SIAM J. DISCRETE MATH.
Vol. 35, No. 1, pp. 267–280

LARGE INDUCED MATCHINGS IN RANDOM GRAPHS*

OLIVER COOLEY†, NEMANJA DRAGANIĆ‡, MIHYUN KANG†, AND BENNY SUDAKOV‡

Abstract. Given a large graph H, does the binomial random graph $G(n,p)$ contain a copy of H as an induced subgraph with high probability? This classical question has been studied extensively for various graphs H, going back to the study of the independence number of $G(n,p)$ by Erdős and Bollobás and by Matula in 1976. In this paper we prove an asymptotically best possible result for induced matchings by showing that if $C/n \leq p \leq 0.99$ for some large constant C, then $G(n,p)$ contains an induced matching of order approximately $2\log_q(np)$, where $q = \frac{1}{1-p}$.

Key words. random graphs, induced matchings, Talagrand's inequality, Paley–Zygmund inequality

AMS subject classifications. 05C80, 05C70

DOI. 10.1137/20M1330609

1. Introduction.

Let $G(n,p)$ denote the binomial random graph on vertex set $[n] := \{1, 2, \ldots, n\}$, where each edge is included independently with probability p. The following classic question has been extensively studied in the theory of random graphs: Given a large graph H, does $G(n,p)$ contain a copy of H as an induced subgraph with high probability (abbreviated to whp, meaning with probability tending to 1 as n tends to infinity)? One of the first instances of this problem is determining the independence number of $G(n,p)$; i.e., when $H = H_k$ is an empty graph on k vertices, how large can k be such that H is an induced subgraph of $G(n,p)$? The study of this particular instance dates back to 1976 when Bollobás and Erdős [2] and Matula [16] showed that the independence number of $G(n,p)$ for constant p is asymptotically $2\log_q(np)$, where $q = 1/(1-p)$. A simple first moment argument shows that the size of this empty subgraph is asymptotically largest possible. Frieze [9] extended this result to the sparse regime, when $p = c/n$ for a large enough constant c, with the same expression $2\log_q(np)$ for the asymptotic size of the largest independent set. Indeed, it can be shown that the same result holds for every $p = p(n) \gg 1/n$ (see, e.g., [12]).

Another classic result which deals with nonempty induced subgraphs of $G(n,p)$ is due to Erdős and Palka [8]. They showed that whp, the largest induced tree in $G(n,p)$ is asymptotically of size $2\log_q(np)$ if p is a constant. Furthermore, they conjectured that the largest induced tree in the sparse regime (when $p = c/n$ for a large constant c) is of linear size. Frieze and Jackson [11], Kučera and Rödl [13], Łuczak and Palka [15], and de la Vega [4] independently proved this conjecture. Subsequently, for $p = c\ln n/n$, where ln denotes the natural logarithm, Palka and Ruciński [17] showed that the largest induced tree is of size between $\log_q(np)$ and $2\log_q(np)$. Finally, de la Vega [5] showed that for $p = c/n$, the largest induced tree has size asymptotically

*Received by the editors April 9, 2020; accepted for publication (in revised form) November 22, 2020; published electronically February 22, 2021.
https://doi.org/10.1137/20M1330609
Funding: The first and third authors were supported by Austrian Science Fund (FWF) grant I3747. The fourth author's research was supported in part by SNSF grant 200021_196965.
†Institute of Discrete Mathematics, Graz University of Technology, 8010 Graz, Austria (cooley@math.tugraz.at, kang@math.tugraz.at).
‡Department of Mathematics, ETH, 8092 Zürich, Switzerland (nemanja.draganic@math.ethz.ch, benjamin.sudakov@math.ethz.ch).

$2\log_q(np) \sim 2\frac{\ln c}{c}n$. Although de la Vega proves his result only for $p = c/n$, one can check that his ideas extend to any larger p as well.

Note that the fact that $G(n,p)$ contains large induced trees does not give much information on what these trees look like. Therefore, a natural question is whether a given *fixed* large tree is an induced subgraph of $G(n,p)$. The first steps in this direction were made by Frieze and Jackson [10] and Suen [19], who showed that the length of the longest induced path in $G(n,c/n)$ is linear for c large enough. Łuczak [14] improved on their results by proving that the length of the longest induced path in $G(n,c/n)$ is between $\log_q(np)$ and $2\log_q(np)$. For constant p, Ruciński [18] showed that the longest induced path in $G(n,p)$ is of length asymptotically $2\log_q(np)$, which was later extended to all $p \geq n^{-1/2}(\ln n)^2$ by Dutta and Subramanian in [7].

When p is small, it is natural to require some restriction on the maximum degree $\Delta(H)$ of H, simply because whp, $G(n,p)$ does not have any vertices of large degree. In particular, when $p = c/n$, a natural case to study is when H is a fixed tree of bounded degree, i.e., a tree with $\Delta(H) < d$ for some constant d. The second author [6] proved that for $n^{-1/2}(\ln n)^2 \leq p \leq 0.99$ and every fixed bounded degree tree T of size asymptotically $2\log_q(np)$, $G(n,p)$ contains T with high probability. On the other hand, in sparser random graphs very little is known about induced copies of general large bounded degree trees.

Another natural class of induced subgraphs to look for in $G(n,p)$ is induced matchings, which are in some sense an interpolation between independent sets and trees. For constant p, it has been shown by Clark [3] that whp, $G(n,p)$ contains induced matchings with $(2 \pm o(1))\log_q(np)$ vertices.

In this paper we establish the following result on induced matchings, stating that the largest induced matching in $G(n,p)$ contains roughly $2\log_q(np)$ vertices, which is an asymptotically optimal result.

THEOREM 1.1. *For all $\varepsilon_0 > 0$, there exists $C = C(\varepsilon_0) > 0$ such that whp, the largest induced matching in $G(n,p)$ contains $(1 \pm \varepsilon_0)\log_q(np)$ edges, where $q = \frac{1}{1-p}$, whenever $\frac{C}{n} \leq p \leq 0.99$.*

As described above, the size of largest independent sets in random graphs is well understood, and there are also several known results on the size of the largest induced tree for various regimes of p as well as for the largest induced matching when p is constant. However, for a *fixed* (i.e., previously specified) induced bounded degree tree in the sparse regime, we know very little. Our result for induced matchings is the first step in understanding this problem.

For $p > \frac{(\ln n)^2}{\sqrt{n}}$, the aforementioned results for independent sets, paths, bounded degree trees, and matchings can be proved using the second moment method. On the other hand, the vanilla second moment calculations break down roughly when $p \sim \frac{1}{\sqrt{n}}$, and for smaller p, all of the aforementioned problems become significantly harder.

Our proof relies on two main ingredients—the second moment method and Talagrand's inequality. Although the second moment method on its own is of little use in sparse regimes, in combination with strong concentration bounds (such as Talagrand's inequality) it yields a nice tool which can be very powerful, as was already demonstrated by Frieze in [9].

2. Large induced matching: Proof of Theorem 1.1. Note that for constant p, Theorem 1.1 is exactly the aforementioned result of Clark [3]. We will explicitly prove Theorem 1.1 for $p \leq \frac{1}{(\ln n)^3}$ and point out that for larger $p = o(1)$, it is enough to use just a standard second moment argument with Chebyshev's inequality, whose calculations are much simpler than those required in the sparser case. Full details can be found in Appendix B of the arXiv submission of this paper, arXiv:2004.03359.

Throughout the paper we will use the standard Landau notations $o(\cdot), O(\cdot), \Theta(\cdot), \Omega(\cdot), \omega(\cdot)$. When not otherwise explicitly stated, the asymptotics in this notation are with respect to n. We will also use this notation with asymptotics with respect to c, in which case we add c explicitly to the notation. For example, $f = o_c(g)$ means that $|f|/|g| \xrightarrow{c\to\infty} 0$. We will omit floors and ceilings when these do not significantly affect the argument.

Given a graph H, we denote by $\mathcal{M}(H)$ the size of (i.e., the number of edges in) the largest induced matching in H. We want to show that $\mathcal{M}(G(n,p)) \geq (1-\varepsilon_0)\log_q(np)$ with high probability. Note that if $p = o(1)$, then this is asymptotically equal to $(1-\varepsilon_0)\frac{\ln(np)}{p}$, so we will work with the latter expression to ease notation.

We use the notation $a \gg b$ to mean that for some implicit function $f : \mathbb{R} \to \mathbb{R}$, we have $a \geq f(b)$. We will not determine the function f that we require explicitly, although it could be deduced from a careful analysis of the calculations. For the rest of the paper, we fix the following parameters. Let $\varepsilon_0 > 0$, let $\varepsilon := \frac{\varepsilon_0}{3}$, and let $p = c/n$, where $c = c(n)$ is a function of n satisfying $\frac{n}{(\ln n)^3} \geq c \gg \varepsilon^{-1}$. Let

$$k := \frac{(1-\varepsilon)\ln c}{c} n. \tag{2.1}$$

Let $G \sim G(n,p)$, and let Y_k be the random variable which counts the number of induced matchings of size k in G. We will prove two lemmas which directly imply our theorem. The first lemma tells us that $\mathcal{M}(G)$ is well concentrated in the sense that it cannot have both upper and lower tails having large probability.

LEMMA 2.1.

$$\mathbb{P}\left(\mathcal{M}(G) \leq k - \varepsilon\frac{\ln c}{c}n\right) \cdot \mathbb{P}\Big(\mathcal{M}(G) \geq k\Big) \leq \exp\left(-\frac{2n}{c}\right).$$

We will prove Lemma 2.1 in section 3 using an application of Talagrand's inequality.

The second lemma gives a rather weak estimate on the probability that a large matching occurs but which, in combination with Lemma 2.1, is enough to show Theorem 1.1.

LEMMA 2.2.

$$\mathbb{P}\left(Y_k > 0\right) \geq \exp\left(-\frac{n}{c}\right).$$

The proof of Lemma 2.2 is based on the second moment method using the Paley–Zygmund inequality and appears in section 4.

Now we show how Theorem 1.1 follows from these two lemmas. We begin with a simple first moment calculation. Analogously to Y_k, for any $r \in \mathbb{N}$, let Y_r denote the number of induced matchings of size r in G.

CLAIM 1. *For any positive integer r, we have*

$$\mathbb{E}[Y_r] = \binom{n}{2r}\frac{(2r)!}{r!2^r}p^r(1-p)^{\binom{2r}{2}-r}.$$

Proof. There are $\binom{n}{2r}(2r-1)!! = \binom{n}{2r}\frac{(2r)!}{r!2^r}$ possible matchings M of size r. Furthermore, in order for M to form an induced matching in $G(n,p)$, all r edges must be present, and furthermore the remaining $\binom{2r}{2} - r$ pairs in $V(M)$ may not be edges of $G(n,p)$, which occurs with probability $p^r(1-p)^{\binom{2r}{2}-r}$. Combining these two terms gives the claim. ∎

Proof of Theorem 1.1. We first prove the upper bound using the first moment method. Setting

$$r = (1+\varepsilon_0)\log_q(np) = (1+\varepsilon_0)\frac{\ln(np)}{-\ln(1-p)} = \Theta(1)\frac{\ln(np)}{p}$$

(where the last equality follows since $p \le 0.99$), let us observe that Claim 1 gives

$$\begin{aligned}\mathbb{E}[Y_r] = \binom{n}{2r}\frac{(2r)!}{r!2^r}p^r(1-p)^{\binom{2r}{2}-r} &\le \frac{n^{2r}}{(2r)!}\frac{(2r)!}{(2r/e)^r}p^r(1-p)^{r(2r-2)} \\ &= \left(\Theta(1)\frac{n^2p(1-p)^{2r}}{r}\right)^r \\ &= \left(\Theta(1)\frac{n^2p(np)^{-2(1+\varepsilon_0)}}{(\ln(np))/p}\right)^r \\ &= \left(\Theta(1)\frac{(np)^{-2\varepsilon_0}}{\ln(np)}\right)^r.\end{aligned}$$

Now if $p = \Theta\left(n^{-1}\right)$, then $r \ge \log_q(np) = \frac{\ln(np)}{-\ln(1-p)} \to \infty$, and recalling that $p \ge \frac{C}{n}$ for some $C = C(\varepsilon_0)$ sufficiently large, we have

$$\mathbb{E}[Y_r] \le \left(\Theta(1)\frac{C^{-2\varepsilon_0}}{\ln(C)}\right)^r \le \left(\frac{1}{2}\right)^r = o(1).$$

On the other hand, if $p = \omega\left(n^{-1}\right)$, then we have

$$\mathbb{E}[Y_r] \le (o(1))^r = o(1).$$

In both cases an application of Markov's inequality shows that whp, $Y_r = 0$, as required.

To prove the lower bound, as mentioned previously we assume that $p < \frac{1}{(\ln n)^3}$.

Recalling that for $p = o(1)$ we have

$$\log_q(np) = \frac{\ln(np)}{\ln\left(\frac{1}{1-p}\right)} = (1+o(1))\frac{\ln(np)}{p},$$

the two lemmas together imply that

$$\begin{aligned}\mathbb{P}\left(\mathcal{M}(G) \le (1-\varepsilon_0)\log_q(np)\right) &\le \mathbb{P}\left(\mathcal{M}(G) \le \left(1-\frac{2\varepsilon_0}{3}\right)\frac{\ln c}{c}n\right) \\ &= \mathbb{P}\Big(\mathcal{M}(G) \le k - \varepsilon\frac{\ln c}{c}n\Big) \\ &\overset{\text{L.2.1}}{\le} \frac{\exp\left(-\frac{2n}{c}\right)}{\mathbb{P}\left(\mathcal{M}(G) \ge k\right)} = \frac{\exp\left(-\frac{2n}{c}\right)}{\mathbb{P}\left(Y_k > 0\right)} \overset{\text{L.2.2}}{\le} \exp\left(-\frac{n}{c}\right),\end{aligned}$$

which tends to zero as required. ∎

3. Concentration using Talagrand's inequality: Proof of Lemma 2.1. Talagrand's inequality is a useful tool to show that under certain conditions, a random variable is tightly concentrated. We will use it in the form which appears in [1].

DEFINITION 3.1. *Let* $\Omega = \prod_{i=1}^{n} \Omega_i$ *be a product of probability spaces such that* Ω *has the product measure. Let* $g : \Omega \to \mathbb{R}$ *and* $f : \mathbb{N} \to \mathbb{N}$ *be functions.*

- *We say that* g *is* Lipschitz *if* $|g(x) - g(y)| \leq 1$ *for every* $x, y \in \Omega$ *which differ in at most one coordinate.*
- *We say that* g *is* f-certifiable *if for any* $x \in \Omega$ *and* $m \in \mathbb{N}$ *such that* $g(x) \geq m$*, there exists a set of coordinates* $I \subset [n]$ *with* $|I| \leq f(m)$ *such that each* $y \in \Omega$ *which agrees with* x *on* I *also satisfies* $g(y) \geq m$*.*

THEOREM 3.2 (Talagrand). *Let* X *be a Lipschitz random variable on* Ω *which is* f*-certifiable. Then for all* $\lambda > 0$ *and* $b \in \mathbb{N}$*, it holds that*

$$\mathbb{P}\left(X < b - \lambda\sqrt{f(b)}\right) \cdot \mathbb{P}\left(X \geq b\right) \leq \exp\left(-\frac{\lambda^2}{4}\right).$$

In order to prove Lemma 2.1 we will regard $G(n, p)$ as a product of $n-1$ probability spaces Z_i, $i \in [n-1]$. Recall that $G(n, p)$ is a graph on vertex set $[n]$. Each Z_i picks uniformly at random a subset of $[i]$ of size $\mathrm{Bi}(i, p)$—these are the neighbors of vertex $i + 1$ within $[i]$. It is easy to see that this is equivalent to $G(n, p)$.

Proof of Lemma 2.1. Let $G \sim G(n, p)$. Then the random variable $\mathcal{M} = \mathcal{M}(G)$ is Lipschitz. Indeed, note that by changing a particular Z_i, one can change $\mathcal{M}$ only by at most 1 since if G' is obtained from G by changing some edges adjacent to vertex $i + 1$ and if M is a largest matching in G, then certainly M', which is obtained from M by deleting $i + 1$ and its partner if it lies in M, is a matching in G', and therefore $\mathcal{M}(G') \geq \mathcal{M}(G) - 1$. By symmetry, also $\mathcal{M}(G') \leq \mathcal{M}(G) + 1$.

Furthermore, if $\mathcal{M} \geq m$, then there exists a set $S \subset V(G)$ of $2m$ vertices which induces a matching of size m. By fixing Z_{i-1} for each $i \in S$ (where we interpret Z_0 as an empty random variable), changing other coordinates can only increase the largest induced matching in G. Therefore, $\mathcal{M}$ is f-certifiable with $f(m) = 2m$. This means that we can apply Theorem 3.2 with parameters $b = k$ and $\lambda = \varepsilon\frac{\ln c}{c\sqrt{2k}}n$, and observing that $\lambda\sqrt{f(b)} = \varepsilon\frac{\ln c}{c}n$ we obtain

$$\begin{aligned}\mathbb{P}\left(\mathcal{M} < k - \varepsilon\frac{\ln c}{c}n\right) \cdot P\left(\mathcal{M} \geq k\right) &\leq \exp\left(-\frac{\lambda^2}{4}\right)\\ &= \exp\left(-\frac{\varepsilon^2(\ln c)^2 n^2}{8c^2 k}\right)\\ &\overset{(2.1)}{=} \exp\left(-\frac{\varepsilon^2}{8(1-\varepsilon)}\frac{\ln c}{c}n\right) \leq \exp\left(-\frac{2n}{c}\right),\end{aligned}$$

which completes the proof. ∎

4. Second moment method: Proof of Lemma 2.2. Consider the family $\{M_i \mid i \in I\}$ of all sets of k unordered disjoint pairs of vertices in G, i.e., the family of possible matchings of size k. For $i \in I$, let X_i be the indicator random variable which indicates that the pairs in M_i form an induced matching in G. In particular, it holds that $Y_k = \sum_{i \in I} X_i$. The main difficulty is to prove the following.

LEMMA 4.1.

$$\frac{\sum_{i \in I} \mathbb{E}[X_i | X_1 = 1]}{\mathbb{E}[Y_k]} \leq \exp\left(\frac{n}{c}\right). \tag{4.1}$$

Proof of Lemma 2.2. We will use the well-known inequality

$$\mathbb{P}(Y_k > 0) \geq \frac{\mathbb{E}[Y_k]^2}{\mathbb{E}[Y_k^2]}, \tag{4.2}$$

which can be deduced, for example, as a special case of the Paley–Zygmund inequality. We now observe that

$$\begin{aligned}\mathbb{E}[Y_k^2] &= \sum_{i,j\in I} \mathbb{E}[X_i X_j] = \sum_{j\in I}\sum_{i\in I} \mathbb{E}[X_i|X_j = 1]\mathbb{E}[X_j]\\ &= \sum_{j\in I}\sum_{i\in I} \mathbb{E}[X_i|X_1 = 1]\mathbb{E}[X_1]\\ &= |I| \cdot \mathbb{E}[X_1]\sum_{i\in I} \mathbb{E}[X_i|X_1 = 1] = \mathbb{E}[Y_k]\sum_{i\in I} \mathbb{E}[X_i|X_1 = 1].\end{aligned}$$

Therefore, using Lemma 4.1, we obtain

$$\mathbb{P}(Y_k > 0) \overset{(4.2)}{\geq} \left(\frac{\mathbb{E}[Y_k^2]}{\mathbb{E}[Y_k]^2}\right)^{-1} = \left(\frac{\sum_{i\in I} E[X_i|X_1 = 1]}{\mathbb{E}[Y_k]}\right)^{-1} \overset{(4.1)}{\geq} \exp\left(-\frac{n}{c}\right)$$

as required. ◻

To prove Lemma 4.1, we will consider the numerator and denominator separately in the next subsection.

4.1. Conditional expectation. In this subsection we will give an explicit expression for the numerator in the left-hand side of (4.1); for the denominator, we will just substitute the expression from Claim 1. We begin with the following lemma, whose proof forms the main part of this subsection.

LEMMA 4.2. *Given $\ell, s \in \mathbb{N}_0$ with $\ell + s \leq k$, let us define $a_{\ell,s} = a_{\ell,s}(n, p, k, c, \varepsilon)$ by*

$$a_{\ell,s} := 2^{\ell+2s-k} \frac{k!}{\ell! s!((k-\ell-s)!)^2} \cdot \frac{(n-2k)!}{(n-4k+2\ell+s)!} p^k \left(\frac{k}{(1-\varepsilon)\ln c}\right)^{\ell} (1-p)^{\binom{2k}{2}-k-\left(\binom{2\ell+s}{2}-\ell\right)}.$$

Then

$$\sum_{i\in I} \mathbb{E}[X_i|X_1 = 1] = \sum_{\ell=0}^{k}\sum_{s=0}^{k-\ell} a_{\ell,s}.$$

We first determine which $i \in I$ give a nonzero contribution to $\sum_{i\in I} \mathbb{E}[X_i|X_1 = 1]$.

DEFINITION 4.3. *We say that M_i is* compatible *with M_1 if*
1. *M_i contains no pair $\{u, v\}$ whose vertices lie in different pairs of M_1;*
2. *M_1 contains no pair $\{u, v\}$ whose vertices lie in different pairs of M_i.*

As a consequence of this definition, we observe that we can classify the pairs of M_i into types according to their intersection with M_1. More precisely, for any $i \in I$, denote by $V(M_i)$ the set of vertices contained in some pair of M_i. Then we have the following.

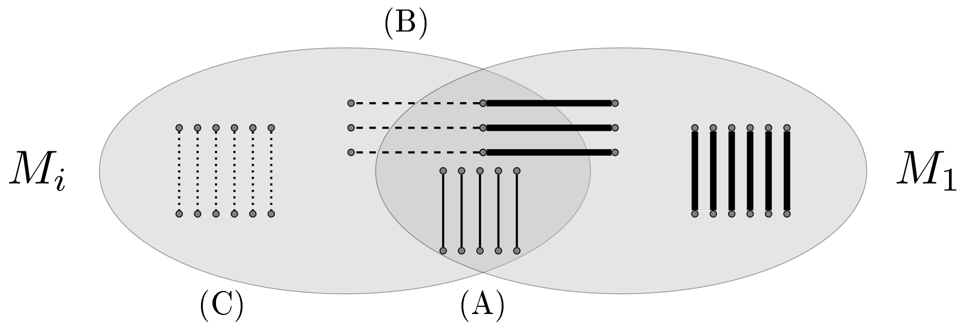

Fig. 1. *Type* (A), (B)*, and* (C) *pairs illustrated as plain, dashed, and dotted lines, respectively. Pairs which are in M_1 but not in M_i are thick.*

Remark 4.4. If M_i is compatible with M_1, then M_i only contains the following types of pairs (see Figure 1):

(A) pairs from M_1;

(B) pairs which contain one vertex from $V(M_1)$ and one vertex outside $V(M_1)$;

(C) pairs with no vertex in $V(M_1)$.

Furthermore, the only pairs of M_1 whose endpoints both lie in $V(M_i)$ are also pairs of M_i (or, equivalently, conditions (A), (B), and (C) also hold with M_1 and M_i switched).

Claim 2. *If M_i is not compatible with M_1, then $\mathbb{E}[X_i|X_1 = 1] = 0$.*

Proof. If M_i violates the first condition of compatibility with M_1, i.e., if M_i contains a pair of vertices $\{u, v\}$ from different pairs in M_1, then under the assumption that $X_1 = 1$, since M_1 is an *induced* matching, certainly $\{u, v\}$ is not an edge in G; hence, $X_i = 0$. By symmetry, if the second condition is violated, it is also not possible that $X_i = X_1 = 1$. ◻

We may therefore restrict our attention to matchings M_i that are compatible with M_1. Next we define an equivalence relation on I (or more precisely on the subset of those $i \in I$ such that M_i is compatible with M_1) such that for each i in the same equivalence class, the expression $\mathbb{E}[X_i|X_1 = 1]$ is the same.

Definition 4.5. *Define $I(\ell, s)$ to be the set of $i \in I$ such that M_i is compatible with M_1 and has exactly ℓ pairs of vertices which are of type* (A) *and s pairs of type* (B) *(see Remark* 4.4*).*

Thus, we have

$$\sum_{i\in I} \mathbb{E}[X_i|X_1 = 1] = \sum_{\ell,s} \sum_{i\in I(\ell,s)} \mathbb{E}[X_i|X_1 = 1]. \tag{4.3}$$

Now for fixed ℓ, s, we can handle the conditional expectation with the following claim.

Claim 3. *For $i \in I(\ell, s)$, we have*

$$\mathbb{E}[X_i|X_1 = 1] = p^k \left(\frac{k}{(1-\varepsilon)\ln c}\right)^{\ell} (1-p)^{\binom{2k}{2}-k-\left(\binom{2\ell+s}{2}-\ell\right)}. \tag{4.4}$$

Proof. Conditioning on the event that $X_1 = 1$, we observe that the ℓ pairs of type (A) are already automatically present as edges because they are pairs of M_1, but we require a further $k-\ell$ pairs to be present as edges, which occurs with probability $p^{k-\ell} = p^k(\frac{k}{(1-\varepsilon)\ln c})^\ell$. Furthermore, we require all of the $\binom{2k}{2}-k$ pairs which lie within $V(M_i)$ but are not in the matching M_i to be nonedges of $G(n,p)$; however, those pairs that lie inside $V(M_1)$, of which there are $\binom{2\ell+s}{2}-\ell$, are already guaranteed to be nonedges by the conditioning on $X_1 = 1$. Thus, the probability that all appropriate pairs are nonedges is $(1-p)^{\binom{2k}{2}-k-(\binom{2\ell+s}{2}-\ell)}$. Multiplying the two terms together, we obtain the statement of the claim. □

The sum over $I(\ell,s)$ in (4.3) is dealt with using the following result.

Claim 4. *Let* $\ell, s \in \mathbb{N}_0$.
(i) *If* $\ell+s>k$, *then* $|I(\ell,s)|=0$.
(ii) *If* $\ell+s\le k$, *then*

$$|I(\ell,s)| = 2^{\ell+2s-k}\frac{k!}{\ell!s!((k-\ell-s)!)^2}\cdot\frac{(n-2k)!}{(n-4k+2\ell+s)!}. \tag{4.5}$$

Proof. The first statement is clear since in order for i to lie in $I(\ell,s)$, the matching M_i must contain ℓ pairs of type (A) and s pairs of type (B) but k pairs in total.

For the second statement, observe that there are $\binom{k}{\ell}$ ways of choosing the ℓ pairs of type (A). We subsequently choose the s endpoints within M_1 of pairs of type (B), for which there are $\binom{k-\ell}{s}2^s$ possible choices. For the other endpoints of these s pairs, we have $\binom{n-2k}{s}$ choices for the vertices outside $V(M_1)$ and $s!$ ways of matching them with the s endpoints already chosen within $V(M_1)$. Finally, we have $\binom{n-2k-s}{2k-2\ell-2s}$ ways of choosing the vertices for the remaining $k-\ell-s$ pairs of type (C) (while avoiding $V(M_1)$ and the further s vertices chosen for pairs of type (B) and $(2k-2\ell-2s-1)!! = \frac{(2k-2\ell-2s)!}{(k-\ell-s)!2^{k-\ell-s}}$ ways of choosing a perfect matching on these vertices. Collecting all these terms gives

$$\begin{aligned}|I(\ell,s)| &= \binom{k}{\ell}\binom{k-\ell}{s}2^s\binom{n-2k}{s}s!\binom{n-2k-s}{2k-2\ell-2s}\frac{(2k-2\ell-2s)!}{(k-\ell-s)!2^{k-\ell-s}}\\ &= \frac{k!}{\ell!s!(k-\ell-s)!}2^s\frac{(n-2k)!}{(n-4k+2\ell+s)!(k-\ell-s)!2^{k-\ell-s}}\\ &= 2^{\ell+2s-k}\frac{k!}{\ell!s!((k-\ell-s)!)^2}\cdot\frac{(n-2k)!}{(n-4k+2\ell+s)!}\end{aligned}$$

as claimed. □

We can combine the two previous claims to prove Lemma 4.2.

Proof of Lemma 4.2. Observe that the statement of Lemma 4.2 follows directly by applying (4.3) and Claims 3 and 4. □

Now let us define $b_{\ell,s} = b_{\ell,s}(n,p,k,c,\varepsilon)$ by

$$\begin{aligned}b_{\ell,s} &:= a_{\ell,s}\frac{k!2^k}{(2k)!\binom{n}{2k}}p^{-k}(1-p)^{-\binom{2k}{2}+k}\\ &= 2^{\ell+2s}\frac{(k!)^2}{\ell!s!((k-\ell-s)!)^2}\cdot\frac{((n-2k)!)^2}{n!(n-4k+2\ell+s)!}\left(\frac{k}{(1-\varepsilon)\ln c}\right)^\ell(1-p)^{-\binom{2\ell+s}{2}+\ell}.\end{aligned} \tag{4.6}$$

Then we have the following immediate consequence of Lemma 4.2 and Claim 1 (applied with $r=k$).

COROLLARY 4.6.

$$\frac{\sum_{i\in I}\mathbb{E}[X_i|X_1=1]}{\mathbb{E}[Y_k]}=\sum_{\ell=0}^{k}\sum_{s=0}^{k-\ell}b_{\ell,s}.$$

4.2. Analyzing the summands. Given Corollary 4.6, we will aim to bound each of the summands $b_{\ell,s}$, which is the goal of this subsection.

LEMMA 4.7. *For any* $\ell,s\in\mathbb{N}_0$ *satisfying* $\ell+s\le k$*, we have*

$$b_{\ell,s}\le\exp\left(\frac{n}{2c}\right).$$

In the main argument we will use some approximations which are not well-defined if any one of s,ℓ or $k-\ell-s$ is 0, so we first deal with such terms by comparing them to others.

CLAIM 5. *For any* $\ell,s\in[k]_0$ *with* $\ell+s\le k$*, we have*

$$b_{\ell,s}\le n^9\max_{\substack{1\le i\le k-2\\1\le j\le k-i-1}}b_{i,j}.$$

To prove this claim, it suffices to compare terms on the "boundary," i.e., when $\ell=0$, when $s=0$, or when $\ell+s=k$, with other terms. For example, it is elementary to check that for $0\le\ell\le k-1$, we have $b_{\ell,0}\le n\cdot b_{\ell,1}$. In other words, we may "move" from $s=0$ to $s=1$ at a cost of a multiplicative factor n. By making at most three such moves, each at a cost of at most n^3, we can reach the interior of the region, i.e., $1\le\ell,s\le k-1$ and $\ell+s\le k-1$, from any point on the boundary. We omit the details, which are an elementary but tedious exercise in calculation. For the interested reader, these details can be found in Appendix A of the arXiv submission of this paper, arXiv:2004.03359.

In particular, Claim 5 allows us to restrict our attention to the case when $\ell,s,k-\ell-s\neq0$.

PROPOSITION 4.8. *If* $1\le\ell,s\in\mathbb{N}$ *and* $\ell+s\le k-1$*, then*

$$b_{\ell,s}\le\exp\left(\frac{n}{3c}\right).$$

Proof. We will write $P(n)$ for any term that is polynomial in n (i.e., there exists a polynomial Q such that $\frac{1}{Q(n)}\le P(n)\le Q(n)$ for sufficiently large n). Since $\ell,s,k-\ell-s\neq0$, we may apply Stirling's approximation to various terms in (4.6) and obtain

$$\begin{aligned}b_{\ell,s}&=2^{\ell+2s}\frac{P(k)\left(\frac{k}{e}\right)^{2k}}{\left(\frac{\ell}{e}\right)^{\ell}\left(\frac{s}{e}\right)^{s}\left(\frac{k-\ell-s}{e}\right)^{2(k-\ell-s)}}\cdot\frac{P(n)\left(\frac{n-2k}{e}\right)^{2(n-2k)}}{\left(\frac{n}{e}\right)^{n}\left(\frac{n-4k+2\ell+s}{e}\right)^{n-4k+2\ell+s}}\\&\quad\cdot\left(\frac{k}{(1-\varepsilon)\ln c}\right)^{\ell}(1-p)^{-\binom{2\ell+s}{2}+\ell}\\&=P(n)2^{\ell+2s}\frac{e^{\ell}k^{2k+\ell}}{\ell^{\ell}s^{s}(k-\ell-s)^{2(k-\ell-s)}}\cdot\frac{(n-2k)^{2(n-2k)}}{n^{n}(n-4k+2\ell+s)^{n-4k+2\ell+s}}\\&\quad\cdot\left(\frac{1}{(1-\varepsilon)\ln c}\right)^{\ell}(1-p)^{-\binom{2\ell+s}{2}+\ell}\end{aligned}$$

$$= P(n) \frac{\left(\frac{k}{\ell}\right)^{\ell} \left(\frac{k}{s}\right)^{s} \left(\frac{k}{k-\ell-s}\right)^{2(k-\ell-s)} \left(\frac{n-2k}{n}\right)^{n} \left(\frac{n-2k}{n-4k+2\ell+s}\right)^{n-4k+2\ell+s}}{\left(\frac{n-2k}{k}\right)^{2\ell+s}} 2^{2s}$$
$$\cdot \left(\frac{2e(1-p)}{(1-\varepsilon)\ln c}\right)^{\ell} (1-p)^{-\binom{2\ell+s}{2}}$$
$$= P(n) \frac{e^{\ell \ln(k/\ell)+s\ln(k/s)}(1+\frac{\ell+s}{k-\ell-s})^{2(k-\ell-s)}(1-\frac{2k}{n})^{n}(1+\frac{2k-2\ell-s}{n-4k+2\ell+s})^{n-4k+2\ell+s}}{(\frac{n-2k}{k})^{2\ell+s}(1-p)^{\binom{2\ell+s}{2}}} 2^{2s}$$
$$\left(\frac{2e(1-p)}{(1-\varepsilon)\ln c}\right)^{\ell}.$$

Let us observe that $\frac{2e(1-p)}{(1-\varepsilon)\ln c} \leq 1$ for sufficiently large c, so for an upper bound we may ignore the final term. For what remains, we will use the inequalities $e^{-x-x^2} \leq 1-x \leq e^{-x-x^2/2}$ and $1+x \leq e^{x-x^2/2+x^3/3} \leq e^x$, which hold for all positive $x < 1/2$, and indeed the inequality $1+x \leq e^x$ holds for any real number x. Thus, observing that $\ln(P(n)) = \Theta(\ln n)$, we obtain

$$\begin{aligned}
\ln(b_{\ell,s}) \leq \Theta(\ln n) &+ \ell \ln\left(\frac{k}{\ell}\right) + s\ln\left(\frac{k}{s}\right) + 2(\ell+s) - \left(2k + \frac{2k^2}{n}\right) \\
&+ (2k-2\ell-s) - \frac{(2k-2\ell-s)^2}{2(n-4k+2\ell+s)} + \frac{(2k-2\ell-s)^3}{3(n-4k+2\ell+s)^2} \\
&+ 2s\ln 2 - \ln\left(\frac{n-2k}{k}\right)(2\ell+s) - \left(-\frac{c}{n} - \frac{c^2}{n^2}\right)\frac{(2\ell+s)^2}{2} \\
\leq \Theta(\ln n) &+ \ell\ln\left(\frac{k}{\ell}\right) + s\ln\left(\frac{k}{s}\right) + s - \frac{2k^2}{n} - \frac{(2k-2\ell-s)^2}{2n} + \frac{3k^3}{n^2} \\
&+ 2s\ln 2 - \ln\left(\frac{n-2k}{k}\right)(2\ell+s) + \frac{c}{n}\left(1+\frac{\varepsilon}{2}\right)\frac{(2\ell+s)^2}{2}
\end{aligned}$$
$$= F + o_c\left(\frac{1}{c}\right)n, \tag{4.7}$$

where we define

(4.8)

$$\begin{aligned}
F &= F(n,k,c,\ell,s) \\
&:= \ell\ln\left(\frac{k}{\ell}\right) + s\ln\left(\frac{k}{s}\right) + s + 2s\ln 2 + \frac{c}{n}\left(1+\frac{\varepsilon}{2}\right)\frac{(2\ell+s)^2}{2} - \ln\left(\frac{n-2k}{k}\right)(2\ell+s).
\end{aligned}$$

Our aim is to bound F from above—we will have two cases depending on how large $2\ell+s$ is.

Case I: $2\ell + s = \omega_c\left(\frac{1}{\ln c}\right)k$.
In this case the last term will outweigh all the positive terms in expression (4.8), so F will be negative. Indeed,

$$\ln\left(\frac{n-2k}{k}\right)(2\ell+s) = (1-o_c(1))(\ln c)(2\ell+s),$$

and we first claim that this expression dominates the first two positive terms. To see this, set $g(x) := x\ln(k/x)$, and observe that $g'(x) = \ln(k/x) - 1$ and $g''(x) = -1/x <$

0. Therefore, g attains its maximum when $\ln(k/x) - 1 = 0$, i.e., when $x = k/e$, which implies that for all x, $g(x) \leq g(k/e) = k/e$. Therefore,

$$\ell \ln\left(\frac{k}{\ell}\right) + s\ln\left(\frac{k}{s}\right) \leq \frac{2k}{e} \leq o_c(\ln c)(2\ell + s).$$

Since $s + 2s\ln 2 \leq 3k = o_c(\ln c)(2\ell + s)$ is also comparatively small, the only term which remains is $\frac{c}{n}\left(1 + \frac{\varepsilon}{2}\right)\frac{(2\ell+s)^2}{2}$, and by recalling that $2\ell + s \leq 2k$, we get that this term is less than $\frac{kc}{n}\left(1 + \frac{\varepsilon}{2}\right)(2\ell + s) \leq \left(1 - \frac{\varepsilon}{2}\right)(\ln c)(2\ell + s)$. Therefore, F is dominated by $-\frac{\varepsilon}{2}(\ln c)(2\ell + s)$, which means that $F \leq 0$, so in this case we are done.

Case II: $2\ell + s = O_c\left(\frac{1}{\ln c}\right)k$.
Here we split F into two parts and separately prove that they are small:

$$F_1 := \ell\ln\left(\frac{k}{\ell}\right) + \frac{c}{n}\left(1 + \frac{\varepsilon}{2}\right)(2\ell^2 + 2\ell s) - \ln\left(\frac{n-2k}{k}\right)2\ell,$$
$$F_2 := s\ln\left(\frac{k}{s}\right) + s + 2s\ln 2 + \frac{cs^2}{2n}\left(1 + \frac{\varepsilon}{2}\right) - \ln\left(\frac{n-2k}{k}\right)s.$$

Upper bound on F_1: We will again use the fact that, by the arguments in Case I, $g(x) = x\ln(k/x)$ is increasing if $x \leq \frac{k}{c\ln c} \leq k/e$. We divide further into two subcases.

Subcase (a): $\ell \leq \frac{k}{c\ln c}$. Here we simply ignore the negative term for an upper bound and also observe that $2\ell^2 + 2\ell s \leq 2(2\ell + s)\ell \leq 2k\ell$, which gives

$$\begin{aligned} F_1 \leq \ell\ln\left(\frac{k}{\ell}\right) + \frac{c}{n}\left(1 + \frac{\varepsilon}{2}\right)2k\ell &\leq \frac{k}{c\ln c}\ln(c\ln c) + \frac{3k^2}{n\ln c} \\ &\leq k\left(\frac{2}{c} + \frac{3k}{n\ln c}\right) \\ &\leq \left(\frac{\ln c}{c}\right)n\cdot\frac{5}{c} \\ &= o_c\left(\frac{1}{c}\right)n. \end{aligned}$$

Subcase (b): $\ell > \frac{k}{c\ln c}$. In this case we bound the positive terms in F_1 by

$$\begin{aligned} \ell\ln\left(\frac{k}{\ell}\right) + \frac{c}{n}\left(1 + \frac{\varepsilon}{2}\right)(2\ell^2 + 2\ell s) &\leq \ell\ln(c\ln c) + \frac{c}{n}3\ell(2\ell + s) \\ &\leq \frac{4\ln c}{3}\ell + \frac{3ck}{n}\ell O_c\left(\frac{1}{\ln c}\right) \\ &\leq \frac{4\ln c}{3}\ell + (\ln c)\ell O_c\left(\frac{1}{\ln c}\right) \\ &\leq \frac{3}{2}\ell\ln c. \end{aligned}$$

Meanwhile, (the absolute value of) the negative term in F_1 is

$$\ln\left(\frac{n-2k}{k}\right)2\ell = \ln\left(\frac{c}{(1-\varepsilon)\ln c} - 2\right)2\ell \geq \frac{3}{2}(\ln c)\ell,$$

and so in total we have $F_1 \leq 0$.

Thus, in both subcases (a) and (b), we certainly have

$$F_1 \le o_c\left(\frac{1}{c}\right) n.$$

Upper bound on F_2: As before, we will consider two subcases.

Subcase (a): $s \le \frac{k}{c \ln c}$. Analogously as for F_1, we get that $F_2 \le o_c\left(\frac{1}{c}\right) n$. More precisely, again using the fact that $g(x) = x\ln(k/x)$ is increasing for $x \le \frac{k}{c\ln c}$, we have

$$\begin{aligned} F_2 &\le s\ln\left(\frac{k}{s}\right) + s + 2s\ln 2 + \frac{cs^2}{2n}\left(1+\frac{\varepsilon}{2}\right) \\ &\le \frac{k}{c\ln c}\left(\ln(c\ln c) + 1 + 2\ln 2 + \frac{k}{(\ln c)n}\right) \\ &\le \frac{n}{c^2}\left(2\ln c + \frac{1}{c}\right) \\ &= o_c\left(\frac{1}{c}\right) n. \end{aligned}$$

Subcase (b): $s > \frac{k}{c\ln c}$. First, we will estimate the last term in F_2 using the assumption of Case II, which implies that $s = O_c\left(\frac{1}{\ln c}\right) k$:

$$\begin{aligned} \ln\left(\frac{n-2k}{k}\right) s &= \ln\left(\frac{c}{(1-\varepsilon)\ln c}\left(1 - \frac{2(\ln c)(1-\varepsilon)}{c}\right)\right) s \\ &\ge \ln\left(\frac{c}{\ln c}\right) s + \ln\left(1 - \frac{2(\ln c)(1-\varepsilon)}{c}\right) s \\ &= \big(\ln c - \ln\ln c\big)s - O_c\left(\frac{\ln c}{c}\right) O_c\left(\frac{1}{\ln c}\right) k \\ &= \big(\ln c - \ln\ln c\big)s - o_c\left(\frac{1}{c}\right) n. \end{aligned} \tag{4.9}$$

Now set $\alpha := \frac{cs}{k\ln c}$, so that $s = \alpha\frac{\ln c}{c}k$, and we have $\frac{1}{(\ln c)^2} < \alpha \le O_c(\frac{c}{(\ln c)^2})$ since $\frac{k}{c\ln c} < s = O_c(\frac{k}{\ln c})$. This implies that

$$\begin{aligned} F_2 &= s\ln\left(\frac{c}{\alpha\ln c}\right) + s + 2s\ln 2 + \frac{cs^2}{2n}\left(1+\frac{\varepsilon}{2}\right) - \big(\ln c - \ln\ln c - \ln(1-\varepsilon)\big)s + o_c\left(\frac{1}{c}\right) n \\ &\le s\ln(1/\alpha) + s + 2s\ln 2 + \frac{cs^2}{n} + o_c\left(\frac{1}{c}\right) n. \end{aligned} \tag{4.10}$$

First, suppose $\alpha > c^{1/4}$. Using the assumption of Case II, we have $\frac{cs^2}{n} = O_c(1)s$, and so we obtain

$$F_2 \le s\left(-\frac{\ln c}{4} + O_c(1)\right) + o_c\left(\frac{1}{c}\right) n \le o_c\left(\frac{1}{c}\right) n.$$

On the other hand, if $\alpha < c^{1/4}$, then we have

$$\frac{cs^2}{n} = c\frac{(\alpha k(\ln c)/c)^2}{n} \le \frac{(\ln c)^2 k^2}{n\sqrt{c}} \le \frac{(\ln c)^4 n}{c^{5/2}} = o_c\left(\frac{1}{c}\right) n,$$

and substituting this into (4.10), we obtain

$$F_2 \leq s\left(\ln(1/\alpha) + 1 + 2\ln 2\right) + o_c\left(\frac{1}{c}\right) n$$
$$= \frac{k \ln c}{c}\left(\alpha \ln(1/\alpha) + \alpha + 2\alpha \ln 2\right) + o_c\left(\frac{1}{c}\right) n,$$

and this function is maximized for $\alpha = 4$. Thus, we have

$$F_2 \leq \frac{k \ln c}{c} \cdot 4 + o_c\left(\frac{1}{c}\right) n = o_c\left(\frac{1}{c}\right) n.$$

Thus, in all cases we have $F_1, F_2 \leq o_c\left(\frac{1}{c}\right) n$ and therefore also $F = F_1 + F_2 = o_c\left(\frac{1}{c}\right) n$. Substituting this into (4.7), we deduce that

$$b_{\ell,s} \leq \exp\left(o_c\left(\frac{1}{c}\right) n\right) \leq \exp\left(\frac{n}{3c}\right)$$

as claimed. □

We have now collected all the auxiliary results we need, and we show that these imply the various previously stated results. First, we show that Proposition 4.8 implies Lemma 4.7.

Proof of Lemma 4.7. Clearly, Lemma 4.7 is implied directly by Proposition 4.8 for any $\ell, s \geq 1$ such that $\ell + s \leq k - 1$. The remaining terms can be dealt with by combining Proposition 4.8 and Claim 5, together with the observation that

$$n^9 \exp\left(\frac{n}{3c}\right) \leq \exp\left(\frac{n}{2c}\right)$$

because $c < \frac{n}{(\ln n)^3}$. □

We can now combine previous results to prove Lemma 4.1.

Proof of Lemma 4.1. Applying Corollary 4.6 and Lemma 4.7, we have

$$\frac{\sum_{i\in I} \mathbb{E}[X_i | X_1 = 1]}{\mathbb{E}[Y_k]} = \sum_{\ell=0}^{k} \sum_{s=0}^{k-\ell} b_{\ell,s} \leq \sum_{\ell=0}^{k} \sum_{s=0}^{k-\ell} \exp\left(\frac{n}{2c}\right) \leq n^2 \exp\left(\frac{n}{2c}\right) \leq \exp\left(\frac{n}{c}\right)$$

as claimed. □

5. Concluding remarks. In this paper we asymptotically determined the size of the largest induced matching of $G(n, p)$. Using a similar approach, one could probably show the existence of large forests with components of bounded size, but the calculations get messy very quickly. When instead of a large induced matching we look for a fixed large induced bounded degree tree, much less is known. Nevertheless, based on the evidence listed in the introduction, we propose the following conjecture.

CONJECTURE 5.1. *Let $\Delta \geq 2$, let $\varepsilon > 0$, and let $p = p(n)$ be a function such that $\frac{C}{n} < p < 0.99$, where $C = C(\Delta, \varepsilon)$ is sufficiently large. Let T be a tree with $(2-\varepsilon)\log_q(np)$ vertices and maximum degree Δ, where $q = 1/(1-p)$. Then with high probability, $G(n, p)$ contains T as an induced subgraph.*

If true, this conjecture would be asymptotically best possible. As mentioned in the introduction, the result was already proved for $p \geq n^{-1/2}(\ln n)^2$ in [6].

REFERENCES

[1] N. Alon and J. H. Spencer, *The Probabilistic Method*, John Wiley & Sons, New York, 2004.
[2] B. Bollobás and P. Erdős, *Cliques in Random Graphs*, in Math. Proc. Cambridge Philos. Soc. 80, Cambridge University Press, Cambridge, 1976, pp. 419–427.
[3] L. Clark, *The strong matching number of a random graph*, Australas. J. Combin., 24 (2001), pp. 47–58.
[4] W. F. de la Vega, *Induced trees in sparse random graphs*, Graphs Combin., 2 (1986), pp. 227–231.
[5] W. F. de la Vega, *The largest induced tree in a sparse random graph*, Random Structures Algorithms, 9 (1996), pp. 93–97.
[6] N. Draganić, *Large Induced Trees in Dense Random Graphs*, preprint, arXiv:2004.02800, 2020.
[7] K. Dutta and C. Subramanian, *On induced paths, holes and trees in random graphs*, in 2018 Proceedings of the Fifteenth Workshop on Analytic Algorithmics and Combinatorics (ANALCO), SIAM, Philadelphia, 2018, pp. 168–177.
[8] P. Erdős and Z. Palka, *Trees in random graphs*, Discrete Math., 46 (1983), pp. 145–150.
[9] A. M. Frieze, *On the independence number of random graphs*, Discrete Math., 81 (1990), pp. 171–175.
[10] A. M. Frieze and B. Jackson, *Large holes in sparse random graphs*, Combinatorica, 7 (1987), pp. 265–274.
[11] A. M. Frieze and B. Jackson, *Large induced trees in sparse random graphs*, J. Combin. Theory Ser. B, 42 (1987), pp. 181–195.
[12] S. Janson, T. Łuczak, and A. Ruciński, *Random Graphs*, Vol. 45, John Wiley & Sons, New York, 2011.
[13] L. Kučera and V. Rödl, *Large trees in random graphs*, Comment. Math. Univ. Carolin., 28 (1987), pp. 7–14.
[14] T. Łuczak, *The size of the largest hole in a random graph*, Discrete Math., 112 (1993), pp. 151–163.
[15] T. Łuczak and Z. Palka, *Maximal induced trees in sparse random graphs*, Discrete Math., 72 (1988), pp. 257–265.
[16] D. W. Matula, *The Largest Clique Size in a Random Graph*, Technical report CS 7608, Southern Methodist University, Dallas, TX, 1976.
[17] Z. Palka and A. Ruciński, *On the order of the largest induced tree in a random graph*, Discrete Appl. Math., 15 (1986), pp. 75–83.
[18] A. Ruciński, *Induced Subgraphs in a Random Graph*, in North-Holland Mathematics Studies 144, Elsevier, Ansterdam, 1987, pp. 275–296.
[19] W. S. Suen, *On large induced trees and long induced paths in sparse random graphs*, J. Combin. Theory, Ser. B, 56 (1992), pp. 250–262.