Random maximal H-free graphs

Deryk Osthus* and Anusch Taraz
Institut für Informatik, Humboldt-Universität zu Berlin
Unter den Linden 6, 10099 Berlin, Germany
E-mail address: {osthus,taraz}@informatik.hu-berlin.de

Abstract

Given a graph H, a random maximal H-free graph is constructed by the following random greedy process. First assign to each edge of the complete graph on n vertices a birthtime which is uniformly distributed in $[0,1]$. At time $p = 0$ start with the empty graph and increase p gradually. Each time a new edge is born, it is included in the graph if this does not create a copy of H. The question is then how many edges such a graph will have when $p = 1$. Here we give asymptotically almost sure bounds on the number of edges if H is a strictly 2-balanced graph, which includes the case when H is a complete graph or a cycle. Furthermore, we prove the existence of graphs with girth greater than ℓ and chromatic number $n^{1/(\ell-1)+o(1)}$, which improves on previous bounds for $\ell > 3$.

1 Introduction and Results

Given a fixed graph H, consider the following procedure which randomly constructs a maximal H-free graph. To each edge of the complete graph K_n on n vertices independently assign a birthtime which is uniformly distributed in $[0,1]$. For $p = 0$ start with the empty graph on n vertices. Now increase p gradually. Each time a new edge is born, add it to the existing graph if this does not create a copy of H. Edges with equal birthtime (which occur with probability zero) are considered in arbitrary order. Denote the graph at time p by $M_{n,p}(H)$ and denote the final graph $M_{n,1}(H)$ by $M_n(H)$.

The first result about the asymptotic structure of $M_n(H)$ is due to Ruciński and Wormald [14]. They proved that if H is a star with $d+1$ edges then with probability tending to one, $M_n(H)$ is in fact an *extremal* H-free graph, i.e., it contains at most one vertex of degree $d-1$, and all other vertices will have degree d. The case when H is a triangle was investigated by Erdős, Suen and Winkler [6] and by Spencer [16]. Here it turns out that $M_n(K_3)$ is *not* extremal. Indeed, with probability tending to one $M_n(K_3)$ has close to $n^{3/2}$ edges, whereas the balanced complete bipartite graph has $\lfloor n^2/4 \rfloor$ edges. Shortly afterwards, Kim [11] made a very detailed study of a similar process which enabled him to obtain his celebrated lower bound on the Ramsey number $R(3,t)$.

To be able to state our results, we first recall the following standard definitions. Given a graph G, $e(G)$ denotes the number of edges in G and $v(G)$

*Graduate school "Algorithmische Diskrete Mathematik", supported by Deutsche Forschungsgemeinschaft grant GRK/3-98

denotes the number of vertices. We say that G is *strictly 2-balanced* if $v(G) \geq 3$ and $e(G) \geq 3$ and if for all proper subgraphs G' of G with $v(G') \geq 3$ we have

$$\frac{e(G)-1}{v(G)-2} > \frac{e(G')-1}{v(G')-2}.$$

Furthermore we say that a graph G is *balanced* if $e(G)/v(G) \geq e(G')/v(G')$ for all subgraphs G' of G. We remark that strictly 2-balanced graphs are balanced, and it is easily seen that cycles, complete graphs, and the complete r-partite graph $K_{t,\dots,t}$ are strictly 2-balanced, whereas trees and disconnected graphs are not. Another example of a strictly 2-balanced graph is the d-dimensional cube (for $d \geq 2$). This fact may easily be derived from a lemma proven in Chung et al. [4], which states that every subgraph of the d-dimensional cube with v' vertices has at most $(v' \log_2 v')/2$ edges.

We write *asymptotically almost surely* or simply *a.a.s.* to mean "with probability tending to one as n tends to infinity". Throughout, H denotes the forbidden graph and if there is no danger of confusion, we write $e = e(H)$ and $v = v(H)$. We write e for the Euler number. All logarithms are base e unless specified otherwise.

The following lower bound on the number of edges in $M_n(H)$ is easily proved.

Theorem 1 *Suppose that H is balanced. Then there exists a constant $c = c(H)$ so that* a.a.s. *$M_n(H)$ has average degree at least*

$$cn^{1-\frac{v-2}{e-1}}.$$

Our main result is the following upper bound on the number of edges in $M_n(H)$ when H is strictly 2-balanced. For any graph G, let $\Delta(G)$ denote the maximum degree of G.

Theorem 2 *Suppose that H is strictly 2-balanced. Then there exists a constant $C = C(H)$ so that* a.a.s. *$M_n(H)$ has maximum degree at most*

$$Cn^{1-\frac{v-2}{e-1}} (\log n)^{1/(\Delta(H)-1)}.$$

The case when H is a triangle in Theorems 1 and 2 is already due (with slightly better bounds) to Erdős, Suen and Winkler [6] and Spencer [16]. The cases when H is a K_4 or a C_4 were proven independently by Bollobás and Riordan [3].

If we forbid for example the complete graph K_s on s vertices, Theorem 2 implies that *a.a.s.* the average degree $d(M_n(H))$ of $M_n(H)$ will satisfy

$$c\, n^{\frac{s-1}{s+1}} \leq d(M_n(H)) \leq C\, n^{\frac{s-1}{s+1}} (\log n)^{1/(s-2)}$$

for some constants c and C depending on s. Up to the logarithmic factor, this solves a problem of Erdős, Suen, and Winkler (see e.g. [5]). Furthermore, the proofs of Theorems 1 and 2 easily generalize to forbidding a bounded number of strictly 2-balanced graphs. If the family of forbidden graphs is $\mathcal{H}$, we define the random maximal $\mathcal{H}$-free graph $M_n(\mathcal{H})$ analogously to $M_n(H)$. We state the following theorem separately, as we shall give only a sketch of its proof.

Theorem 3 *Given $s \in \mathbb{N}$ and a family of strictly 2-balanced graphs $\mathcal{H} = \{H_0, H_1, \dots, H_{s-1}\}$ which is ordered so that*

$$\frac{e(H_0)-1}{v(H_0)-2} \leq \frac{e(H_1)-1}{v(H_1)-2} \leq \cdots \leq \frac{e(H_{s-1})-1}{v(H_{s-1})-2},$$

then a.a.s. *$M_n(\mathcal{H})$ has average degree at least*

$$cn^{1-\frac{v(H_0)-2}{e(H_0)-1}}$$

and maximum degree at most

$$Cn^{1-\frac{v(H_0)-2}{e(H_0)-1}}(\log n)^{1/(\Delta(H_0)-1)}$$

for some constants c and C depending on $\mathcal{H}$.

As a special case this answers (up to the factor of $\log n$ again) another problem listed in [5], namely:

Corollary 4 *Let $\mathcal{H}$ be the set of all cycles of length up to ℓ. Then* a.a.s. *$M_n(\mathcal{H})$ has average degree at least $cn^{1/(\ell-1)}$ and maximum degree at most $Cn^{1/(\ell-1)}\log n$ for some constants c and C depending on ℓ.*

Note that if one interprets the birthtimes of the edges as edge weights (which are uniformly and independently distributed in $[0,1]$), then the random greedy algorithm which forbids the set of *all* cycles produces a minimum spanning tree of the complete graph K_n. The asymptotically almost sure weight of this tree was determined by Frieze [8], and other properties of this tree were considered by Aldous [1].

Also, it is worth pointing out that Erdős, Suen and Winkler [6] proved that if one forbids the set of *all odd* cycles, then the graph obtained from the corresponding random greedy process *a.a.s.* has $\Theta(n^2)$ edges. On the other hand, Theorem 3 implies that if we forbid only all odd cycles up to length ℓ (where ℓ is fixed and odd) the resulting graph *a.a.s.* has far fewer edges – namely at most $Cn^{1+1/(\ell-1)}\log n$, for some constant C.

The common idea of the proofs is that as long as p is not too large, a random graph $G_{n,p}$ with edge probability p does not contain too many copies of H and thus $G_{n,p}$ and $M_{n,p}(H)$ will have similar properties. Beyond this critical point, the growth of $M_{n,p}(H)$ slows down very quickly. Quite naturally, when H is strictly 2-balanced the critical value of p turns out to be the point at which an edge in $G_{n,p}$ has a small but constant probability of being contained in a copy of H. We will obtain our results by studying the random graph obtained from $G_{n,p}$ by deleting all edges which lie in a copy of H, where p is suitably chosen. In fact, it is easily seen that the proofs yield error bounds which decay like $O(n^{-c})$ for any fixed constant c. Also, since birthtimes are distinct with probability one, we also obtain corresponding results for the processes where instead of random birthtimes we are given a random permutation of the edge set of the K_n. The proofs build upon ideas introduced in Spencer [15] and Krivelevich [12]. As a simple application of the methods used in [12], we also obtain the following result.

Theorem 5 *Given $\ell \in \mathbb{N}$ with $\ell \geq 3$, there is a graph G on n vertices with girth greater than ℓ and chromatic number*

$$\chi(G) \geq cn^{1/(\ell-1)}/\log n,$$

where c depends only on ℓ.

For $\ell = 3$, this is improved by [11], but for $\ell > 3$, the exponent is an improvement over the best previous exponent of $1/\ell$. The latter bound is contained in the original proof of the existence of graphs with high girth and high chromatic number due to Erdős (see, e.g., Alon and Spencer [2]). Note that Kostochka (see e.g. Jensen and Toft [10]) proved that for fixed $\ell \geq 3$, the chromatic number of a graph on n vertices and girth greater than ℓ is $O(n^{2/(\ell-1)})$.

We have made no attempt to optimize the constants involved in our results, as we believe that none of them give the correct order of magnitude. Indeed, Kim's proof of the lower bound on $R(3,t)$ in [11] shows that the average degree of the final random maximal triangle-free graph in a similar process is *a.a.s.* of order $\sqrt{n \log n}$. The result of Kim is obtained through an extremely sophisticated application of the so-called "nibble method" introduced by Rödl. It is likely that these techniques could be applied to improve our results, but in view of the complexity of Kim's proof, closing the gap between the upper and lower bounds even for triangles seems to be a formidable task indeed.

Another related result is that in [13] it is proven that for some (explicit) constant c and any odd ℓ, asymptotically, almost all graphs not containing an odd cycle C_ℓ and whose average degree is $(c'n \log n)^{1/(\ell-1)}$ are bipartite if $c' > c$, while this is not the case if $c' < c$ (where c' is a constant). We believe that in fact, the average degree of $M_n(C_\ell)$ may be *a.a.s.* of order $(n \log n)^{1/(\ell-1)}$.

This paper is organized as follows. In Section 2, we introduce some large deviation inequalities which we shall need in the proofs. In Section 3, we prove Theorem 5. In Section 4, we prove Theorem 1. In Section 5, we illustrate the ideas and techniques involved in the proof of Theorem 2 by proving the special case when the forbidden graph is a K_4. In Section 6, we then prove Theorem 2. In Section 7, we show how the proof of Theorem 2 may be modified to prove Theorem 3. In the final section, we discuss the open problem of forbidding an arbitrary graph H.

2 Tools and Notation

Let $G_{n,p}$ be the graph which has as its edges all those edges whose birthtime is at most p and let $G_{n,p}(H)$ be the subgraph of $G_{n,p}$ that is obtained by deleting *all* edges of $G_{n,p}$ that lie in a copy of H. We emphasize the immediate but important fact that the edge set of $G_{n,p}(H)$ is contained in that of $M_{n,p}(H)$. Thus

$$G_{n,p}(H) \subseteq M_{n,p}(H) \subseteq G_{n,p}. \tag{1}$$

For a graph G, the vertex set of G is denoted by $V(G)$. Furthermore we shall omit floors and ceilings throughout, as these will not affect the proof. Also, we shall always assume that n is sufficiently large for our estimates to hold.

We shall make use of two bounds on the number of disjoint subgraphs of a random graph. A slightly more general form of these inequalities, together with their proofs, may be found in Chapter 8 of [2].

Let $\mathcal{S}$ be a set of subsets of the edge set of the complete graph K_n on n vertices and for $\alpha \in \mathcal{S}$ let I_α be the corresponding indicator variable which equals one if and only if the edges of α are contained in the random graph $G_{n,p}$. Let

$$X = \sum_{\alpha \in \mathcal{S}} I_\alpha$$

and $\mu = \mathbb{E}[X]$. For α, $\alpha' \in \mathcal{S}$ with $\alpha \neq \alpha'$ write $\alpha \sim \alpha'$ if they have at least one edge in common and define

$$\Delta = \sum_{\alpha \sim \alpha'} \mathbb{E}[I_\alpha I_{\alpha'}],$$

where the summation is over all ordered pairs $\alpha \sim \alpha'$ in $\mathcal{S}$.

Let Y denote the size of a largest edge-disjoint family of elements of $\mathcal{S}$ in $G_{n,p}$. Erdős and Tetali [7] proved the following simple but very useful upper tail bound on Y. (It is also proved in Chapter 8 of [2] and is a special case of Lemma 2.46 in [9].) For every $k \in \mathbb{N}$,

$$\mathbb{P}[Y \geq k] \leq \left(\frac{\mathrm{e}\mu}{k}\right)^k. \tag{2}$$

We will also need a lower tail bound on Y. Let $\eta = \max_{\alpha \in \mathcal{S}} \mathbb{E}[I_\alpha]$ and

$$\nu = \max_{\alpha \in \mathcal{S}} \sum_{\alpha' \sim \alpha} \mathbb{E}[I_{\alpha'}].$$

Then Lemma 4.2 in Chapter 8 of [2] states that for any $s \in \mathbb{N}$,

$$\mathbb{P}[Y = s] \leq \frac{\mu^s}{s!} \exp\left\{-\mu + s\nu + \frac{\Delta}{2(1-\eta)}\right\}. \tag{3}$$

Thus for every $\varepsilon > 0$,

$$\begin{aligned} \mathbb{P}[Y \leq (1-\varepsilon)\mu] &\leq \exp\left\{\frac{\Delta}{2(1-\eta)}\right\} \sum_{s \leq (1-\varepsilon)\mu} \frac{\mu^s}{s!}\mathrm{e}^{-\mu+s\nu} \\ &\leq \exp\left\{\frac{\Delta}{2(1-\eta)} + (1-\varepsilon)\mu\nu\right\} \sum_{s \leq (1-\varepsilon)\mu} \frac{\mu^s}{s!}\mathrm{e}^{-\mu} \\ &= \exp\left\{\frac{\Delta}{2(1-\eta)} + (1-\varepsilon)\mu\nu\right\} \mathbb{P}[P \leq (1-\varepsilon)\mu], \end{aligned}$$

where P is a Poisson random variable with mean μ. But Theorem A.15 in [2] states that

$$\mathbb{P}[P \leq (1-\varepsilon)\mu] \leq \mathrm{e}^{-\varepsilon^2\mu/2}$$

which implies that

$$\mathbb{P}[Y \leq (1-\varepsilon)\mu] \leq \exp\left\{(1-\varepsilon)\mu\nu + \frac{\Delta}{2(1-\eta)} - \varepsilon^2\mu/2\right\}. \tag{4}$$

The above derivation of (4) from (3) is taken from [12], we have included it here for completeness.

Alternatively, we could have applied the following inequality, which follows from Talagrand's inequality using the same argument as in the proof of inequality (2.43) in Janson, Łuczak, and Ruciński [9]. For $t \geq 0$ and some constant γ depending only on the maximum number of edges of an element in $\mathcal{S}$, we have

$$\mathbb{P}[|Y - \mathbb{E}[Y]| \geq t] \leq 4\mathrm{e}^{-\gamma t^2/(\mathbb{E}[Y]+t)}.$$

3 Graphs with large girth

The idea of the proof is to show that deleting some maximal collection of edge-disjoint cycles from $G_{n,p}$, where p is suitably chosen, yields a graph with no large independent sets. An analogous idea was used by Krivelevich [12] to give lower bounds on Ramsey numbers.

Proof of Theorem 5. Let $p = c_0 n^{-\beta}$, where $\beta = (\ell-2)/(\ell-1)$. Let $G'_{n,p}$ be the graph obtained from $G_{n,p}$ by deleting some maximal collection of edge-disjoint cycles of length at most ℓ. Thus $G'_{n,p}$ has girth greater than ℓ. The theorem will follow if we can show that *a.a.s.* $G'_{n,p}$ has no independent set T of size t, where $t = |T| = Cn^{\beta}\log n$, where C is some sufficiently large constant depending on ℓ (see below). So for given T, let Y denote the number of edges in $G_{n,p}$ with both endpoints in T. Note that Y is binomially distributed with mean $\mu = \mathbb{E}[Y] = \binom{t}{2}p$. Hence a simple Chernoff tail estimate (e.g. Theorem A.13 in [2] with $a = \mathbb{E}[Y]/2$ or Theorem 2.1 in [9] with $t = \mathbb{E}[Y]/2$) yields

$$\mathbb{P}[Y \leq \mu/2] \leq \mathrm{e}^{-\mu/8}.$$

Let X' be the number of cycles in $G_{n,p}$ of length at most ℓ with at least one edge in T and let Y' be the maximum number of edge-disjoint cycles in $G_{n,p}$ of length at most ℓ with at least one edge in T. Then the number of edges of $G'_{n,p}$ which lie in T is at least $Y - \ell Y'$, so we are done if we can find a sufficiently sharp upper tail bound on Y'. If we choose the constant c_0 sufficiently small, then whatever the value of C we have that

$$\mathbb{E}[X'] \leq \sum_{j=3}^{\ell} t^2 n^{j-2} p^j \leq \ell\, t^2 n^{\ell-2} p^{\ell} \leq \mu/(\mathrm{e}\, 4\ell\, \mathrm{e}^{\ell/2}).$$

Now denote by $\mathcal{S}$ the set of all cycles of length at most ℓ with at least one edge in T and apply inequality (2) to see that

$$\mathbb{P}[Y' \geq \mu/(4\ell)] \leq \left(\frac{\mathrm{e}\, 4\ell\, \mathbb{E}[X']}{\mu}\right)^{\mu/(4\ell)} \leq \mathrm{e}^{-\mu/8}.$$

Putting the above together, the probability that $G'_{n,p}$ has an independent set of size t is at most

$$\binom{n}{t} 2\mathrm{e}^{-\mu/8} \leq 2\mathrm{e}^{t\log n - \mu/8} = 2\mathrm{e}^{t(\log n - (t-1)p/16)},$$

which tends to zero provided that C is sufficiently large. $\square$

4 Theorem 1 – the lower bound

The lower bound is an immediate consequence of the following well-known result of Erdős and Rényi (see Theorem 4.2 in Chapter 4 of [2] or Theorem 3.4 and Remarks 3.7 and 3.8 in [9]) which we state here in a form convenient to us.

Theorem 6 (Erdős and Rényi) *Let H be a balanced graph and let X denote the number of copies of H in $G_{n,p}$. If $pn^{v(H)/e(H)} \to \infty$, then* a.a.s. $X \leq 2\mathbb{E}[X]$.

Proof of Theorem 1. Let $p = c_0 n^{-(v-2)/(e-1)}$, where c_0 satisfies $c_0^{e-1}\, e = 1/8$, which is certainly large enough to be able to apply Theorem 6. Clearly, $e(G_{n,p}) - eX$ is a lower bound for the number of edges in $G_{n,p}(H)$ and thus also in $M_{n,p}(H)$ or $M_n(H)$. But, crudely,

$$e\,\mathbb{E}[X] \leq e\, n^v p^e \leq e\, n^2 p\, c_0^{e-1} = pn^2/8,$$

and the result follows. □

5 Forbidding a K_4 – the upper bound

To illustrate the ideas and techniques involved in the proof of Theorem 2, we now give a detailed proof of Theorem 2 for the case when the forbidden graph H is K_4. Let $p = n^{-2/5}/4$ and let

$$t = 10^4 n^{3/5} \sqrt{\log n}. \tag{5}$$

Suppose that we have a vertex y in the final random greedy graph $M_n(K_4)$ whose degree is at least t. Consider any set T of vertices which are adjacent to y in $M_n(K_4)$. Since $M_n(K_4)$ contains no copy of K_4 and since $M_{n,p}(K_4) \subseteq M_n(K_4)$, the graph $M_{n,p}(K_4)$ constructed by the random greedy process up to time p contains no copy of a K_4, the subgraph of $M_{n,p}(K_4)$ induced by T is triangle-free. By (1), it follows that the subgraph of $G_{n,p}(K_4)$ induced by T is also triangle-free. Thus it suffices to prove that *a.a.s.*, for all $T \subseteq V(K_n)$ of size t, $G_{n,p}(K_4)[T]$ contains a triangle – since this would imply that *a.a.s.*, $M_n(K_4)$ has maximum degree less than t, as stated in Theorem 2.

Given an edge f in K_n and $k \in \mathbb{N}$ we say that a sequence $H_1, \ldots, H_k$ of copies of K_4 forms a (k, f)*-cluster* in K_n if they all contain f and have the property that for all i with $1 \leq i \leq k$, H_i contains an edge which is not contained in any of the other H_j with $j < i$. Let $\mathcal{A}$ denote the event that $G_{n,p}$ does not contain a $(\log n, f)$-cluster for any edge f. Below, we will prove that $\mathbb{P}[\mathcal{A}] \to 1$. Also, we will prove that for any fixed $T \subseteq V(K_n)$ with $|T| = t$, we have

$$\mathbb{P}[\{G_{n,p}(K_4)[T] \text{ is } K_3\text{-free}\} \cap \mathcal{A}] \leq n^{-2t}. \tag{6}$$

These two results imply that the probability that there exists a set of t vertices which is triangle-free in $G_{n,p}(K_4)$ is at most

$$\mathbb{P}[\mathcal{A}^c] + \binom{n}{t}\mathbb{P}[\{G_{n,p}(K_4)[T] \text{ is } K_3\text{-free}\} \cap \mathcal{A}] \leq o(1) + n^t\, n^{-2t} = o(1),$$

which proves the theorem.

It remains to prove that $\mathbb{P}[\mathcal{A}] \to 1$ and (6). First we consider $\mathbb{P}[\mathcal{A}]$. For a fixed edge f in K_n and $k \in \mathbb{N}$ with $k \le \log n$, let $Z_{k,f}$ denote the number of (k,f)-clusters in $G_{n,p}$. Clearly, a $(1,f)$-cluster is a copy of a K_4 containing f, and thus

$$\mathbb{E}[Z_{1,f}] \le n^2 p^6 = 4^{-5} p \le \mathrm{e}^{-4}. \tag{7}$$

Note that every $(k+1,f)$-cluster can be viewed as the union of a (k,f)-cluster $C_{k,f}$ and a $(1,f)$-cluster which is not contained in $C_{k,f}$. Fix some (k,f)-cluster $C_{k,f}$ and let $Z_{1,f,C_{k,f}}$ denote the number of those $(1,f)$-clusters in $G_{n,p}$ which are not contained in $C_{k,f}$. By separately considering those $(1,f)$-clusters having zero, one, or two vertices which are not contained in $C_{k,f}$, we have (recalling $k \le \log n$)

$$\mathbb{E}[Z_{1,f,C_{k,f}} \mid C_{k,f} \subseteq G_{n,p}] \le O(k^2 p) + O(knp^3) + n^2p^5 = o(1) + 4^{-5} \le \mathrm{e}^{-4}.$$

Thus, if we let $\sum_{C_{k,f}}$ denote the sum over all (k,f)-clusters in K_n, we have

$$\begin{aligned}
\mathbb{E}[Z_{k+1,f}] &\le \sum_{C_{k,f}} \sum_{C_{1,f} \not\subseteq C_{k,f}} \mathbb{P}[\{C_{k,f} \subseteq G_{n,p}\} \cap \{C_{1,f} \subseteq G_{n,p}\}] \\
&= \sum_{C_{k,f}} \mathbb{P}[C_{k,f} \subseteq G_{n,p}]\, \mathbb{E}[Z_{1,f,C_{k,f}} \mid C_{k,f} \subseteq G_{n,p}] \\
&\le \sum_{C_{k,f}} \mathbb{P}[C_{k,f} \subseteq G_{n,p}]\, \mathrm{e}^{-4} = \mathbb{E}[Z_{k,f}]\, \mathrm{e}^{-4}.
\end{aligned}$$

Together with (7) this shows that $\mathbb{E}[Z_{k,f}] \le \mathrm{e}^{-4k}$ and thus that $\mathbb{E}[Z_{\log n,f}] \le n^{-4}$. Summing over all edges of the complete graph, we have that

$$\mathbb{P}[\mathcal{A}^c] \le \binom{n}{2} n^{-4} = o(1).$$

Now we prove (6), which is done in two steps. In the first step, we show that p is sufficiently large to ensure that with sufficiently high probability, T contains a large set of *edge-disjoint* triangles in $G_{n,p}$. In the second step, we will then show that p is sufficiently small to ensure that if $\mathcal{A}$ holds, then with sufficiently high probability, at least one of these triangles is also contained in $G_{n,p}(K_4)[T]$.

Turning to the first step, let X denote the number of triangles in $G_{n,p}[T]$ and let $\mu = \mathbb{E}[X]$. Note that

$$\mu = \binom{t}{3} p^3 \overset{(5)}{=} \frac{(1+o(1))t}{6} \left(10^4 n^{3/5} \sqrt{\log n}\right)^2 \left(n^{-2/5}/4\right)^3 \ge 10^5 t \log n. \tag{8}$$

Let Y denote the size of a largest family of edge-disjoint triangles in $G_{n,p}[T]$. We want to use (4) to give a lower bound on Y. Thus, let $\mathcal{S}$ be the set of triples of edges forming a triangle in $K_n[T]$. Δ, as defined in Section 2, is then the expected number of ordered pairs of triangles in $G_{n,p}[T]$ having exactly an edge in common. Thus

$$\Delta = O\left(t^4 p^5\right) = O\left(\mu t p^2\right) \le \mu n^{-1/5+o(1)} = o(\mu).$$

Also, ν is the expected number of triangles in $G_{n,p}[T]$ containing an edge of a fixed triangle and thus $\nu \leq 3tp^2 = n^{-1/5+o(1)} = o(1)$. η is the probability that a fixed triangle is contained in $G_{n,p}[T]$ and thus $\eta = p^3 = o(1)$. Thus by (4), with $\varepsilon = 1/2$, we have

$$\mathbb{P}[Y \leq \mu/2] \leq \mathrm{e}^{-(1+o(1))\mu/8} \overset{(8)}{\leq} n^{-3t}. \tag{9}$$

Now we turn to the second step. We define an auxiliary graph G_{K_4} on $G_{n,p}$ as follows. Let $\mathcal{S}' \subseteq \mathcal{S}$ be the lexicographically first set of edge-disjoint triangles in $G_{n,p}[T]$ of maximum size (i.e. $|\mathcal{S}'| = Y$). We say that a subgraph of K_n is a 1-*cluster* if it is the union of a K_3 in T and a K_4 which share at least an edge. For each triangle in $\mathcal{S}'$ which has an edge in common with a K_4 in $G_{n,p}$, pick the lexicographically first of these K_4's and let the 1-cluster which is the union of the triangle and this K_4 be a vertex of G_{K_4}. Two (distinct) vertices in G_{K_4} are joined by an edge if the corresponding 1-clusters share at least an edge. Since the number of vertices of G_{K_4} is equal to the number of elements of $\mathcal{S}'$ which share at least an edge with a K_4 in $G_{n,p}$, (6) will then follow if we can show that if $\mathcal{A}$ holds, then the probability that G_{K_4} has $|\mathcal{S}'| = Y$ vertices is sufficiently small.

To bound the number of vertices of G_{K_4}, we will use the fact that for any graph G, we have $v(G) \leq \alpha(G) + 2\Delta(G)\gamma(G)$, where $\alpha(G)$ denotes the size of a largest independent set of vertices in G, and $\gamma(G)$ denotes the size of a largest induced matching in G.

First we bound $\Delta(G_{K_4})$. For this, we will use the fact that any collection of size $6k$ of 1-clusters where the K_4's all contain a fixed edge f must contain a (k, f)-cluster. (To see this, consider the 1-clusters as given in a sequence $F_1, \dots, F_{6k}$ and denote by e_i the (lexicographically first) edge in F_i that is shared by the K_3 and the K_4. Now for each i (in increasing order) delete those F_j with $j > i$ from the current sequence for which e_j lies H_i, provided F_i was not previously deleted. Since the triangles in different F_j are edge-disjoint and $e(K_4) = 6$, a subsequence containing at least k of the F_i remains, and it is clear that for each such F_i the edge e_i is not contained in the K_4 of any of the F_j in that subsequence with $j < i$.) Also a 1-cluster has at most 8 edges. So if $\mathcal{A}$ holds, we have $\Delta(G_{K_4}) \leq 48 \log n$.

Next we bound $\alpha(G_{K_4})$. Let X_1 denote the number of 1-clusters in $G_{n,p}$ and let $\mu_1 = \mathbb{E}[X_1]$. A 1-cluster is either the union of a K_3 in T and a K_4 which have exactly an edge in common, or a K_4 with at least three vertices in T. Thus

$$\mu_1 \leq \binom{t}{3} n^2 p^8 + \binom{t}{3} np^6 = \mu \left(n^2p^5 + np^3\right) = \mu \left(4^{-5} + O(n^{-1/5})\right) \leq \mu/80.$$

Let Y_1 denote the size of a largest family of edge-disjoint 1-clusters in $G_{n,p}$. Note that $\alpha(G_{K_4}) \leq Y_1$. Then we have

$$\mathbb{P}[\alpha(G_{K_4}) \geq \mu/8] \leq \mathbb{P}[Y_1 \geq \mu/8] \overset{(2)}{\leq} (\mathrm{e}/10)^{\mu/8} \leq \mathrm{e}^{-\mu/8} \overset{(8)}{\leq} n^{-3t}.$$

Now we bound $\gamma(G_{K_4})$. The details of the argument we give here are tailored towards the K_4 case and thus slightly different than in the general case. Let

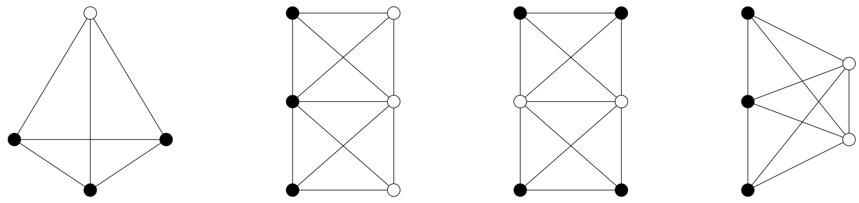

Figure 1: Minimal structures in a 2-cluster

$\mathcal{G}$ be the set of graphs $G_1, \ldots, G_4$ which are shown in Figure 1. We say that a copy of G_i in K_n is a T*-copy* of G_i if the black vertices are all contained in T (the white vertices are allowed to be anywhere). Let $X_{\mathcal{G}}$ denote the number of T-copies of graphs in $\mathcal{G}$ in $G_{n,p}$ and let $\mu_2 = \mathbb{E}[X_{\mathcal{G}}]$. Then

$$\mu_2 \le t^3 np^6 + t^3 n^3 p^{11} + t^4 n^2 p^{11} + t^3 n^2 p^9 = tn^{-1/5+o(1)} = \mu n^{-1/5+o(1)}. \qquad (10)$$

We say that a 2*-cluster* is the union of two 1-clusters which share at least an edge and which have been obtained from two edge-disjoint triangles in T. Let Y_2 denote the size of a largest family of edge-disjoint 2-clusters in $G_{n,p}$ and let Y_2' denote the size of a largest family in $G_{n,p}$ of edge-disjoint T-copies of graphs from $\mathcal{G}$. Now one may check that every 2-cluster must contain a T-copy of a graph from $\mathcal{G}$. Thus $Y_2 \le Y_2'$. Also note that $\gamma(G_{K_4}) \le Y_2$. Let $y_2 = \mu/(2000 \log n)$. Then

$$\mathbb{P}[\gamma(G_{K_4}) \ge y_2] \le \mathbb{P}[Y_2' \ge y_2] \overset{(2)}{\le} \left(\frac{\mathrm{e}\mu_2}{y_2}\right)^{y_2} \overset{(10)}{\le} n^{-(1+o(1))y_2/5} \overset{(8)}{\le} n^{-3t}.$$

Since for $i = 1,\ 2$, $\mathbb{P}[\{Y_i \ge y_i\} \cap \mathcal{A}] \le \mathbb{P}[Y_i \ge y_i]$, we have thus proved that if $\mathcal{A}$ holds, then with probability at least $1 - 2n^{-3t}$ we have $v(G_{K_4}) \le \mu/8 + 2\,(48 \log n) y_2 \le \mu/4$. Combining this with (9), this implies (6).

6 Theorem 2 – the upper bound

Let H be the forbidden graph, and suppose that H is strictly 2-balanced. Let x be a vertex of maximum degree in H, and let $\Gamma(x)$ denote the neighbourhood of x in H. Let $H \backslash x$ be the subgraph of H induced by $V(H) \backslash \{x\}$. Let $H_{\Gamma(x)}$ be the subgraph of $H \backslash x$ induced by $\Gamma(x)$. We may assume that $H \backslash x$ contains an edge (and thus also that $\Delta(H) > 1$), since otherwise H is a star, and hence not strictly 2-balanced.

We now define a generalized notion of the (non)-independence of a set. Let H and x be chosen as above. Suppose we are given a graph G together with a subset T of its vertices and a vertex $y \in V(G) \backslash T$. Then we say that (T, y) is (H, x)*-extendible in* G if G contains a copy of $H \backslash x$, which we call an *extension (of* (T, y) *in* G*)*, so that $H_{\Gamma(x)}$ is contained in T and all other vertices are contained in $V(G) \backslash \{T \cup \{y\}\}$. In other words, if y is connected to all vertices of T and (T, y) is (H, x)-extendible, then y may be extended to a copy of H where

x is mapped to y, the neighbours of x are mapped into T, and all other vertices are mapped outside $T \cup \{y\}$. Thus for example (T, y) is (K_3, x)-extendible in G if and only if T is not an independent set in G. (T, y) is (K_4, x)-extendible in G if and only if T contains a triangle, and (T, y) is (C_ℓ, x)-extendible in G if and only if G contains a path on $\ell - 1$ vertices whose endvertices are contained in T and whose remaining vertices are not in $T \cup \{y\}$. Finally, let $\mathcal{S} = \mathcal{S}_{T,y}$ denote the set of all extensions of (T, y) in K_n. Let

$$\beta = (v-2)/(e-1)$$

and note that $0 < \beta < 1$. Let

$$t = Cn^{1-\beta} (\log n)^{1/(\Delta(H)-1)}, \tag{11}$$

where C is some sufficiently large constant (see Proposition 10 below) depending only on H. Also let $p = c_0 n^{-\beta}$, where c_0 satisfies $c_0^{e-1} e^2 = 1/100$. So $t = pn^{1+o(1)}$ and

$$n^{v-2} p^{e-1} = 1/(100e^2). \tag{12}$$

The proof of Theorem 2 relies on Lemmas 7 and 8. Lemma 7 asserts that *a.a.s.* the copies of H in $G_{n,p}$ do not cluster too much. Lemma 8 implies that for the above choice of t and p, *a.a.s.* any set of size t will be extendible in $G_{n,p}(H)$, and thus in $M_{n,p}(H)$. Given an edge f in K_n and $k \in \mathbb{N}$ we say that a sequence $H_1, \dots, H_k$ of copies of H forms a (k, f, H)-*cluster* in K_n if they all contain f and have the property that for all i with $1 \le i \le k$, H_i contains an edge which is not contained in any of the other H_j with $j < i$. Let $\mathcal{A}$ denote the event that $G_{n,p}$ does not contain a $(\log n, f, H)$-cluster for any edge f.

Lemma 7 *$\mathcal{A}$ holds* a.a.s..

We defer the proof of this lemma, and first show how it may be used to prove the theorem. Given a subset T of vertices of K_n with $|T| = t$ and a vertex $y \in V(K_n) \backslash T$, let $X_{M(H)}$ be the number of extensions of (T, y) in $M_{n,p}(H)$, let $X_{G(H)}$ denote their number in $G_{n,p}(H)$, and let X_G denote their number in $G_{n,p}$. Let $\mu = \mathbb{E}[X_G]$. (We will suppress the dependence of our random variables on T and y throughout). Note that by (1), we have

$$X_{G(H)} \le X_{M(H)} \le X_G.$$

The second (and main) ingredient in the proof of the upper bound is the following lemma.

Lemma 8 *For a fixed pair (T, y) as above, we have*

$$\mathbb{P}[\{X_{G(H)} = 0\} \cap \mathcal{A}] \le n^{-2t},$$

provided n is sufficiently large.

The proof of this lemma is also deferred to the end of this section.

Proof of Theorem 2 (modulo Lemmas 7 and 8). Suppose we have a vertex y in $M_n(H)$ whose degree is at least t. Let T be a set of neighbours of y with $|T| = t$. As $M_n(H)$ contains no copies of H and $M_{n,p}(H) \subseteq M_n(H)$, $M_{n,p}(H)$ contains no copies of H, and so (T, y) is not extendible in $M_{n,p}(H)$. But since $X_{M(H)} \geq X_{G(H)}$, Lemma 8 tells us that this happens with probability tending to zero: there are (very crudely) at most n^{t+1} choices for (T, y) and thus the probability that there exists a pair (T, y) which is not extendible is at most

$$1 - \mathbb{P}[\mathcal{A}] + n^{t+1}\, n^{-2t}.$$

Applying Lemma 7 completes the proof of the theorem. □

We now prove Lemma 7 and Lemma 8. In the proofs we will repeatedly make use of the following three propositions. Throughout, for a graph F, we write Z_F for the number of copies of F in $G_{n,p}$.

Proposition 9 *There exists a constant $\delta > 0$, depending only on H, so that for any proper subgraph H' of H with $e(H') \geq 1$ and $v(H') \geq 3$, we have*

$$\mathbb{E}[Z_{H'}] \geq n^{2+\delta+o(1)} p = n^{\delta+o(1)}\mathbb{E}[Z_H].$$

Moreover, $\mathbb{E}[Z_H] \leq n^2 p/(100e^2)$.

Proof. Let

$$\delta = \min_{H' \subset H, e(H') \geq 2} \frac{v(H') - 2}{e(H') - 1} - \beta, \tag{13}$$

where the minimum is over all proper subgraphs H' of H with $e(H') \geq 2$. Note that the fact that H is strictly 2-balanced implies that $0 < \delta < 1$.

Let H' be a proper subgraph of H with $e(H') \geq 1$ and $v(H') \geq 3$. First suppose that $e(H') = 1$. Then for some sufficiently small constant $c > 0$, we have

$$\mathbb{E}[Z_{H'}] \geq cn^{v(H')}p \geq cn^{2+\delta}p.$$

Now suppose that $e(H') > 1$. Then for some sufficiently small constant $c > 0$, we have (recalling that $p = c_0 n^{-\beta}$)

$$\begin{aligned}
\mathbb{E}[Z_{H'}] &\geq cn^{v(H')}p^{e(H')} \\
&= cn^2 p \left(n^{(v(H')-2)/(e(H')-1)}p\right)^{e(H')-1} \\
&\geq cn^2 p \left(n^{\delta+\beta} c_0 n^{-\beta}\right)^{e(H')-1} \\
&\geq c\, c_0\, n^{2+\delta} p.
\end{aligned}$$

The right hand side and the "moreover" part follow from the fact that for some constant $c' \leq 1$, we have

$$\mathbb{E}[Z_H] = (1 + o(1))c' n^v p^e \stackrel{(12)}{=} (1 + o(1))c' n^2 p/(100e^2),$$

where the $o(1)$ term is negative. □

For the remainder of this section we fix δ as in (13). Let

$$c_2 = \min\{\delta, \beta, 1-\beta\}, \tag{14}$$

and note that $c_2 < 1$. Recall that μ is the expected number of extensions of (T, y) in $G_{n,p}$.

Proposition 10

$$\mu \geq \frac{80e^2}{c_2} t \log n.$$

Moreover, $\mu = n^{1+o(1)}p$.

Proof. For some sufficiently small c_1 depending on H, we have

$$\begin{aligned}
\mu &= |\mathcal{S}|\, p^{e-\Delta(H)} && (15)\\
&\geq c_1 t^{\Delta(H)} (n-t-1)^{v-\Delta(H)-1}\, p^{e-\Delta(H)} \\
&\overset{(11)}{\geq} c_1 C^{\Delta(H)-1}\, t \log n\, (n-t-1)^{v-2}\, p^{e-1} \\
&\overset{(12)}{=} \frac{c_1 C^{\Delta(H)-1}}{100e^2} t \log n \left(\frac{n-t-1}{n}\right)^{v-2}. && (16)
\end{aligned}$$

Recalling that $\Delta(H) > 1$ and $t = o(n)$, the assertion then follows if we choose C sufficiently large. The proof of the "moreover" part is similar. $\square$

The following proposition will imply that intersecting copies of H will typically only have an edge in common.

Proposition 11 *Given a copy F of a subgraph in K_n with $e(F) = O(\log n)$ and a proper subgraph H' of H with $e(H') \geq 1$, let $Z'_{H\cap F=H'}$ denote the number of copies of H in $G_{n,p} \cup F$ so that their intersection with F is isomorphic to H'. If H' is an edge, then*

$$\mathbb{E}[Z'_{H\cap F=H'}] \leq \frac{e(F)}{100e}.$$

If H' is not an edge, then

$$\mathbb{E}[Z'_{H\cap F=H'}] \leq n^{-\delta+o(1)}.$$

Proof. Suppose first that H' is an edge. Thus there are $e(F)$ possibilities for choosing H' in F and e possibilities for choosing H' in H. Thus in this case

$$\mathbb{E}[Z'_{H\cap F=H'}] \leq e(F) e n^{v-2} p^{e-1} \overset{(12)}{\leq} \frac{e(F)}{100e}.$$

Now suppose that H' is not an edge. Then

$$\mathbb{E}[Z'_{H\cap F=H'}] = O\left(e(F)^{e(H')} n^{v-v(H')} p^{e-e(H')}\right) = \frac{O\left(e(F)^{e(H')}\mathbb{E}[Z_H]\right)}{\mathbb{E}[Z_{H'}]}.$$

The result now follows from Proposition 9. $\square$

Proof of Lemma 7. Let $Z_{k,f,H}$ denote the number of (k, f, H)-clusters in $G_{n,p}$. By applying Proposition 11 with $F = f$, we have,

$$\mathbb{E}[Z_{1,f,H}] \leq 1/(100e) \leq \mathrm{e}^{-4}. \tag{17}$$

Note that every $(k+1, f, H)$-cluster can be viewed as the union of a (k, f, H)-cluster $C_{k,f}$ and a $(1, f, H)$-cluster which is not contained in $C_{k,f}$. For a fixed (k, f, H)-cluster $C_{k,f}$ in K_n (with $k < \log n$), the expected number of $(1, f, H)$-clusters in $G_{n,p} \cup C_{k,f}$ intersecting $C_{k,f}$ only in f is at most

$$en^{v-2}p^{e-1} \overset{(12)}{=} 1/(100e).$$

By Proposition 11, the expected number of $(1, f, H)$-clusters in $G_{n,p} \cup C_{k,f}$ intersecting $C_{k,f}$ in some proper subgraph H' of H which is not an edge is at most $n^{-\delta+o(1)}$. Let $Z_{1,f,C_{k,f}}$ denote the number of $(1, f, H)$-clusters in $G_{n,p}$ which are not contained in $C_{k,f}$. Then by the above,

$$\mathbb{E}[Z_{1,f,C_{k,f}} \mid C_{k,f} \subseteq G_{n,p}] \leq 1/(99e) \leq \mathrm{e}^{-4}.$$

Thus, if we let $\sum_{C_{k,f}}$ denote the sum over all (k, f, H)-clusters in K_n, we have

$$\begin{aligned}
\mathbb{E}[Z_{k+1,f,H}] &\leq \sum_{C_{k,f}} \mathbb{P}[C_{k,f} \subseteq G_{n,p}]\, \mathbb{E}[Z_{1,f,C_{k,f}} \mid C_{k,f} \subseteq G_{n,p}] \\
&\leq \sum_{C_{k,f}} \mathbb{P}[C_{k,f} \subseteq G_{n,p}]\, \mathrm{e}^{-4} \\
&= \mathbb{E}[Z_{k,f,H}]\, \mathrm{e}^{-4}.
\end{aligned}$$

Together with (17) this shows that $\mathbb{E}[Z_{k,f,H}] \leq \mathrm{e}^{-4k}$ and thus that $\mathbb{E}[Z_{\log n,f,H}] \leq n^{-4}$. Summing over all edges of the complete graph, the result now follows.

□

It now remains to prove Lemma 8. To this end, we first show that for our choice of p and of the size of T, there should be many *edge-disjoint* extensions of (T, y) in $G_{n,p}$ for a fixed pair (T, y). The next lemma will be the key towards proving this fact.

Lemma 12 *Consider a given copy F of a graph with at most $3e$ edges in K_n. Let $Z^*_{F \cup H \setminus x}$ denote the number of extensions in $G_{n,p} \cup F$ having at least an edge in common with F and which are not contained in F. Then*

$$\mathbb{E}[Z^*_{F \cup H \setminus x}] \leq n^{-\delta+o(1)}.$$

Proof. The main problem is that we have to take care of how many vertices of the intersection of an extension and F lie in T. Fix a proper subgraph H' of $H \setminus x$. Let $Z'_{H'}$ denote the number of those copies of H' in $G_{n,p}$ for which the vertices corresponding to vertices in $V(H') \cap \Gamma(x)$ are contained in T and the vertices not corresponding to vertices in $V(H') \cap \Gamma(x)$ do not lie in T. Let $\phi = |V(H') \cap \Gamma(x)|$. Then the expected number of extensions in $G_{n,p} \cup F$ whose intersection with F is H' (and thus exactly ϕ vertices of the intersection are contained in T) is

$$\begin{aligned}
&O\left(t^{\Delta(H)-\phi}\, n^{v-\Delta(H)-1-(v(H')-\phi)}\, p^{e-\Delta(H)-e(H')}\right)\\
&\overset{(15)}{=} O\left(\mu t^{-\phi}\, n^{-(v(H')-\phi)}\, p^{-e(H')}\right)\\
&= O\left(\mu/\mathbb{E}[Z'_{H'}]\right).
\end{aligned}$$

Let $\sum_{H'}$ denote the summation over all subgraphs H' of $H\backslash x$. Then by the above,

$$\mathbb{E}[Z^*_{F\cup H\backslash x}] = \sum_{H'} \frac{O(\mu)}{\mathbb{E}[Z'_{H'}]}. \tag{18}$$

Since $t = n^{1+o(1)}p$,

$$\mathbb{E}[Z'_{H'}] = O\left(\mathbb{E}[Z_{H'}](t/n)^{\phi}\right) = \mathbb{E}[Z_{H'}]p^{\phi}n^{o(1)} = \mathbb{E}[Z_{H'\cup x}]n^{-1+o(1)},$$

where $H' \cup x$ is the subgraph of H obtained from H' by adding a vertex x and connecting it to those vertices of H' that are contained in $\Gamma(x)$. Since H' is a proper subgraph of $H\backslash x$ containing at least one edge, $H' \cup x$ is a proper subgraph of H which is not an edge. Proposition 9 then implies that $\mathbb{E}[Z_{H'\cup x}] \geq n^{2+\delta+o(1)}p$ and thus that $\mathbb{E}[Z'_{H'}] \geq n^{1+\delta+o(1)}p$. Combining this with equation (18), and recalling that $\mu = n^{1+o(1)}p$ (Proposition 10), we have

$$\mathbb{E}[Z^*_{F\cup H\backslash x}] \leq \frac{\mu}{n^{1+\delta+o(1)}p} \leq n^{-\delta+o(1)}.$$

□

Let Y be the maximum number of edge-disjoint extensions of (T, y) in $G_{n,p}$ and recall that μ is the expected number of extensions of (T, y) in $G_{n,p}$.

Lemma 13 *For n sufficiently large,*

$$\mathbb{P}[Y \leq \mu/2] \leq n^{-3t}.$$

Proof. We shall apply inequality (4) to $\mathcal{S}$, the set of extensions of (T, y) in K_n. In order to be able to do this, we will apply Lemma 12 to show that η, ν and Δ (as defined in Section 2) are sufficiently small.

We first consider Δ. Since the extensions are symmetric (more formally, for all α, $\alpha' \in \mathcal{S}$ there is an automorphism of the underlying probability space which sends the event I_α to $I_{\alpha'}$), we have $\Delta = \mu\Delta^*$, where for some fixed $\alpha \in \mathcal{S}$,

$$\Delta^* = \sum_{\alpha' \sim \alpha} \mathbb{E}[I_{\alpha'} | I_\alpha = 1].$$

But, defining $Z^*_{\alpha\cup H\backslash x}$ as in Lemma 12 (with α playing the role of F), one may observe that $\Delta^* = \mathbb{E}[Z^*_{\alpha\cup H\backslash x}]$, and thus by Lemma 12, we have $\Delta^* \leq n^{-\delta+o(1)} = o(1)$. Thus $\Delta = o(\mu)$.

We now bound ν. Fix some $\alpha \in \mathcal{S}$ again. Note that, since $\alpha' \subseteq G_{n,p}$ certainly implies that $\alpha' \subseteq G_{n,p} \cup \alpha$, we have

$$\nu = \sum_{\alpha' \sim \alpha} \mathbb{E}[I_{\alpha'}] \leq \mathbb{E}[Z^*_{\alpha \cup H \backslash x}].$$

Thus by Lemma 12 again, we have $\nu \leq n^{-\delta + o(1)} = o(1)$.

Finally, since an extension contains at least an edge, we have $\eta = \mathbb{E}[I_\alpha] = p^{e-\Delta(H)} = o(1)$, and thus

$$\mathbb{P}[Y \leq \mu/2] \overset{(4)}{\leq} \mathrm{e}^{-(1+o(1))\mu/8} \leq n^{-3t},$$

where the final inequality follows from Proposition 10. $\square$

We now show that at least some of the extensions which are present in $G_{n,p}$ are also present in $G_{n,p}(H)$. In what follows, we will say that a graph $G \subset K_n$ can be obtained by *merging* two graphs G_1 and G_2 if G_1 and G_2 share at least an edge and if $G = G_1 \cup G_2$. We say that a subgraph F of K_n is a 1-*cluster* $E \cup H_1$ if it can be obtained by merging an extension E with a copy H_1 of H. A subgraph F of K_n is a 2-*cluster* if it can be obtained by merging two 1-clusters $E_1 \cup H_1$ and $E_2 \cup H_2$ where E_1 and E_2 are edge-disjoint extensions. We say that a 2-cluster is a *minimal* 2-*cluster* if it contains no 2-cluster as a proper subgraph.

For $a = 1, 2$, let Y_a be the maximum size of a family of pairwise edge-disjoint a-clusters in $G_{n,p}$. Let Y_2' denote the maximum size of a family of pairwise edge-disjoint *minimal* 2-clusters in $G_{n,p}$. Note that since every 2-cluster contains a minimal 2-cluster, we have $Y_2' = Y_2$. Let μ_1 denote the expected number of 1-clusters in $G_{n,p}$ and let μ_2 denote the expected number of minimal 2-clusters in $G_{n,p}$. Let $y_1 = \mu/8$ and $y_2 = \mu/(20e^2 \log n)$.

Lemma 14

$$\mathbb{P}\left[Y_1 \geq y_1\right] \leq n^{-3t}.$$

Proof. Fix an extension F in K_n and consider a subgraph H' (which contains at least an edge) of $H \backslash x$. As in Proposition 11, let $Z'_{H \cap F = H'}$ denote the number of copies of H in $G_{n,p} \cup F$ whose intersection with F is isomorphic to H'. Then, since a 1-cluster is the union of an extension and a copy of H sharing at least an edge, we have

$$\mu_1 \leq \mu \sum_{H' \subseteq H \backslash x} \mathbb{E}[Z'_{H \cap F = H'}],$$

where the summation is over all subgraphs H' of $H \backslash x$ with $e(H') \geq 1$. Thus Proposition 11 implies that

$$\mu_1 \leq \mu \left(1/100 + n^{-\delta + o(1)}\right) \leq \mu/80. \tag{19}$$

Combining this with inequality (2), we have

$$\mathbb{P}\left[Y_1 \geq y_1\right] \leq \left(\frac{\mathrm{e}\mu_1}{y_1}\right)^{y_1} \leq \mathrm{e}^{-\mu/8}.$$

The result now follows from Proposition 10. $\square$

Lemma 15

$$\mathbb{P}[Y_2 \ge y_2] \le n^{-3t}.$$

Proof. We will actually show that $\mathbb{P}[Y_2' \ge y_2] \le n^{-3t}$, which implies the lemma by our previous observation that $Y_2' = Y_2$. We first show that

$$\mu_2 \le \mu n^{-c_2+o(1)}, \tag{20}$$

where c_2 is defined as in (14). To this end, note that a (minimal) 2-cluster may be obtained from a 1-cluster F_1 by first merging F_1 with a copy H_2 of H (call the resulting graph an *augmented* 1-cluster), and then merging the augmented 1-cluster with an extension E. Write $\mu_2 = \mu_2' + \mu_2''$, where μ_2' denotes the expected number of minimal 2-clusters where the extension E is not contained in the augmented 1-cluster and μ_2'' denotes the expected number of minimal 2-clusters where it is (see Fig. 6). First we bound μ_2'. By (19), the expected number of 1-clusters is $O(\mu)$. We claim that also the expected number of augmented 1-clusters is $O(\mu)$. Indeed, the expected number of augmented 1-clusters with $H_2 \subset F_1$ is certainly $O(\mu)$. So consider those with $H_2 \not\subset F_1$. Since a 1-cluster has less than $2e$ edges, applying Proposition 11 to all proper subgraphs H' of H (with the 1-cluster playing the role of F in the statement of the proposition) shows that the expected number of augmented 1-clusters is also $O(\mu)$. Thus by Lemma 12 (with the augmented 1-cluster playing the role of F in the statement of the lemma), we have

$$\mu_2' \le \mu n^{-\delta+o(1)}.$$

It remains to estimate μ_2''. Unfortunately, in this case we cannot use Lemma 12 (since here we have $E \subset F$), so we have to work harder. Note that the augmented 1-clusters are obtained by merging a 1-cluster F_1 (which by definition must contain a copy H_1 of H) with a copy H_2 of H which may be contained in F_1.

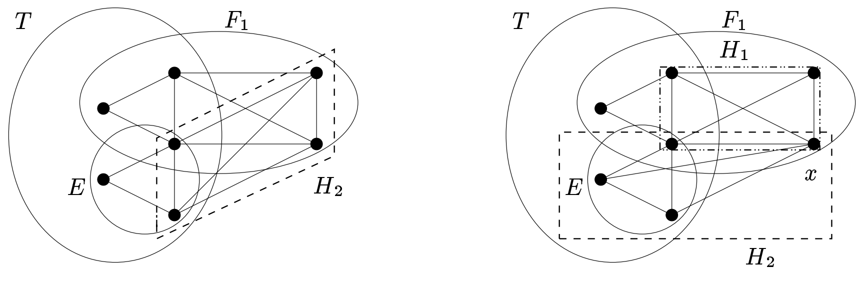

Figure 2: Minimal 2-clusters when $H = K_4$: the one on the left illustrates the estimation of μ_2' and the one on the right illustrates the estimation of μ_2'' for the case when $H_2 \not\subseteq F_1$.

Suppose first that $H_2 \not\subseteq F_1$. By definition of μ_2'', the augmented 1-cluster must contain the extension E. Since we are considering only minimal 2-clusters, the extension E must be edge-disjoint from H_1, otherwise $F_1 \cup E$ would already be a 2-cluster. As $H_2 \not\subseteq F_1$, this would contradict minimality. Thus $E \subset H_2$.

Now $V(H_2)\backslash V(E)$ contains only a single vertex x (which is the same vertex as in the definition of an extension) and so the fact that E is edge-disjoint from H_1 implies that $x \in F_1$ and that $F_1 \cap H_2$ consists only of edges running between x and E and hence is a star with i edges, where $1 \le i \le \Delta(H)$. Since for the extension E the vertices belonging to $\Gamma_{H_2}(x)$ must be contained in T, the vertices of E which are adjacent to the centre x must be mapped to T, while the remaining $v - \Delta(H) - 1$ vertices of E can be mapped anywhere. Since by (19), the expected number of 1-clusters is $O(\mu)$, the contribution to μ_2'' from the case $H_2 \not\subseteq F_1$ is thus

$$\sum_{i=1}^{\Delta(H)} O\left(\mu\, t^{\Delta(H)-i} n^{v-\Delta(H)-1} p^{e-i}\right). \tag{21}$$

Now consider the case when $H_2 \subseteq F_1$. Then the minimal 2-cluster is actually a 1-cluster containing two edge-disjoint extensions. (This can occur e.g. if H is a triangle and F_1 is a triangle in T.) Such a minimal 2-cluster can be obtained by merging an extension E with a copy H_1 of H which contains an extension. Again, by the edge-disjointness of the extensions the intersection $E \cap H_1$ is a star with at most $\Delta(H)$ edges. As the expected number of extensions is μ, arguing as above but starting from the extension E rather than the 1-cluster F_1, we find that the contribution to μ_2'' from this case is also

$$\sum_{i=1}^{\Delta(H)} O\left(\mu\, t^{\Delta(H)-i} n^{v-\Delta(H)-1} p^{e-i}\right). \tag{22}$$

Using the facts that $\Delta(H) > 1$ and that $\mu = n^{1+o(1)}p = tn^{o(1)}$ (see Proposition 10), (21) and (22) imply that

$$\begin{aligned}
\mu_2'' &\overset{(12)}{=} \sum_{i=1}^{\Delta(H)} O\left(\mu\, t^{\Delta(H)-i} n^{1-\Delta(H)} p^{1-i}\right) \\
&\le \mu\, n^{o(1)} \sum_{i=1}^{\Delta(H)} t^{1-i} p^{\Delta(H)-i} \le \mu\, n^{o(1)} \left(n^{-\beta} + n^{-(1-\beta)}\right),
\end{aligned}$$

where the first summand comes from the case $i = 1$ and the second one from the case $i = \Delta(H)$ (note that if $\Delta(H) > 2$ the terms with $1 < i < \Delta(H)$ are dominated by the ones with $i = 1$ and $i = \Delta(H)$). Thus (20) follows by the definition of c_2 in (14).

Combining inequalities (2) and (20) this time, we have

$$\begin{aligned}
\mathbb{P}\left[Y_2 \ge y_2\right] &\le \left(\frac{\mathrm{e}\mu_2}{y_2}\right)^{y_2} \\
&\le n^{-(1+o(1))c_2\mu/(20e^2 \log n)} \\
&\le n^{-(1+o(1))4t},
\end{aligned}$$

where the last line follows by Proposition 10. □

Proof of Lemma 8. Let $\mathcal{S}' \subseteq \mathcal{S}$ be the lexicographically first set of edge-disjoint extensions in $G_{n,p}$ of maximum size (so that $|\mathcal{S}'| = Y$). We now define an auxiliary graph G_H on $G_{n,p}$ as follows. For each extension in $\mathcal{S}'$ which has an edge in common with a copy of H in $G_{n,p}$ (and thus lies in a 1-cluster), pick the lexicographically first of these copies of H and let the 1-cluster which is the union of the extension and this copy of H be a vertex of G_H. Two (distinct) vertices are connected by an edge if the 1-clusters have at least an edge in common. Since the number of vertices of G_H is equal to the number of elements of $\mathcal{S}'$ which share at least an edge with a copy of H, the lemma will then follow if the probability that G_H has $|\mathcal{S}'| = Y$ vertices is sufficiently small.

To bound the number of vertices of G_H, we will use the fact that for any graph G, we have $v(G) \leq \alpha(G) + 2\Delta(G)\gamma(G)$, where $\alpha(G)$ denotes the size of a largest independent set of vertices in G and $\gamma(G)$ denotes the size of a largest induced matching in G.

First we bound $\Delta(G_H)$. For this, we will use the fact that any collection of size ek of 1-clusters where the copies of H all contain a fixed edge f must contain a (k, f, H)-cluster. (To see this, consider the 1-clusters as given in a sequence $F_1, \ldots, F_{ek}$, where $F_i = E_i \cup H_i$ is obtained by merging an extension E_i with a copy H_i of H. Also denote by e_i the (lexicographically first) edge in $E_i \cap H_i$. For each i (in increasing order) delete those F_j with $j > i$ from the current sequence for which e_j lies H_i, provided F_i was not previously deleted. Since the E_j are edge-disjoint, a subsequence $\mathcal{F}$ containing at least k of the F_i remains, and for each $F_i \in \mathcal{F}$ the edge e_i is not contained in any of the H_j with $j < i$ and $F_j \in \mathcal{F}$.) Also, a 1-cluster has at most $2e$ edges. So if $\mathcal{A}$ holds, we have $\Delta(G_H) \leq 2e^2 \log n$.

Furthermore, it is easily verified that the definitions of Y_1, Y_2, and G_H imply that $\alpha(G_H) \leq Y_1$ and $\gamma(G_H) \leq Y_2$. By Lemmas 14 and 15, $\mathbb{P}[\{Y_a \geq y_a\} \cap \mathcal{A}] \leq \mathbb{P}[Y_a \geq y_a] \leq n^{-3t}$ for $a = 1, 2$. Thus if $\mathcal{A}$ holds, with probability at least $1 - 2n^{-3t}$ our auxiliary graph has at most

$$y_1 + 2\,(2e^2 \log n)\, y_2 \leq (1/8 + 4/20)\mu \leq 2\mu/5 \tag{23}$$

vertices. On the other hand, by Lemma 13, we have

$$\mathbb{P}[\{Y \leq \mu/2\} \cap \mathcal{A}] \leq \mathbb{P}[Y \leq \mu/2] \leq n^{-3t}. \tag{24}$$

Since $X_{G(H)} \geq Y - v(G_H)$, this completes the proof of the lemma. □

7 Forbidding several strictly 2-balanced graphs

Here we indicate how the proofs of Sections 4 and 6 can easily be modified in order to establish Theorem 3. Denote by v_i and e_i the number of vertices and edges of H_i respectively and let $\beta = (v_0 - 2)/(e_0 - 1)$.

For the lower bound, denote by X_i the number of copies of H_i in $G_{n,p}$ and let $X = \sum_{i=0}^{s-1} e_i X_i$. Let $p = c_0 n^{-\beta}$, where c_0 satisfies $\max_i\{c_0^{e_i - 1} e_i\} = \frac{1}{8s}$. Then an upper bound on $\mathbb{E}[X]$, very similarly to the proof of Theorem 1, implies the result.

As for the proof of the upper bound, let $p = c_0 n^{-\beta}$, where c_0 now satisfies

$$\max_i \{c_0^{e_i - 1} e_0 e_i\} = \frac{1}{100s}.$$

In the definitions of (H,x)-*extendible*, *extension*, $X_{G(H)}$, $X_{M(H)}$ and X_G we substitute H_0 for H. Instead of (k,f,H)-clusters we consider $(k,f,\mathcal{H})$-*clusters*, where a $(k,f,\mathcal{H})$-cluster is now the union of k copies $H^1,\ldots,H^k$ of possibly different elements in $\mathcal{H}$ which all contain the edge f and which have the property that for all i with $1 \le i \le k$, H^i contains an edge which is not contained in any of the other H^j with $j < i$. It is then clear that the theorem follows from Lemmas 7 and 8 as before. In the proof of (the analogue of) Lemma 8 and within this proof Lemmas 14 and 15 in particular, the definition of a cluster is modified as follows: a 1-cluster is obtained by merging an extension and a copy of a graph in $\mathcal{H}$. A 2-cluster is again obtained by merging two 1-clusters, where we require disjointness of the extensions as before. Now it suffices to check (19) and (20).

With regard to (19), the main contribution again is from the case when each copy of an element H_i of $\mathcal{H}$ has exactly an edge in common with the extension. Note that each H_i contributes $\mu n^{v_i-2} p^{e_i-1} e_0 e_i$, which is less than $\mu/(100s)$ due to the choice of p. Hence the same bound as in (19) holds. (20) is also proven using the same arguments as in Section 6.

8 Forbidding an arbitrary subgraph – an open question

We remark that lower bounds on $e(M_n(H))$ when H is an arbitrary graph may be obtained by choosing p so that approximately half of the edges of $G_{n,p}$ lie in a copy of H. Deleting these edges as in Section 4 yields the lower bound. This value of p may be determined using the methods introduced by Spencer in [15]. We do not give any details however, as there are graphs for which the lower bounds obtained in this way are far from tight, as the example in the following paragraph will show.

Concerning upper bounds on the number of edges in $M_n(H)$, we have seen that if H is strictly 2-balanced, then $e(M_n(H))$ is not much larger that $e(G_{n,p})$, where p was chosen so that the number of copies of H in $G_{n,p}$ is roughly $e(G_{n,p})$. However, this is not true in general, as the following example shows. Let H be the graph which is obtained from K_4 by adding a new vertex and joining it to one of the vertices of the K_4. Then it is easily seen that *a.a.s.* $e(M_n(H))$ equals $e(M_n(K_4))$. (Indeed, suppose not and consider the first edge which is admitted in one process and not in the other.) The proof of Theorem 2 implies that $e(M_n(K_4))$ is *a.a.s.* close to $e(G_{n,p})$ for $p = n^{-2/5}$. However the number of copies of H in $G_{n,p}$ is close to $e(G_{n,p})$ only if p is close to $n^{-1/2}$.

Alternatively, one might think that for all graphs H with $v(H) \ge 3$ and $e(H) \ge 2$, for any $\varepsilon > 0$ *a.a.s.* we have

$$n^{2-1/m_2(H)-\varepsilon} \le e(M_n(H)) \le n^{2-1/m_2(H)+\varepsilon}, \qquad (25)$$

where for a graph H with $v(H) \geq 3$ and $e(H) \geq 2$ its 2-*density* $m_2(H)$ is defined by

$$m_2(H) = \max_{H' \subseteq H,\ v(H')>2} \frac{e(H')-1}{v(H')-2}.$$

The following example shows that this too need not be true. Furthermore, it shows that $e(M_n(H))$ need not necessarily be concentrated on a single interval. Let H be the disjoint union of a triangle and a cycle of length four. Then with probability bounded away from zero, a triangle appears in the random greedy process before a four-cycle, and thus one may check that the proof of Theorem 2 shows that in this case, *a.a.s.* $e(M_n(H))$ will be close to $n^{4/3}$. With probability bounded away from zero, it will be the other way round, and then *a.a.s.* $e(M_n(H))$ will be close to $n^{3/2}$. Thus it appears to be impossible to give good bounds on $e(M_n(H))$ in general. Nevertheless, we conjecture that for an arbitrary graph H with $v(H) \geq 3$ and any $\varepsilon > 0$, *a.a.s.* we have

$$e(M_n(H)) \leq n^{2-1/m_2(H)+\varepsilon}.$$

9 Acknowledgements

We would like to thank Hans Jürgen Prömel, who told us about the problem, and Andreas Marwinski, whose computer simulations pointed us in the right direction.

We would also like to thank both referees for their detailed comments on an earlier version of this paper and for pointing out several mistakes. Finally, we are indebted to Andrzej Ruciński for pointing out several simplifications in the proof.

References

[1] D. Aldous, A random tree model associated with random graphs, *Random Struct. Algorithms* **1** (1990), 383–402.

[2] N. Alon and J.H. Spencer, *The Probabilistic Method*, Wiley-Interscience, New York, 1992.

[3] B. Bollobás and O.M. Riordan, Constrained graph processes, *Electronic Journal of Combinatorics* **7** (2000), # R18.

[4] F. Chung, Z. Füredi, R.L. Graham, and P.D. Seymour, On induced subgraphs of the cube, *J. Combin. Theory Ser. A* **49** (1988), 180–187.

[5] F. Chung and R.L. Graham, *Erdős on Graphs: His Legacy of Unsolved Problems*, A. K. Peters, Wellesley, 1998.

[6] P. Erdős, S. Suen, and P. Winkler, On the size of a random maximal graph, *Random Struct. Algorithms* **6** (1995), 309–318.

[7] P. Erdős and P. Tetali, Representation of integers as the sum of k terms, *Random Struct. Algorithms* **1** (1990), 245–261.

[8] A. Frieze, On the value of a minimum spanning tree problem, *Discrete App. Math.* **10** (1985), 47–56.

[9] S. Janson, T. Łuczak, and A. Ruciński, *Random Graphs*, Wiley-Interscience, New York, in press.

[10] T.R. Jensen and B. Toft, *Graph Coloring Problems*, Wiley-Interscience, New York, 1995.

[11] J.H. Kim, The Ramsey number $R(3,t)$ has order of magnitude $t^2/\log t$, *Random Struct. Algorithms* **7** (1995), 173–207.

[12] M. Krivelevich, Bounding Ramsey numbers through large deviation inequalities, *Random Struct. Algorithms* **7** (1995), 145–155.

[13] D. Osthus, H.J. Prömel, and A. Taraz, On the evolution of triangle-free graphs, 1999, submitted.

[14] A. Ruciński and N.C. Wormald, Random graph processes with degree restrictions, *Combin. Probab. Comput.* **1** (1992), 169–180.

[15] J.H. Spencer, Counting extensions, *J. Combin. Theory Ser. A* **53** (1990), 247–255.

[16] J.H. Spencer, Maximal triangle-free graphs and Ramsey $R(3,t)$, unpublished manuscript, 1995.