $|\mathcal{C}_1^{(2)}| \leq 2\varepsilon^2 n$, provided ε_0 is chosen small enough. It suffices then to show that there exist constants $C, \varepsilon_0 > 0$ such that for every fixed $\varepsilon \in (0, \varepsilon_0)$ and for $\ell := \frac{C}{\varepsilon} \ln\left(\frac{1}{\varepsilon}\right)$ we have whp that

$$X_\ell < \varepsilon^3 n \quad \text{and} \quad Z_\ell < 29\varepsilon^5 n.$$

since in that case we have whp that the maximum number of vertices of G covered by vertex disjoint paths of length at least ℓ is at most

$$X_\ell + Y_\ell \leq X_\ell + 6|\mathcal{C}_1^{(2)}| + 6Z_\ell < \varepsilon^3 n + 6 \cdot 2\varepsilon^2 n + 6 \cdot 29\varepsilon^5 n \leq 13\varepsilon^2 n.$$

provided ε_0 is chosen sufficiently small. Lemmas 9 and 10 complete the proof. □

Lemma 9. *There exist constants $C, \varepsilon_0 > 0$ such that for every fixed $\varepsilon \in (0, \varepsilon_0)$ and for $\ell := \frac{C}{\varepsilon} \ln\left(\frac{1}{\varepsilon}\right)$ we have $X_\ell < \varepsilon^3 n$ whp.*

Proof of Lemma 9. Given a set $S \subseteq [n]$ of size t, let $\mathcal{S}_\ell(S)$ (resp. $\mathcal{T}_\ell(S)$) denote the set of possible connected graphs (resp. spanning trees) on the vertex set S which contain a path of length at least ℓ. Let X_S denote the indicator random variable of the event that $G[S] \in \mathcal{S}_\ell(S)$ and that there are no edges in G between S and $[n] \setminus S$. Note that $G[S] \in \mathcal{S}_\ell(S)$ if and only if there exists $T \in \mathcal{T}_\ell(S)$ such that $T \subseteq G[S]$. Thus, by the union bound we have

$$\mathbb{E}\left[X_S\right] \leq |\mathcal{T}_\ell(S)| \cdot p^{t-1} \cdot (1-p)^{t(n-t)} \tag{8}$$

where the first term accounts for taking a union bound over all $T \in \mathcal{T}_\ell(S)$, the second term accounts for the probability that the edges in T are present in $G[S]$ and the last term accounts for the probability that none of the edges between S and $[n] \setminus S$ are present in G. Note that $|\mathcal{T}_\ell(S)|$ does not depend on the set S and is equal to the number of labeled trees on t vertices which contain a path of length at least ℓ. More specifically, if $p_{t,\ell}$ denotes the proportion of labeled trees on t vertices which contain a path of length at least ℓ, then $|\mathcal{T}_\ell(S)| = p_{t,\ell} \cdot t^{t-2}$. Observe now that the random variable X_ℓ satisfies the following:

$$X_\ell \leq \sum_{t=\ell}^{\frac{20}{\varepsilon^2}\ln n} \sum_{S \in \binom{[n]}{t}} t \cdot X_S.$$

We claim that for $\ell := \frac{C}{\varepsilon} \ln\left(\frac{1}{\varepsilon}\right)$, where $C > 0$ is a large constant, and for some constant $\varepsilon_0 > 0$, if $\varepsilon \in (0, \varepsilon_0)$ is fixed then $\Pr\left[X_\ell \geq \varepsilon^3 n\right] = o(1)$. To prove this claim we start by estimating $\mathbb{E}[X_\ell]$. Setting $t_0 := \frac{15}{\varepsilon^2} \ln\left(\frac{1}{\varepsilon}\right)$, we have by the linearity of expectation and by (8) that if ε_0 is sufficiently small then:

$$\mathbb{E}[X_\ell] \leq \sum_{t=\ell}^{\frac{20}{\varepsilon^2}\ln n} t \cdot \binom{n}{t} \cdot p_{t,\ell} \cdot t^{t-2} \cdot p^{t-1} \cdot (1-p)^{t(n-t)}$$

$$\leq \sum_{t=\ell}^{\frac{20}{\varepsilon^2}\ln n} t \cdot \left(\frac{en}{t}\right)^t \cdot p_{t,\ell} \cdot t^{t-2} \cdot \left(\frac{1+\varepsilon}{n}\right)^{t-1} \left(1 - \frac{1+\varepsilon}{n}\right)^{t(n-t)}$$

$$\leq \sum_{t=\ell}^{\frac{20}{\varepsilon^2}\ln n} e^t \cdot t^{-1} \cdot n \cdot p_{t,\ell} \cdot \frac{e^{\varepsilon t - \frac{\varepsilon^2}{3}t}}{1+\varepsilon} \cdot e^{-(1+\varepsilon)t + \frac{(1+\varepsilon)t^2}{n}}$$

$$\leq \frac{(1+o(1))n}{\ell(1+\varepsilon)} \cdot \sum_{t \geq \ell} p_{t,\ell} \cdot e^{-\frac{\varepsilon^2}{3}t}$$

$$\leq \frac{n}{14} \cdot \left(\sum_{\ell \leq t \leq t_0} p_{t,\ell} + \sum_{t \geq t_0} e^{-\frac{\varepsilon^2}{3}t} \right) \tag{9}$$

where in the third inequality we used the fact that $(1+\varepsilon)^t \leq e^{\varepsilon t - \frac{\varepsilon^2}{3}t}$ for sufficiently small $\varepsilon > 0$. By Lemma 4 there exist constants $C, \varepsilon_0 > 0$ such that the first sum in (9) is at most ε^3. Moreover, by (3) the second sum in (9) is at most $6\varepsilon^3$. Thus, all in all, we conclude that there exist constants $C, \varepsilon_0 > 0$ such that

$$\mathbb{E}\left[X_\ell\right] \leq \frac{n}{14} \cdot (\varepsilon^3 + 6\varepsilon^3) = \frac{\varepsilon^3 n}{2}.$$

Note that if G and H differ in precisely one edge then $|X_\ell(G) - X_\ell(H)| \leq \frac{40}{\varepsilon^2}\ln n$ because one edge affects at most two connected components of size at most $\frac{20}{\varepsilon^2}\ln n$. Thus, by Lemma 2 it follows that

$$\Pr\left[X_\ell > \varepsilon^3 n\right] \leq \Pr\left[|X_\ell - \mathbb{E}[X_\ell]| > \frac{\varepsilon^3 n}{2}\right] \leq e^{-\Omega\left(\frac{n}{(\ln n)^2}\right)} = o(1).$$

□

Remark. An alernative approach to the proof of Lemma 9 would be to invoke the so called symmetry rule (see, e.g., Chapter 5.6 of [11]), postulating that in the supercritical regime $p = \frac{1+\varepsilon}{n}$, the subgraph of $G \sim \mathcal{G}(n,p)$ outside the giant component behaves typically as a random graph with subcritical edge probability. One can then estimate the likely contribution of paths of length at least $\ell = \frac{C}{\varepsilon}\ln\left(\frac{1}{\varepsilon}\right)$ coming from the small components to the total volume of vertex disjoint paths of length at least ℓ and to show it to be $O(\varepsilon^2 n)$ whp, using a direct first moment argument. Since we still need to treat the paths residing in the giant component outside the 2-core (the random variable Z_ℓ), we chose to adopt a unified approach using the machinery of Galton-Watson trees developed in Section 2.2, and to apply it here as well.

Lemma 10. *There exist constants $C, \varepsilon_0 > 0$ such that for every fixed $\varepsilon \in (0, \varepsilon_0)$ and for $\ell := \frac{C}{\varepsilon}\ln\left(\frac{1}{\varepsilon}\right)$ we have $Z_\ell < 29\varepsilon^5 n$ whp.*

Proof of Lemma 10. Recall that Z_ℓ counts the maximum number of vertices covered by vertex disjoint paths of length at least $\frac{\ell}{3}$ in $\mathcal{C}_1 \backslash \mathcal{C}_1^{(2)}$. Let $0 < \mu < 1$ be such that $\mu e^{-\mu} = (1+\varepsilon)e^{-(1+\varepsilon)}$ and consider $2\varepsilon^2 n$ independent Poisson(μ)-Galton-Watson trees $\mathcal{T}_1, \ldots, \mathcal{T}_{2\varepsilon^2 n}$. By Lemma 8 it suffices for our purposes to show that whp the maximum number of vertices covered by vertex disjoint paths of length at least $\frac{\ell}{3}$ in the disjoint union of $\mathcal{T}_1, \ldots, \mathcal{T}_{2\varepsilon^2 n}$ is less than $29\varepsilon^5 n$, for appropriate $C, \varepsilon_0 > 0$.

For each $1 \leq i \leq 2\varepsilon^2 n$ consider the following random variable:

$$T_{i,\ell} := \begin{cases} |\mathcal{T}_i| & \text{if } \mathcal{T}_i \text{ contains a path of length at least } \frac{\ell}{3} \\ 0 & \text{otherwise} \end{cases}$$

and set $T_\ell = \sum_{i=1}^{2\varepsilon^2 n} T_{i,\ell}$. Clearly T_ℓ is an upperbound on the maximum number of vertices covered by vertex disjoint paths of length at least $\frac{\ell}{3}$ in in the disjoint union of $\mathcal{T}_1, \dots, \mathcal{T}_{2\varepsilon^2 n}$. To finish the proof, we show that whp $T_\ell < 29\varepsilon^5 n$, provided $C, \varepsilon_0 > 0$ are chosen appropriately.

By Lemma 5 we know that there exist constants $C, \varepsilon_0 > 0$ such that for every $\varepsilon \in (0, \varepsilon_0)$ and for $\ell = \frac{C}{\varepsilon} \ln\left(\frac{1}{\varepsilon}\right)$ we have $\mathbb{E}[T_{i,\ell}] \leq 14\varepsilon^3$ and $\mathrm{Var}[T_{i,\ell}] \leq \frac{8}{\varepsilon^3}$. Thus, since the random variables $T_{i,\ell}$ are independent, we have that

$$\mathbb{E}[T_\ell] \leq 14\varepsilon^3 \cdot 2\varepsilon^2 n = 28\varepsilon^5 n \quad \text{and} \quad \mathrm{Var}[T_\ell] \leq \frac{8}{\varepsilon^3} \cdot 2\varepsilon^2 n = \frac{16n}{\varepsilon}.$$

Thus, by Chebyshev's Inequality (see, e.g., [4]) we conclude that

$$\Pr\left[T_\ell \geq 29\varepsilon^5 n\right] \leq \Pr\left[|T_\ell - \mathbb{E}[T_\ell]| \geq \varepsilon^5 n\right] \leq \frac{\mathrm{Var}[T_\ell]}{\varepsilon^{10} n^2} \leq \frac{16}{\varepsilon^{11} n} = o(1).$$

□

5 Concluding remarks

We have shown that in order to find a path of length $\ell = \Omega\left(\frac{\log\left(\frac{1}{\varepsilon}\right)}{\varepsilon}\right)$ in $G \sim \mathcal{G}(n,p)$ with at least some constant probability, where $p = \frac{1+\varepsilon}{n}$ with $\varepsilon > 0$ fixed, one needs to query at least $\Omega\left(\frac{\ell}{p\varepsilon \log\left(\frac{1}{\varepsilon}\right)}\right)$ pairs of vertices. This is close to best possible since a randomised depth first search algorithm from [14] finds whp a path of length ℓ after querying at most $O\left(\frac{\ell}{p\varepsilon}\right)$ pairs of vertices. A natural question, which remains open, is to close the gap between these bounds. We believe that every adaptive algorithm which reveals whp a path of length ℓ in $G \sim \mathcal{G}(n,p)$, where $p = \frac{1+\varepsilon}{n}$ with $\varepsilon > 0$ fixed, has to query $\Omega\left(\frac{\ell}{p\varepsilon}\right)$ pairs of vertices.

Recall that, to prove our main result, in Theorem 2 we bounded the total number of vertices covered by vertex disjoint paths of size at least $\Omega\left(\frac{1}{\varepsilon}\log\left(\frac{1}{\varepsilon}\right)\right)$ in a typical graph sampled from $\mathcal{G}(n,p)$, $p = \frac{1+\varepsilon}{n}$, by $O\left(\varepsilon^2 n\right)$. Since a graph $G \sim \mathcal{G}(n,p)$ contains whp a path of length $\Theta(\varepsilon^2 n)$ (see e.g. [11]), this is best possible up to a multiplicative constant. If one can show that a similar statement holds for paths of length $\Omega\left(\frac{1}{\varepsilon}\right)$ then one can modify our proof to obtain a $\Omega\left(\frac{\ell}{p\varepsilon}\right)$ bound in Theorem 1.

In the proof of Theorem 2 we needed to bound the number of vertices covered by vertex disjoint paths of a prescribed length ℓ in a random tree of fixed size t (Lemma 5). Our estimate was a bit wasteful because for trees which contained a path of length ℓ we used their total number of vertices t instead of the number of vertices covered by vertex disjoint paths of length ℓ, which is most likely significantly smaller. A way to fix this is to obtain good bounds for the following question:

Question. *Given $a = a(t) \in \mathbb{N}$ and $b = b(t) \in \mathbb{N}$ what is the probability that a random tree on t vertices contains b vertex disjoint paths, each of length at least a?*

Note that, since the diameter of a random tree on t vertices is whp $\Theta(\sqrt{t})$ (see e.g. [1]), the only interesting regime is when $ab \geq C\sqrt{t}$ for some constant $C > 0$. Moreover, by splitting paths of length larger than $2a$ into smaller subpaths of length at least a, we may consider only paths of length between a and $2a$.

One possible approach to this problem would be through a nice argument of Joyal ([12], see also [2]). It shows that a random tree $\mathcal{T}$ on t vertices can be obtained from a random map $f : [t] \to [t]$

as follows. First we create the directed graph D (possibly with loops) on vertex set $[t]$ with edges $i \to f(i)$ for each $i \in [t]$. Then we look at a maximal set of vertices $M = \{i_1, \ldots, i_m\} \subseteq [t]$ such that $f|_M$ is a permutation. We remove the directed edges inside M and replace them by the path $f(i_1) \to f(i_2) \to \ldots \to f(i_m)$ (where $i_1 < i_2 < \ldots < i_m$). By ignoring the orientations of the edges we obtain the desired tree $\mathcal{T}$. Note that, since the vertices in M form a path in $\mathcal{T}$, we must have $|M| = O(\sqrt{t})$ whp. Moreover, if we have a path P in $\mathcal{T}$ then a moment's thought reveals that either P has at least $\frac{|V(P)|}{3}$ vertices in M or there are $\frac{|V(P)|}{3}$ vertices of P which form a directed path in D. Thus, it follows that if we have a collection of b vertex disjoint paths in $\mathcal{T}$ each of length between a and $2a$ then D contains a collection of vertex disjoint directed paths each of length between $\frac{a+1}{3}$ and $2a$ covering at least $\frac{(a+1)b}{3} - |M|$ vertices. Since $|M| = O(\sqrt{t})$ whp and since we are interested only in the case when $ab \geq C\sqrt{t}$ for some large constant $C > 0$, it follows that in that case we have, say, at least $\frac{b}{10}$ such paths. Thus, up to changing a and b by constant multiplicative factors, it is enough to estimate the probability that the directed graph D obtained from a random map $f : [t] \to [t]$ contains at least b vertex disjoint directed paths, each of length (at least) a.

We can give a simple upper bound on this probability by taking the union bound over all collections of b vertex disjoint directed paths of length a. This shows that the probability that we want to estimate is at most

$$\frac{t!}{(t-(a+1)b)!b!}\left(\frac{1}{t}\right)^{ab} = \frac{t^b}{b!}\prod_{i=1}^{(a+1)b-1}\left(1-\frac{i}{t}\right) \leq e^{b+b\ln(t/b)-\binom{(a+1)b}{2}/t}.$$

Unfortunately, this upper bound is not strong enough to allow us to prove Theorem 2 for paths of length at least $\Omega\left(\frac{1}{\varepsilon}\right)$ because when b is roughly a constant and a is close to $\sqrt{t}$ the positive term $b\ln(t/b)$ in the exponent is much larger than the negative term $\binom{(a+1)b}{2}/t$. Thus, it would be nice to obtain tighter bounds for the probability in question.

Acknowledgement. We would like to thank the anonymous referees for carefully reading our paper as well as for their helpful comments. Parts of this work were carried out when the third author visited the School of Mathematical Sciences of Tel Aviv University, Israel. We would like to thank this institution for its hospitality and for creating a stimulating research environment.

References

[1] L. Addario-Berry, L. Devroye, S. Janson, Sub-Gaussian tail bounds for the width and height of conditioned Galton-Watson trees, *Annals of Probability*, Vol. 41 (2), pp. 1072–1087, 2013.

[2] M. Aigner and G. M. Ziegler, **Proofs from THE BOOK**, 5^{th} Ed., Springer, 2014.

[3] M. Ajtai, J. Komlós and E. Szemerédi, The longest path in a random graph, *Combinatorica*, Vol. 1 (1), pp. 1–12, 1981.

[4] N. Alon and J. H. Spencer, **The Probabilistic Method**, 3^{rd} Ed., John Wiley & Sons, 2008.

[5] O. Angel, I. Benjamini, E. Ofek and U. Wieder, Routing complexity of faulty networks, *Random Structures & Algorithms*, Vol. 32 (1), pp. 71–87, 2008.

[6] B. Bollobás, **Random Graphs**, 2^{nd} Ed., Cambridge University Press, 2001.

[7] L. Devroye, Branching processes and their applications in the analysis of tree structures and tree algorithms, In: *Probabilistic methods for algorithmic discrete mathematics*, Springer, pp. 249–314, 1998.

[8] J. Ding, E. Lubetzky and Y. Peres, Anatomy of the giant component: The strictly supercritical regime, *European Journal of Combinatorics*, Vol. 35, pp. 155–168, 2014.

[9] A. Ferber, M. Krivelevich, B. Sudakov and P. Vieira, Finding Hamilton cycles in random graphs with few queries, *Random Structures & Algorithms*, to appear.

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

[11] S. Janson, T. Łuczak, and A. Ruciński, **Random graphs**, John Wiley & Sons, 2000.

[12] A. Joyal, Une théorie combinatoire des séries formelles, *Advances in Mathematics*, Vol. 42 (1), pp. 1–82, 1981.

[13] D. P. Kennedy, The Galton-Watson process conditioned on the total progeny, *Journal of Applied Probability*, Vol. 12, pp. 800-806, 1975.

[14] M. Krivelevich and B. Sudakov, The phase transition in random graphs: a simple proof, *Random Structures & Algorithms*, Vol. 43 (2), pp. 131–138, 2013.

[15] B. Pittel, On tree census and the giant component in sparse random graphs, *Random Structures & Algorithms*, Vol. 1 (3), pp. 311–342, 1990.