# A variant of the Erdős-Rényi random graph process

Adam Logan* Mike Molloy† Paweł Prałat‡

September 15, 2021

**Abstract**

We consider a natural variant of the Erdős-Rényi random graph process in which $k$ vertices are special and are never put into the same connected component. The model is natural and interesting on its own, but is actually inspired by the multiway cut problem that itself is connected to a number of important problems in graph theory. We will show that a phase transition for appearance of the giant component occurs when the number of special vertices is roughly $n^{1/3}$, where $n$ is the number of vertices.

## 1 Introduction

The study of the *random graph process* was initiated by Erdős and Rényi in their celebrated paper from 1959 [7]. The process starts with an empty graph on $n$ vertices and adds all $\binom{n}{2}$ edges in a random order. The model is now well understood, though there are still some challenging questions waiting to be answered (see the following monographs on the topic: [5,10,14]). On the other hand, relatively little is known about variants of this process. In particular, a natural variant of the model is the *constrained random graph process* in which, after the edge to be inserted is chosen at random, we check whether the graph at this stage of the process together with this edge satisfies some properties; if so, we accept it, otherwise we reject it and never consider it again.

The first result on the constrained random graph process is due to Ruciński and Wormald, who answered a question of Erdős regarding the process in which we maintain a bound on the maximum degree [18]. Erdős, Suen, and Winkler considered both the odd-cycle-free process [8] and the triangle-free process that was later analyzed by Bohman [2] (see also [3, 9]). Other special cases that were considered include the properties of being cycle-free [1], $H$-free [4, 17], and planarity [12].

In this paper, we consider another natural constrained random graph process, the $k$-process, in which $k$ vertices are special and never put into the same connected component. The main question we address is the following: how large is a largest component at the end of this process? This problem was motivated by considering a natural greedy algorithm for the multiway cut problem that is a special case of the combinatorial data fusion problem.

The paper is structured as follows. In Section 2 we introduce necessary definitions (including a formal definition of the $k$-process) and state main results. Connections to the combinatorial data fusion problem are discussed in Section 3. The original random graph process mentioned above is formally introduced in Section 4 where we also make a connection between the two models (the random graph process and the $k$-process) and

*The Tutte Institute for Mathematics and Computing, Ottawa, ON, Canada.
†Department of Computer Science, University of Toronto, Toronto, ON, Canada.
‡Department of Mathematics, Ryerson University, Toronto, ON, Canada.

list all properties of the original one that we need to understand our model. In Section 5, we develop some concentration tools that will be used in the proofs. The process shows two different behaviours: $k \ll n^{1/3}$ is considered in Section 6; Section 7 is devoted to $k \gg n^{1/3}(\log n)^{4/3}(\log\log n)^{1/3}$. Final conclusions are in Section 8.

# 2 Definitions and main results

In this section, we introduce the $k$-process we are interested in, the asymptotic notation used throughout the paper, and state the main results.

## 2.1 $k$-process

Let $1 \le k \le n$ be any two integers ($k = k(n)$ may be and usually is a function of $n$). The *$k$-process* starts with $\mathbb{P}_k(n,0)$, the empty graph on $n$ vertices, where $k$ of the vertices are *special*. For integer $m \ge 1$, create $\mathbb{P}_k(n,m)$ from $\mathbb{P}_k(n,m-1)$ as follows. Choose a random pair of vertices not yet considered (in particular, they are not connected by an edge); connect these two vertices unless doing so would put two of the special vertices in the same component (in which case we say that a *collision* occurs). Keep repeating these steps, if needed, until one edge is added. In particular, $\mathbb{P}_k(n,m)$ has $m$ edges. The process stops at time $M = M(n,k)$ when $\mathbb{P}_k(n,M)$ has precisely $k$ connected components, each of which is a complete graph. (Of course, $M$ is a random variable counting the number of edges at the end of the process.) Alternatively, one can stop the process much earlier, at time $\hat{M} = \hat{M}(n,k)$ when $\mathbb{P}_k(n,\hat{M})$ has $k$ connected components for the first time, as $M$ and $\mathbb{P}_k(n,M)$ are already determined at this point.

The main question raised in this paper is the following one. What can be said about the distribution of sizes of the connected components of $\mathbb{P}_k(n,M)$? Another natural question is: what can be said about $M(n,k)$ as a function of $k$? What about $\hat{M}(n,k)$?

## 2.2 Asymptotics

As typical in random graph theory, we shall consider only asymptotic properties of $\mathbb{P}_k(n,m)$ (and $\mathbb{G}(n,m)$, $\hat{\mathbb{G}}_k(n,m)$ defined below) as $n \to \infty$, where $m = m(n)$ depends on $n$. We emphasize that the notations $o(\cdot)$ and $O(\cdot)$ refer to functions of $n$, not necessarily positive, whose growth is bounded. We use the notations $f \ll g$ for $f = o(g)$ and $f \gg g$ for $g = o(f)$. We also write $f(n) \sim g(n)$ if $f(n)/g(n) \to 1$ as $n \to \infty$ (that is, when $f(n) = (1+o(1))g(n)$). We say that an event in a probability space holds *asymptotically almost surely* (*a.a.s.*) if its probability tends to one as $n$ goes to infinity.

## 2.3 Results

In this subsection, we summarize the main results proved in this paper. It turns out that the $k$-process changes its behaviour around $k = n^{1/3}$. If $k \ll n^{1/3}$, a single large component is formed before collisions start affecting the process. In particular, when the first special vertex joins a largest component, its size is much larger than the total size of all other special components. As a result, this component will continue growing and at the end of the process it will have size $n(1-o(1))$. On the other hand, if $k \gg n^{1/3}$, collisions will start affecting the process much earlier, namely, when each component has size smaller than the total size of all special components. As a result, no component is able to dominate all the others and the largest component in the end has size $o(n)$. For technical reasons (see the

final section for a longer discussion), our proofs require slightly larger values of $k$, namely, $k \gg n^{1/3}(\log n)^{4/3}(\log\log n)^{1/3}$. Below we state these results precisely.

For a given graph $G$ (deterministic or random) with $r = r(G)$ connected components, let $C_1(G), \ldots, C_{r(G)}(G)$ be the connected components in order of decreasing size with ties arbitrarily broken, and let $L_i(G)$ be the size of $C_i(G)$ $(i = 1, 2, \ldots, r)$. Then the following holds.

**Theorem 2.1.** *Let $\omega = \omega(n)$ be any function tending to infinity as $n \to \infty$ (sufficiently slowly). Then a.a.s. the following holds:*

*(a) If $k \ll n^{1/3}$, then*

$$L_1(\mathbb{P}_k(n, M)) = n - O(k^3\omega \log(n^{1/3}/k)) \sim n.$$

*As a result,*

$$M(n, k) \sim \binom{n-k+1}{2} \sim \frac{n^2}{2}.$$

*(b) If $k \gg n^{1/3}(\log n)^{4/3}(\log\log n)^{1/3}$ and $k \ll n/\log n$, then*

$$L_1(\mathbb{P}_k(n, M)) = O\left(\left(\frac{n\log(k/n^{1/3})\log^4 n}{k^3}\right)^{1/2} n\right) = o(n).$$

*As a result,*

$$M(n, k) = O\left(\left(\frac{n\log(k/n^{1/3})\log^4 n}{k^3}\right)^{1/2} n^2\right) = o(n^2).$$

*(c) If $n/(\omega\log n) \le k \le n$, then $L_1(\mathbb{P}_k(n, M)) = O(\omega^{3/2}\log^4 n) = o(n)$. As a result, $M(n, k) = O(n\omega^{3/2}\log^4 n) = o(n^2)$.*

Part (a) is proved in Section 6; part (b) is proved in Section 7; part (c) follows immediately from part (b) and Observation 4.2 (see Subsection 7.3 for more details).

Note that, trivially and deterministically, we get that $M(n, k) \le \binom{n-k+1}{2}$; the upper bound holds for the extremal graph on $n$ vertices, $k$ components, and maximum number of edges (union of a complete graph on $n - k + 1$ vertices and $k - 1$ isolated vertices). Understanding whether $M(n, k)$ is close to this trivial bound has important implications for the applications we consider below.

# 3 Motivation

We feel that the $k$-process is natural and inherently interesting in its own right. That said, we will now outline its relationship to the multiway cut problem and the combinatorial data fusion problems as it was these problems that first led us to consider this process.

Before we formally state the combinatorial data fusion problem, let us start with an important special case. The *multiway cut* problem [19, Chapter 4] is a standard NP-hard problem in graph theory. We state it as follows. Let $G$ be a nonnegatively weighted undirected graph with vertex set $V$ and edge set $E$. Let $S \subseteq V$. Find the set $R$ of edges of least total weight such that no two vertices in $S$ are in the same connected component of the graph $G_R$ obtained by removing the edges in $R$ from $G$.

There is a natural greedy algorithm that can be used to construct an approximate solution to this problem:

**Algorithm 3.1.** *(Edge-first greedy algorithm)*

1. *Begin with an empty graph $G_0$ with the same vertices as $G$.*
2. *Order the edges of $G$ as $e_1, e_2, \ldots, e_m$ by decreasing weight, breaking ties uniformly at random.*
3. *For each $i$ with $1 \leq i \leq m$, construct $G_i$ as follows. If adding $e_i$ to $G_{i-1}$ does not create a connected component containing more than one element of $S$, then $G_i$ is $G_{i-1}$ with $e_i$ added; otherwise $G_i = G_{i-1}$.*
4. *Return $R_{greedy} = E(G) \setminus E(G_m)$.*

Clearly this algorithm runs in polynomial time; thus we cannot expect it to find the optimal solution. In fact, it does not approximate the best solution within any constant factor.

**Example 3.2.** *Let $G = K_{n^2}$, and index the vertices as $v_{i,j}$ for $1 \leq i, j \leq n$. Let $S = \{v_{1,j} : 1 \leq j \leq n\}$, and let the weight of the edge joining $v_{i,j}$ to $v_{k,\ell}$ be $1 + \epsilon$ if $j = \ell$ and 1 otherwise. Note that the total weight is equal to $\binom{n^2}{2} + \binom{n}{2} n\epsilon \sim n^4/2$. The greedy algorithm adds all edges of weight $1 + \epsilon$ and then stops, thus producing a graph $G_m$ of total weight $n\binom{n}{2}(1 + \epsilon) \sim n^3(1 + \epsilon)/2 = o(n^4)$. (The order in which edges are added to the graph depends on the ordering of the edges chosen in step 2 of the algorithm, but the final result does not.) Hence, a solution $R_{greedy}$ returned by the algorithm has weight that is asymptotically equal to $n^4/2$, the weight of the whole graph. On the other hand, let $R$ be the set of all edges incident to at least one vertex of $S \setminus \{v_{1,1}\}$. Since $|R| = \binom{n-1}{2} + (n-1)(n^2 - (n-1)) + (n-1)^2\epsilon \sim n^3$, this solution is better by a multiplicative factor of $n/2$ than the solution found by the greedy algorithm.*

Nevertheless, understanding the performance of the greedy algorithm may give some insight into the general problem. As a starting point we consider how it performs on the complete graph with random edge-weights.

Let $G = K_n$, let $S$ be any set of $k$ vertices, and let the edges have i.i.d. random weights. If any edges have the same weight then when running Algorithm 3.1 we break ties uniformly at random. Since the weights are i.i.d., the order that the edges are processed is a uniform permutation. The weights have no other affect on the choice of which edges are selected and so it is not hard to see that the choice of edges in this algorithm is equivalent to the $k$-process defined in Section 2.1.

The main theorem (Theorem 2.1) shows that a.a.s. for $k \ll n^{1/3}$ we obtain an output where all but one of the components are very small, while for $k \gg n^{1/3}(\log n)^{4/3}(\log\log n)^{1/3}$ we do not. One simple case of i.i.d. variables is where every edge is given weight 1. In that case, the optimal solution has $k - 1$ components of size one and one of size $n - k + 1$; so for $k \ll n^{1/3}$ the greedy solution is close to optimal and for larger $k$ it is far from optimal.

The multiway cut problem is a special case of the combinatorial data fusion problem which was introduced by Darling et al. [6], and can be used to describe many situations that are important in applications.

**The Combinatorial Data Fusion Problem:** Let $G$ be a nonnegatively weighted undirected graph with vertex set $V$ and edge set $E$. Fix a set $S \subseteq 2^V$ of sets of vertices, called the *forbidden sets*. A set $T \subseteq E$ of edges of $G$ is a solution to the *combinatorial data fusion problem* associated to $(G, S)$ if, after removing the edges of $T$, no set of vertices in $S$ is wholly contained in a single connected component. As before, our goal is to find a set $T$ of least total weight.

The combinatorial data fusion problem generalizes a number of standard graph-theoretic problems. For example, suppose that $S$ consists of all 2-element subsets of a given subset $U \subseteq V$. Then we have the *multiway cut* problem described above. More generally, if $S$ only contains sets of order 2, then we have the *multicut* problem [19, Problem 18.1]. In a rather different direction, let $G$ be a star. Without loss of generality we may assume that no element of $S$ contains the central vertex: if there is such a set with 2 elements, then they are joined by an edge, so we delete the edge and the non-central vertex on it and proceed. If there is such a set with $> 2$ elements, it is disconnected if and only if the subset obtained by removing the central vertex is disconnected. In this case, disconnecting a set of vertices is the same as removing the edge incident on one of them. Thus we have the *minimum hitting set* problem, which is another standard NP-complete problem [11, SP8, p. 222]: given a collection of subsets $S$ of a finite set $U$, a *hitting set* is a subset of $U$ that meets every element of $S$. The problem is to determine whether there is a hitting set of size less than $k$.

Considering the more general setting of the combinatorial data fusion problem leads to a broad class of random graph processes:

**The CDF-process:** Begin with $n$ vertices and specify a collection of *forbidden* subsets of those vertices. Add random edges one at a time by repeatedly choosing a pair of vertices uniformly from all non-edges whose addition to the graph would not form a component containing a forbidden set of vertices.

The same questions arise: how many edges will be added until the process is complete? What will the components look like? We have not studied this more general process at all, beyond the special case of the $k$-process.

# 4 The random graph process and its properties

In this section, we introduce the random graph process $\mathbb{G}(n,m)$ which will be very useful in the analysis of $\mathbb{P}_k(n,M)$. In particular, a good understanding of the process of forming a giant component in $\mathbb{G}(n,m)$ is needed. We summarize our knowledge on this topic in the last two subsections.

## 4.1 Random graph models

We fix $n$ vertices. $\mathbb{G}(n,m)$ is the random graph selected uniformly from all graphs on those vertices and with exactly $m$ edges. Equivalently, we can select $\mathbb{G}(n,m)$ by the following process:

**Random graph process**: $e_1,\dots,e_{\binom{n}{2}}$ is a sequence of pairs of vertices selected uniformly without repetition. $\mathbb{G}(n,m)$ is the graph formed by edges $e_1,\dots,e_m$.

We can couple this process to a process that is essentially identical to $\mathbb{P}_k(n,m)$, differing only in what $m$ counts:

$\hat{\mathbb{G}}_k(n,m)$**-process**: $S$ is a set of $k$ *special* vertices from amongst our $n$ vertices. As before, $e_1,\dots,e_{\binom{n}{2}}$ is a sequence of pairs of vertices selected uniformly without repetition. $\hat{\mathbb{G}}_k(n,m)$ is the graph formed by starting with the empty graph on $n$ vertices and considering edges $e_1,\dots,e_m$ one-at-a-time; each time, we add $e_i$ unless it joins two components that each contain a special vertex.

So the number of edges in $\hat{\mathbb{G}}_k(n,m)$ is not $m$, rather it is $m$ minus the number of edges that were skipped. However, note that $\hat{\mathbb{G}}_k(n,m=\binom{n}{2})$ is identical to $\mathbb{P}_k(n,M)$ (recall that

$M = M(n,k)$ is defined to be the number of edges in $\mathbb{P}_k$ once no more edges can be added; that is, after all $\binom{n}{2}$ edges are considered). So it suffices to prove the properties stated in Theorem 2.1 (namely, the sizes of largest components) for $\hat{\mathbb{G}}_k(n, m = \binom{n}{2})$ instead of $\mathbb{P}_k(n, M)$.

**Remark 4.1.** *It will be useful to note that the sequence of edges $e_1, \ldots, e_m$ is independent of the set of special vertices. By symmetry, we can assume that the special vertices are chosen uniformly from amongst the $n$ vertices, and so we can first choose $e_1, \ldots, e_m$ and then choose the $k$ vertices uniformly.*

As mentioned above, $\hat{\mathbb{G}}_k(n,m)$ is coupled to the random graph process $\mathbb{G}(n,m)$ by using the same sequence $e_1, \ldots, e_m$. This coupling will be valuable as it will allow us to apply some deep and technical results regarding the giant component of $\mathbb{G}(n,m)$. However, there are also some implications that are easy and straightforward, but useful at the same time.

For example, one can couple $\hat{\mathbb{G}}_k(n,m)$ and $\hat{\mathbb{G}}_{k+1}(n,m)$ by making sure that each special vertex in $\hat{\mathbb{G}}_k(n,m)$ is also special in $\hat{\mathbb{G}}_{k+1}(n,m)$. Then, it is clear that a collision in $\hat{\mathbb{G}}_k(n,m)$ is also a collision in $\hat{\mathbb{G}}_{k+1}(n,m)$. Moreover, at every step the set of components of $\hat{\mathbb{G}}_k(n,m)$ differs from the set of components of $\hat{\mathbb{G}}_{k+1}(n,m)$ only in that possibly two of the special components in the latter process are joined into one component in the former. This yields the following monotonicity result:

**Observation 4.2.** *For any $1 \le k_1 \le k_2 \le n$, and for any $m$:*

*(a) the largest special component in $\hat{\mathbb{G}}_{k_1}(n,m)$ is at least as big as the largest component in $\hat{\mathbb{G}}_{k_2}(n,m)$; and*

*(b) $\hat{M}(n,k_1) > \hat{M}(n,k_2)$, where $\hat{M}(n,k)$ is defined to be the number of edges in $\hat{\mathbb{G}}_k(n, m = \binom{n}{2})$.*

Here is another implication. Note that if $\mathbb{G}(n,m)$ is connected, then $\hat{\mathbb{G}}_k(n,m)$ has $k$ components. Since $\mathbb{G}(n,m)$ is a.a.s. connected for $m = n\log n/2 + n\omega$ (where $\omega = \omega(n)$ is any function tending to infinity as $n \to \infty$), a.a.s. $\hat{M}(n,k) \le n\log n/2 + n\omega$ for any $2 \le k \le n$. Since, $\mathbb{G}(n,m)$ is a.a.s. disconnected for $m = n\log n/2 - n\omega$, this bound is sharp for $k = 1$: $\hat{M}(n,1) \sim n\log n/2$. In fact, it is straightforward to prove that $\hat{M}(n,k) \sim n\log n/2$ for $k \ll n^{1/3}$. Indeed, a.a.s. many (precisely $(1+o(1))e^{2\omega}$) non-special vertices in $\hat{\mathbb{G}}_k(n,m)$ are still isolated at time $m = n\log n/2 - n\omega$; hence, a.a.s. there are more than $k$ components in $\hat{\mathbb{G}}_k(n,m)$. It follows that a.a.s. $\hat{M}(n,k)$ is at least $m$ minus the number of collisions up to this point of the process. More importantly, by Theorem 2.1(a), a.a.s. the giant component in $\mathbb{P}_k(n,\hat{M})$ has size $n(1-o(1))$, and so a.a.s. all collisions that occurred in $\hat{\mathbb{G}}_k(n,m)$ must involve vertices from a small set of size $o(n)$. Finally, one can show that a.a.s. removing any set of size $o(n)$ from $\mathbb{G}(n,m)$ decreases the number of edges by $o(n\log n)$. As this is a trivial bound for the number of collisions in $\hat{\mathbb{G}}_k(n,m)$, we get that a.a.s. $\hat{M}(n,k) \sim n\log n/2$, provided that $k \ll n^{1/3}$. We do not provide a formal proof here as it seems that understanding the behaviour of $\hat{M}(n,k)$ for $k \gg n^{1/3}$ requires more work and a better understanding of the process. $\hat{M}(n,k)$ continues to decrease as $k$ increases (see Observation 4.2) and, trivially, $\hat{M}(n,n) = 0$. But the behaviour of this is unknown.

We will also make use of the classical $\mathcal{G}(n,p)$ model: we begin with $n$ vertices and then decide to include each of the $\binom{n}{2}$ possible edges independently with probability $p$. A standard and very useful fact is that we can typically translate a.a.s. properties between $\mathbb{G}(n,m)$ and $\mathcal{G}(n,p)$ when $m \approx p\binom{n}{2}$. For example, in this paper we will use the following, which comes from (1.6) in [14]. Since the proof is short, we provide it for completeness.

**Lemma 4.3.** *Let $m = m(n)$ be any function such that $n/4 \le m \le n$, let $\gamma = \gamma(n)$ be any function such that $\sqrt{n} \ll \gamma \le m$, and take $p = p(n) = (m-\gamma)/\binom{n}{2}$. Then one can couple the two processes such that a.a.s. $\mathcal{G}(n,p) \subseteq \mathbb{G}(n,m)$.*

*Proof.* Note that $\mathcal{G}(n,p)$ can be generated in two steps. First, we expose the total number of edges $M$, *without* exposing any edges. $M$ is the binomial random variable $\mathrm{Bin}(\binom{n}{2}, p)$ with $\mathbb{E}[M] = p\binom{n}{2} = m - \gamma$. It follows from the well-known Chernoff's bound (see also Lemma 5.1 that generalizes this bound) that

$$\Pr(M \ge m) = \Pr(M \ge \mathbb{E}[M] + \gamma) \le \exp\left(-\frac{\gamma^2}{2(m-\gamma+\gamma/3)}\right) \le \exp\left(-\frac{\gamma^2}{2m}\right) = o(1),$$

as $\gamma^2 \gg n \ge m$. Since we aim for a statement that holds a.a.s., we may assume that $M < m$.

Clearly, once the number of edges is fixed, then $\mathcal{G}(n,p) = \mathbb{G}(n,M)$. Indeed, in this conditional probability space, every graph on $n$ vertices and $M$ edges has the same probability of being generated. We get that a.a.s. $\mathcal{G}(n,p) = \mathbb{G}(n,M) \subseteq \mathbb{G}(n,m)$, and the proof of the lemma is finished. □

## 4.2 Largest component in $\mathbb{G}(n,m)$

We will need the following well-known result on the component sizes of $G(n,m)$ when $m$ is close to the critical point $n/2$. These bounds follow immediately from Theorems 5 and 6, and Corollary 1 of [15].

**Lemma 4.4.** *A.a.s. the random graph process is such that:*

*(a) For every integer $m$ where $m = \lfloor \frac{n}{2}(1 - \lambda n^{-1/3}) \rfloor$ for some $0 \ll \lambda \ll n^{1/3}$, the largest component in $\mathbb{G}(n,m)$ has size $\Theta(n^{2/3}\lambda^{-2}\log\lambda)$.*

*(b) For every integer $m$ where $m = \lfloor \frac{n}{2}(1 + \lambda n^{-1/3}) \rfloor$ for some $0 \ll \lambda \ll n^{1/3}$*

*(i) the largest component in $\mathbb{G}(n,m)$ has size $(2+o(1))\lambda n^{2/3}$;*

*(ii) the second largest component in $\mathbb{G}(n,m)$ has size $\Theta(n^{2/3}\lambda^{-2}\log\lambda)$;*

*(iii) the excess (the difference between the number of edges and vertices) of the largest component in $\mathbb{G}(n,m)$ is $(2/3+o(1))\lambda^3 = o(\lambda n^{2/3})$ (in particular, the number of edges in the largest component is asymptotic to the number of vertices).*

The range of $m$ covered in part (a) is referred to as the *subcritical range*; the range covered in part (b) is the *supercritical range*. We also know that one of the components becomes substantially larger than its competitors during the so-called *critical phase* when $m = \lfloor \frac{n}{2}(1 + \Theta(n^{-1/3})) \rfloor$. Indeed, during the critical phase, both the largest component and the second largest component have cardinalities of order $n^{2/3}$. The ratio of their corresponding sizes, however, increases as $\lambda$ increases. The largest component keeps growing but the size of the second largest component is decreasing (note that these components are absorbed from time to time by the largest component). This behaviour continues during the supercritical range and the largest component eventually becomes the giant component. that is, a component that contains a positive fraction of all vertices. (See Theorem 7 in [15].) At that point, the second largest component has only logarithmic size. (See, for example, Theorem 5.4 in [14].) We will often abuse slightly the terminology and (prematurely) call the largest component the giant if we are guaranteed that this component eventually becomes giant a.a.s.

### 4.3 Susceptibility

The susceptibility $\chi(G)$ of a graph $G$ (deterministic or random) is defined as the expected size of the component containing a random vertex. If the list of component sizes is $s_1, s_2, \ldots, s_r$, then

$$\chi(G) = \sum_{i=1}^{r} \frac{s_i}{n} s_i = n^{-1} \sum_{i=1}^{r} s_i^2.$$

Without loss of generality, we may assume that $s_1 \geq s_2 \geq \ldots \geq s_r$.

For the supercritical case, one can show that the giant component will dominate all other terms in the sum and so a.a.s. $\chi(\mathbb{G}(n,m)) \sim s_1^2/n \sim 4\lambda^2 n^{1/3}$ (see Appendix A in [13]). Similarly, for the critical phase, there are several components of order $n^{2/3}$ but a.a.s. $\chi(\mathbb{G}(n,m)) = \Theta(s_1^2/n) = \Theta(n^{1/3})$ (see Appendix B in [13]). The biggest challenge is to analyze the subcritical phase and this is the main focus of [13], where the following is proved (see Theorem 1.1).

**Lemma 4.5.** *If $n/2 - m \gg n^{2/3}$, then a.a.s.*

$$\chi(\mathbb{G}(n,m)) \sim \frac{n/2}{n/2 - m}.$$

## 5 Concentration tools

Let us start this section with the following result which is a generalization of a well-known Chernoff bound.

**Lemma 5.1.** *Let $\mathcal{C} = (c_1, c_2, \ldots, c_r)$ be a sequence of natural numbers with $c = \max_i c_i$. Let $S_j = \sum_{i=1}^{j} c_i Z_i$, where $Z_i, i \in [r]$ are independent Bernoulli(p) random variables. Let $\mu_j = \mathbb{E}[S_j] = p\sum_{i=1}^{j} c_i$, and let $\mu = \mu_r = \mathbb{E}[S_r]$. Then for $t \geq 0$ we have that*

$$\begin{aligned} \mathbb{P}\left(\max_{1\leq j\leq r}(S_j - \mu_j) \geq t\right) &\leq \exp\left(-\frac{t^2}{2c(\mu + t/3)}\right) \text{ and} \\ \mathbb{P}\left(\max_{1\leq j\leq r}(\mu_j - S_j) \geq t\right) &\leq \exp\left(-\frac{t^2}{2c\mu}\right). \end{aligned}$$

*In particular, for $\varepsilon \leq 3/2$ we have that*

$$\mathbb{P}\left(\max_{1\leq j\leq r}|S_j - \mu_j| \geq \varepsilon\mu\right) \leq 2\exp\left(-\frac{\varepsilon^2\mu}{3c}\right).$$

To prove this lemma, one can easily adjust the proof of the classic Chernoff bound. Alternatively, the same bounds come from [16]. In that paper, the counterpart of Lemma 5.1 is stated for $S_r - \mu$ (see Theorem 2.3); however, the author comments that $S_r - \mu$ can be replaced with $\max_{1\leq j\leq r}(S_j - \mu_j)$ (which is a slightly stronger version than we need here) as follows. A standard martingale bound shows that eg. for any $h > 0$:

$$\mathbb{P}\left(\max_{1\leq j\leq r}(S_j - \mu_j) \geq t\right) \leq e^{-ht}\mathbb{E}\left[e^{h(S_r - \mu_r)}\right].$$

Then plugging this into the appropriate place in the proof of Theorem 2.3 yields the desired bounds.

## 5.1 A rich-get-richer process

Understanding the following process will be crucial in our analysis. Let $x$ and $y$ be any natural numbers (typically $x = x(n)$, $y = y(n)$, and other values defined here are functions of $n$ and tend to infinity as $n \to \infty$). Let $\mathcal{C} = (c_1, c_2, \ldots, c_r)$ be a sequence of natural numbers and let $c = \max_i c_i$. Finally, for any $0 \le q \le r$, let $t_q = x + y + \sum_{i=1}^{q} c_i$. Clearly, $t_q$ is an increasing sequence with $t_0 = x + y$.

We define the $(\mathcal{C}, x, y)$-*process* as follows. The process starts with $X(t_0) = x$ and $Y(t_0) = y$. For any $1 \le q \le r$, with probability $p_q$ where

$$p_q := \frac{X(t_{q-1})}{X(t_{q-1}) + Y(t_{q-1})}, \tag{1}$$

the two random variables are updated as follows:

$$\begin{aligned} X(t_q) &= X(t_{q-1}) + c_q \\ Y(t_q) &= Y(t_{q-1}); \end{aligned}$$

otherwise,

$$\begin{aligned} X(t_q) &= X(t_{q-1}) \\ Y(t_q) &= Y(t_{q-1}) + c_q. \end{aligned}$$

Note that for any $0 \le q \le r$ we have $X(t_q) + Y(t_q) = t_q$.

In expectation, the ratio $p_q$ does not change throughout the process. Indeed,

$$\begin{aligned} \mathbb{E}[p_{q+1} \mid p_q] &= p_q \frac{X(t_{q-1}) + c_q}{X(t_{q-1}) + Y(t_{q-1}) + c_q} + (1 - p_q) \frac{X(t_{q-1})}{X(t_{q-1}) + Y(t_{q-1}) + c_q} \\ &= \frac{p_q c_q + X(t_{q-1})}{X(t_{q-1}) + Y(t_{q-1}) + c_q} \\ &= \frac{p_q(c_q + X(t_{q-1}) + Y(t_{q-1}))}{X(t_{q-1}) + Y(t_{q-1}) + c_q} \qquad \text{by (1)} \\ &= p_q. \end{aligned}$$

The following lemma shows that this ratio is concentrated around $p_1 = x/(x+y)$.

**Lemma 5.2.** *Consider the $(\mathcal{C}, x, y)$-process for some sequence $\mathcal{C}$ and natural numbers $x, y$ such that $\sum_{i=1}^{r} c_i < \frac{1}{2}(x+y)$. Then, for any $w \ge 1$,*

$$\mathbb{P}\left(\frac{X(t_r)}{t_r} > \frac{x}{x+y}\left(1 + \frac{1}{w}\right)\right) \le \mathbb{P}\left(X(t_r) > \frac{x}{x+y} t_r + \frac{x}{w}\right) \le \exp\left(-\frac{x}{12cw^2}\right).$$

*Proof.* Note that if $X(t_{q-1}) \le \frac{x}{x+y} t_{q-1} + \frac{x}{w}$ then

$$\begin{aligned} p_q &= \frac{X(t_{q-1})}{X(t_{q-1}) + Y(t_{q-1})} \\ &\le \frac{x(t_{q-1}/t_0 + 1/w)}{t_{q-1}} = \frac{x}{t_0}\left(1 + \frac{t_0/t_{q-1}}{w}\right) \le \frac{x}{t_0}\left(1 + \frac{1}{w}\right) =: p. \end{aligned} \tag{2}$$

We define $Z_1, \ldots, Z_r$ to be independent Bernoulli$(p)$ variables. Let $X'(t_0) = X(t_0) = x$. For each $1 \le q \le r$ we define $X'(t_q) = X(t_q)$ if $X(t_{q'}) \le \frac{x}{x+y} t_{q'} + \frac{x}{w}$ for every $q' < q$ and otherwise $X'(t_q) = X'(t_{q-1}) + c_q Z_q$. Thus, defining $S_q = \sum_{i=1}^{q} c_i Z_i$, (2) implies that we can couple the process with $S_q$ so that $X'(t_q) \le x + S_q$ for all $0 \le q \le r$.

So we can apply Lemma 5.1 to bound the probability that $X'$ ever deviates much from its mean. Noting that $\mu_r = p(t_r - t_0) \le pt_0/2 = (x/2)(1 + 1/w) \le x$ since $w \ge 1$, and setting $t = x/(2w) \le x/2$ we get that

$$\begin{aligned}\mathbb{P}\left(\max_{1\le q\le r}(S_q - \mu_q) > \frac{x}{2w}\right) &\le \exp\left(-\frac{t^2}{2c(\mu_r + t/3)}\right)\\ &\le \exp\left(-\frac{(x/(2w))^2}{2c(x + (x/2)/3)}\right) = \exp\left(-\frac{x}{12cw^2}\right). \qquad (3)\end{aligned}$$

This implies the lemma since if $Q$ is the smallest $q \le r$ for which $X(t_q) > \frac{x}{x+y}t_q + \frac{x}{w}$ then $X'(t_q) = X(t_q)$ for all $q \le Q$ by definition and so

$$S_Q \ge X'(t_Q) - x > \frac{x}{t_0}(t_Q - t_0) + \frac{x}{w} = p(t_Q - t_0) + \frac{x}{w}\cdot\frac{2t_0 - t_Q}{t_0} \ge p(t_Q - t_0) + \frac{x}{2w},$$

since $t_Q \le t_r \le \frac{3}{2}t_0$. □

We finish this subsection with the following result.

**Lemma 5.3.** *Suppose that $x \le x'$ for some $x'$ such that $c \ll x' \ll y$. Then, at the end of the $(\mathcal{C}, x, y)$-process described above, with probability at least $1 - \exp(-x'/(20c))$,*

$$X(t_r) = O(x't_r/y).$$

Before we move to the proof of the lemma, note that we can assume that $x = x'$. Clearly, running $(\mathcal{C}, x', y)$ rather than $(\mathcal{C}, x, y)$ only decreases the probability that $X \le Z$ for any given $Z$; in particular, it decreases the probability that $X(t_r) = O(x't_r/y)$. So if the lemma holds for $x'$ then it holds for $x$.

*Proof.* Lemma 5.2 requires that $\sum_{i=1}^{r} c_i \le (x+y)/2$. In order to apply this lemma, we split the process into phases. To simplify the notation, set $r_0 = 0$. For the first phase we take the longest sub-sequence $\mathcal{C}_1 = (c_1, c_2, \dots, c_{r_1})$ such that $\sum_{i=1}^{r_1} c_i \le (x+y)/2$. Next, we pick the longest subsequence $\mathcal{C}_2 = (c_{r_1+1}, c_{r_1+2}, \dots, c_{r_2})$ such that $\sum_{i=r_1+1}^{r_2} c_i \le \frac{1}{2}(x+y+\sum_{i=1}^{r_1} c_i)$, and so on, for each $j$ picking the longest subsequence $\mathcal{C}_j = (c_{r_{j-1}+1}, c_{r_{j-1}+2}, \dots, c_{r_j})$ such that

$$\sum_{i=r_{j-1}+1}^{r_j} c_i \le \frac{1}{2}\left(x + y + \sum_{i=1}^{r_{j-1}} c_i\right). \qquad (4)$$

The last phase, phase $\ell$, deals with the sequence $\mathcal{C}_\ell = (c_{r_{\ell-1}+1}, c_{r_{\ell-1}+2}, \dots, c_{r_\ell} = c_r)$. Now the $(\mathcal{C}, x, y)$-process can be treated as a series of $\ell$ processes, each on the sequence $\mathcal{C}_j$ and with initial values taken from the end of the previous sequence; i.e. phase $j$ is the $(\mathcal{C}_j, X(t_{r_{j-1}}), Y(t_{r_{j-1}}))$-process, where $X(t_0) := x, Y(t_0) := y$.

Recall that for every $q$, $X(t_q) + Y(t_q) = x + y + \sum_{i=1}^{q} c_i$. So (4) implies that we can apply Lemma 5.2 to each phase. For any $j \in \mathbb{N}$, let $w_j = 1.1^j$, and for any $j \in \mathbb{N} \cup \{0\}$, let

$$E_j = \prod_{i=1}^{j}\left(1 + \frac{1}{w_i}\right) = \Theta(1),$$

since

$$1 \le \prod_{i=1}^{j}\left(1 + \frac{1}{w_i}\right) \le \exp\left(\sum_{i=1}^{\infty} 1/w_i\right) = e^{10}.$$

(In particular, $E_0 = 1$.) We will prove that with probability at least $1 - \exp(-x/(20c))$, at the end of every phase $j$ we have:

$$X(t_{r_j}) \le \frac{x}{x+y}E_j t_{r_j}. \qquad (5)$$

If (5) holds at the end of phase $j-1$ (or if $j=1$; note that (5) trivially holds for $j=0$), then the probability that (5) holds at the end of phase $j$ is at least the probability that it holds if we adjust the initial values to

$$x_j := \frac{x}{x+y}E_{j-1}t_{r_{j-1}}; \qquad y_j := t_{r_{j-1}} - x_j.$$

So we get a lower bound on the probability of (5) by applying Lemma 5.2 to the $(\mathcal{C}_j, x_j, y_j)$-process.

We require two bounds. For the first one, recall that by definition $t_q = x+y+\sum_{i=1}^{q} c_i$. Hence, $t_{r_0} = t_0 = x+y$ and, since we choose the longest $\mathcal{C}_j$ satisfying (4) and each $c_i \le c \ll x+y \le t_{r_{j-1}}$, for each $1 \le j < \ell$ we have

$$t_{r_j} = t_{r_{j-1}} + \sum_{i=r_{j-1}+1}^{r_j} c_i > 1.5t_{r_{j-1}} - c = (1.5-o(1))t_{r_{j-1}} > 1.4t_{r_{j-1}} > 1.4^j(x+y).$$

Therefore,

$$x_j > x(1.4)^{j-1}E_{j-1} \ge x(1.4)^{j-1} \qquad \text{for any } 1 \le j \le \ell$$

and, since $c \ll x$,

$$\frac{x}{17c}\left((1.1)^j - 1\right) > j.$$

Note that, since $x_j + y_j = t_{r_{j-1}}$,

$$\frac{x_j}{x_j+y_j}\left(1+\frac{1}{w_j}\right)t_{r_j} = \frac{x}{x+y}E_{j-1}t_{r_{j-1}} \times \frac{1}{t_{r_{j-1}}} \times \left(1+\frac{1}{w_j}\right)t_{r_j} = \frac{x}{x+y}E_j t_{r_j}.$$

So Lemma 5.2 (applied with $x = x_j$, $y = y_j$, and $t_r = t_{r_j}$; $c$ remains the same for all applications of the lemma) yields that the probability that (5) fails to hold at the end of phase $j$ is at most

$$\exp\left(-\frac{x_j}{12cw_j^2}\right) \le \exp\left(-\frac{x1.4^{j-1}}{12c1.1^{2j}}\right) \le \exp\left(-\frac{x1.1^j}{17c}\right) \le 2^{-j}\exp(-x/(17c)).$$

Hence, since $\sum_{j\ge 0} 2^{-j}\exp(-x/(17c)) \le 2\exp(-x/(17c)) \le \exp(-x/(20c)) \to 0$, with the desired probability (5) holds at the end of every phase. Since the last phase ends at $r_\ell = r$ and $E_\ell = \Theta(1)$, this implies

$$X(t_r) = X(t_{r_\ell}) \le \frac{x}{x+y}E_\ell t_{r_\ell} = \Theta\left(\frac{xt_r}{x+y}\right) = O\left(\frac{xt_r}{y}\right),$$

as $x \ll y$. □

# 6 The giant has enough time to be born: $k \ll n^{1/3}$

Suppose that $k \ll n^{1/3}$. As mentioned earlier, we will prove that for this range of the parameter $k$, the giant component is formed before collisions start affecting the process. In particular, when the first special vertex joins the giant its size is much larger than the total size of all other special components—see Lemma 6.2. As a result, the giant will continue growing and at the end of the process it will have size $n - o(n)$—see Theorem 2.1(a).

## 6.1 Early phase

Let $\omega = \omega(n)$ be any function that grows with $n$ sufficiently slowly to satisfy various bounds that follow. In particular, it will grow more slowly than $n^{1/3}/k$; we may then assume that $\omega^2 \le n^{1/3}/k$.

It will be also convenient to assume that $k = k(n)$ tends to infinity faster than $\omega$ so let us assume for now that $k \ge \omega^2$; we will discuss how to translate the results to other values of $k$ (including the case when $k$ is a constant) at the end of this section. Define:

$$\begin{aligned} \lambda_1 &= \lambda_1(n) = n^{1/3}/(k\omega) \ll n^{1/3}/k \\ m_1 &= (n/2)(1 + \lambda_1 n^{-1/3}). \end{aligned}$$

So $m_1 \sim n/2$ and $\mathbb{G}(n, m_1)$ is in the supercritical phase (note that $\omega$ tends to infinity slowly enough so that $\lambda_1$ tends to infinity).

Recall that for a given graph $G$ with $r$ connected components, $L_i(G)$ is the size of an $i$-th largest component ($i = 1, 2, \ldots, r$). Similarly, let $\hat{L}_i(G)$ be the size of an $i$-th largest special component ($i = 1, 2, \ldots, k$).

Let us start with the following observation.

**Lemma 6.1.** *Suppose that $\omega^2 \le k \le n^{1/3}/\omega^2$ for some $\omega = \omega(n) \to \infty$ as $n \to \infty$. Let $\lambda_1$ and $m_1$ be defined as above. Then a.a.s. the following properties hold.*

*(a)* $L_1(\hat{\mathbb{G}}_k(n, m_1)) \sim 2n/(k\omega) = o(n)$;

*(b)* $L_2(\hat{\mathbb{G}}_k(n, m_1)) = \Theta(k^2\omega^2 \log(n^{1/3}/k))$;

*(c)* $\hat{L}_1(\hat{\mathbb{G}}_k(n, m_1)) = O(k^2\omega^2 \log(n^{1/3}/k))$.

*Proof.* The proof follows easily from Lemma 4.4 applied to $\mathbb{G}(n, m_1)$. We get that a.a.s. the largest component of $\mathbb{G}(n, m_1)$ has size asymptotic to $2\lambda_1 n^{2/3} = 2n/(k\omega) = o(n)$. Moreover, a.a.s. the size of the second largest component is of order $n^{2/3}\lambda_1^{-2} \log \lambda_1 = \Theta(k^2\omega^2 \log(n^{1/3}/k))$. Since we aim for a statement that holds a.a.s. we may assume that $\mathbb{G}(n, m_1)$ has these properties. Now, using the coupling between $\mathbb{G}(n, m_1)$ and $\hat{\mathbb{G}}_k(n, m_1)$, we select $k$ special vertices at random to translate these observations to $\hat{\mathbb{G}}_k(n, m_1)$, as described in Remark 4.1 in Section 4. The expected number of special vertices that belong to the largest component is asymptotic to $k \cdot \frac{2n}{k\omega} \cdot \frac{1}{n} = o(1)$, so a.a.s. no special vertex appears in that component. Finally, note that if at most one special vertex appears in a given component of $\mathbb{G}(n, m_1)$, then that component remains unaffected in $\hat{\mathbb{G}}_k(n, m_1)$. On the other hand, if $\ell \ge 2$ special vertices appear in a component, then it is split into $\ell$ components by deleting some of its edges. This implies (a) and (c). The same argument shows that a.a.s. no special vertex belongs to the second largest component which implies (b). The proof of the lemma is finished. □

So we can assume that $\hat{\mathbb{G}}_k(n, m_1)$ satisfies the properties of Lemma 6.1. In particular, the largest component of $\hat{\mathbb{G}}_k(n, m_1)$ contains no special vertices and so it is identical to the largest component of $\mathbb{G}(n, m_1)$ under the coupling described in Section 4. Since $\hat{\mathbb{G}}_k(n, \binom{n}{2})$ has exactly $k$ components, each with one special vertex, there will be a step of the $k$-process when the largest component is joined to a component containing a special vertex; we define this step as:

$m_2$ is the first step following $m_1$
in which the largest component of $\mathbb{G}(n, m_2)$ contains a special vertex.

It is worth noting that a.a.s. $m_2 \sim n/2$. One can show easily, using Remark 4.1, that a.a.s. the largest component of $\mathbb{G}(n, m = (n/2)(1 + \lambda_2 n^{-1/3}))$ contains at least one special vertex if

$$\lambda_2 = \lambda_2(n) = n^{1/3}\omega/k \gg n^{1/3}/k.$$

Indeed, by Lemma 4.4, a.a.s. the largest component has size $(2 + o(1))\lambda_2 = (2 + o(1))n\omega/k$ and so, conditioning on this event, the probability that it contains no special vertex is equal to

$$\begin{aligned} \frac{\binom{n-(2+o(1))n\omega/k}{k}}{\binom{n}{k}} &\leq \left(\frac{n-(2+o(1))n\omega/k}{n}\right)^k = \left(1 - \frac{(2+o(1))\omega}{k}\right)^k \\ &\leq \exp\Big(-(2+o(1))\omega\Big) = o(1). \end{aligned}$$

Therefore, a.a.s. $m_2 \leq (n/2)(1 + \lambda_2 n^{-1/3})$. Hence, a.a.s. $m_2 \sim n/2$ (since it is assumed that $k \geq \omega^2$).

**Lemma 6.2.** *Suppose that $\omega^2 \leq k \leq n^{1/3}/\omega^2$ for some $\omega = \omega(n) \to \infty$ as $n \to \infty$. Let $m_2$ be defined as above. Then, a.a.s. the following properties hold*

*(a)* $\hat{L}_1(\hat{\mathbb{G}}_k(n, m_2)) = L_1(\hat{\mathbb{G}}_k(n, m_2)) \geq n/(k\omega)$;

*(b)* $\sum_{i=2}^{k} \hat{L}_i(\hat{\mathbb{G}}_k(n, m_2)) \leq k^2\omega^3 \log(n^{1/3}/k)$.

*Moreover, for every $m \geq m_2$:*

*(c) the size of any non-special component in $\hat{\mathbb{G}}_k(n, m)$ is at most $k^2\omega^3 \log(n^{1/3}/k)$.*

*Proof.* It follows from Lemma 6.1(a) that the giant component of $\hat{\mathbb{G}}_k(n, m_1)$ has size at least $n/(k\omega)$. It keeps growing from that point on and so the same lower bound holds at time $m_2$. Property (a) holds. Moreover, using Lemma 4.4 and the discussion right after the lemma, not only at time $m_1$ (as indicated by Lemma 6.1(b)) but also if one continues the random graph process from time $m_1$ on, a.a.s. the size of the second largest component of $\mathbb{G}(n, m)$ is always at most $\Theta(k^2\omega^2 \log(n^{1/3}/k)) < k^2\omega^3 \log(n^{1/3}/k)$. Since any non-special component in $\hat{\mathbb{G}}_k(n, m)$ is a component in $\mathbb{G}(n, m)$, this proves property (c).

Now, in order to show that property (b) holds, we must study the subgraph induced by the vertices not in the largest component. We define:

$$\mathbb{G}^L(n, m) \text{ is the graph obtained by deleting the largest component from } \mathbb{G}(n, m).$$

This is of particular interest when $m$ is in the supercritical range. We define:

$$\begin{aligned} n' &= \text{the number of vertices in } \mathbb{G}^L(n, m_2); \\ m' &= \text{the number of edges in } \mathbb{G}^L(n, m_2); \\ k' &= \text{the number of special vertices in } \mathbb{G}^L(n, m_2). \end{aligned}$$

By Lemma 4.4, after setting $m_2 = n/2 + \lambda_2 n^{2/3}/2$, we have a.a.s.

$$\begin{aligned} n' &= n - (2 + o(1))\lambda_2 n^{2/3} \\ &= n - (4 + o(1))(m_2 - n/2); \\ m' &= m_2 - (2 + o(1))\lambda_2 n^{2/3} \\ &= n/2 - (3/2 + o(1))\lambda_2 n^{2/3} \\ &= n'/2 - (1/2 + o(1))\lambda_2 n^{2/3} \\ &= n'/2 - (1 + o(1))(m_2 - n/2); \end{aligned} \tag{6}$$

and the component of $\mathbb{G}(n, m_2 - 1)$ that is added to the largest component of $\mathbb{G}(n, m_2 - 1)$ in step $m_2$ contains exactly $k - k' \geq 1$ special vertices.

Conveniently, the distribution of $\mathbb{G}^L(n, m_2)$ and of its $k'$ special vertices is nearly uniform, despite the conditioning implied by the definition of $m_2$. Formally, we need the following claim.

**Claim:** *Expose the values of $m_2, k'$, and the largest component of $\mathbb{G}(n, m_2)$; denote that largest component by $\Theta^*$. (Note that this determines the vertex set of $\mathbb{G}^L(n, m_2)$ and $m'$.) Conditional on that exposure:*

*(i) Every graph on the $n'$ vertices of $\mathbb{G}^L(n, m_2)$ that has $m'$ edges and no component at least as large as $\Theta^*$ is equally likely to be $\mathbb{G}^L(n, m_2)$.*

*(ii) Every set of $k'$ vertices in $\mathbb{G}^L(n, m_2)$ is equally likely to be the special vertices.*

**Proof of the claim:** Consider (i) any set $S$ of $k$ special vertices where exactly $k'$ are outside of $\Theta^*$, and (ii) any random graph process $e_1, \ldots, e_{m_2}$ in which $m_2$ is the first step following $m_1$ where the largest component contains a member of $S$ and $\Theta^*$ is that largest component. Let $L$ be the graph formed by removing $\Theta^*$.

Let $L'$ be any graph on the same vertex set as $L$ with $m'$ edges and with no component larger than $\Theta^*$. Replace the edges of $L$ in the process with the edges of $L'$, in any order; let $G'_i$ be the graph formed by the first $i$ edges of the resulting sequence. Replace the $k'$ special vertices in $L$ by any set of $k'$ vertices in $V(L)$, and do not change the $k - k'$ special vertices in $\Theta^*$; denote the resulting set of $k$ special vertices as $S'$. It is straightforward to check that (1) the largest component of $G'_{m_2-1}$ contains no vertex of $S'$, and (2) the largest component of $G'_{m_2}$ is $\Theta^*$ and hence contains a vertex of $S'$. So $m_2, k'$ and the largest component at step $m_2$ are the same in both processes. Furthermore, each sequence of edges is equally likely to be selected. This implies the claim.

By part (ii) of our claim, and arguing as in Remark 4.1, we can first expose the graph $\mathbb{G}^L(n, m_2)$ and then choose the $k'$ special vertices from that graph by selecting a set of size $k'$ uniformly at random from all sets of that size. We will do it in the following way. We independently select $k'$ vertices uniformly at random, one by one, allowing potential repetitions. Let $E$ be the event that all selected vertices are unique. For each of those vertices, the expected size of the component containing it is $\chi(\mathbb{G}^L(n, m_2))$. Hence, the expected total size of the components containing those vertices is at most $k'\chi(\mathbb{G}^L(n, m_2))$ (note that some of them may end up in the same component). By part (i), we can treat $\mathbb{G}^L(n, m_2)$ as $\mathbb{G}(n', m')$ which, by Lemma 4.5 and (6) a.a.s. has susceptibility:

$$\chi(\mathbb{G}^L(n, m_2)) \sim \frac{n'/2}{n'/2 - m'} \sim \frac{n/2}{n'/2 - m'} \sim k\omega.$$

So the expected total size of the components of $\mathbb{G}^L(n, m_2)$ containing selected vertices is at most $k' \times (1 + o(1))k\omega \leq (1 + o(1))k^2\omega$. Let $Q$ be the event that the total size is at most $k^2\omega^2$. It follows from Markov's inequality that $P(Q) = 1 - o(1)$. Finally, note that the probability of no repetition during the selection process is equal to

$$P(E) = \prod_{i=1}^{k'-1} \left(1 - \frac{i}{n}\right) = \prod_{i=1}^{k'-1} \exp\left(-\frac{i}{n} + O(i^2/n^2)\right) = \exp\left(-O(k'^2/n)\right) = 1 - o(1),$$

since $k' \leq k \ll n^{1/3}$. Hence, a.a.s. the selected vertices form a set of size $k'$ taken uniformly at random from all sets of that size. Since

$$P(Q|E) = \frac{P(Q \cap E)}{P(E)} \geq \frac{1 - P(Q^c) - P(E^c)}{1 - P(E^c)} = \frac{1 - o(1)}{1 - o(1)} = 1 - o(1),$$

we conclude that a.a.s. the total size of the components of $\mathbb{G}^L(n, m_2)$ containing special vertices is at most $k^2\omega^2$.

This bounds the total size of all special components, other than the largest, in $\mathbb{G}(n, m_2)$. But we actually need to bound the total size in $\hat{\mathbb{G}}_k(n, m_2)$. If $k' = k - 1$, i.e. if only one special vertex joins the largest component in step $m_2$, then these two totals are the same. Otherwise, let $\Phi$ be the component of $\mathbb{G}(n, m_2 - 1)$ that contains $k - k'$ special vertices and is merged with the largest component in step $m_2$. In $\hat{\mathbb{G}}_k(n, m_2 - 1)$, $\Phi$ is partitioned into exactly $k - k'$ components. One of them is joined to the largest component in step $m_2$; the others have total size at most $|\Phi|$ which, by Lemma 4.4(b) and since $m_2 > m_1$, is at most $\Theta(n^{2/3}\lambda_1^{-2}\log\lambda_1) = \Theta(\omega^2k^2\log(n^{1/3}/k\omega))$. $\hat{L}_2(\hat{\mathbb{G}}_k(n, m_2)), \ldots, \hat{L}_k(\hat{\mathbb{G}}_k(n, m_2))$ consists of those $k - k' - 1$ components, along with the $k'$ special components contained in the components of $\mathbb{G}^L(n, m_2)$ that contain special vertices. The total size of the latter set was bounded above, and so

$$\sum_{i=2}^{k} \hat{L}_i(\hat{\mathbb{G}}_k(n, m_2)) \le k^2\omega^2 + \Theta(\omega^2k^2\log(n^{1/3}/k\omega)) < \omega^3k^2\log(n^{1/3}/k),$$

thus proving part (b). □

## 6.2 Modelling $\hat{\mathbb{G}}_k(n, m)$ with a $(\mathcal{C}, x, y)$-process

We continue assuming that $\omega^2 \le k \le n^{1/3}/\omega^2$ for some $\omega = \omega(n) \to \infty$ as $n \to \infty$. Beginning at time $m = m_2$, we do not consider $\mathbb{G}(n, m)$ and instead focus directly on $\hat{\mathbb{G}}_k(n, m)$. Recall that a component is called *special* if it contains a special vertex, and so we always have exactly $k$ special components.

Let $j_1, \ldots, j_\ell > m_2$ denote the steps in the $\mathbb{G}_k$-process during which we choose an edge joining a special component to a non-special component. Of course, we accept that edge. Let $\Theta_i$ be the non-special component joining a special component at time $j_i$, and set $c_i = |\Theta_i|$.

**Observation 6.3.** *After $m = m_2$, the sizes of the special components only change during steps $j_1, \ldots, j_\ell$.*

Now expose the components $\Theta_1, \ldots, \Theta_\ell$ but not the edges selected at times $j_1, \ldots, j_\ell$.

**Observation 6.4.** *Conditional on any choice for $\Theta_1, \ldots, \Theta_\ell$, at each step $j_i$, the probability that $\Theta_i$ is joined to a particular special component is proportional to the size of that special component.*

So we can model the growth of the largest component with a $(\mathcal{C}, x, y)$ process. Let

$$y = L_1(\hat{\mathbb{G}}_k(n, m_2)) = \hat{L}_1(\hat{\mathbb{G}}_k(n, m_2)); \qquad x = \sum_{i=2}^{k} \hat{L}_i(\hat{\mathbb{G}}_k(n, m_2));$$

i.e. $y$ is the size of the largest special component and $x$ is the total size of all other special components at step $m_2$. Then setting $\mathcal{C} = (c_1, \ldots, c_\ell)$ we see that our two observations yield:

**Observation 6.5.** *The size of the largest special component at steps $j_1, \ldots, j_\ell$ follows random variable $Y$ in the $(\mathcal{C}, x, y)$ process.*

Note that in this process, we have $t_\ell = n$. By Lemma 6.2, we have:

$$\begin{aligned} y &\ge n/(k\omega); \\ x \le k^2\omega^3\log(n^{1/3}/k) \le x' &:= k^2\omega^4\log(n^{1/3}/k); \qquad \text{and} \\ c_i \le c &:= k^2\omega^3\log(n^{1/3}/k) \qquad \text{for every } i. \end{aligned}$$

From that, it is easily verified that $c \ll x' \ll y$. Indeed, clearly $c/x' = 1/\omega = o(1)$ and since $k \le n^{1/3}/\omega^2$, we have $x'/y \le k^3\omega^5 \log(n^{1/3}/k)/n \le 2\log\omega/\omega = o(1)$. Therefore, we can apply Lemma 5.3 to show that at the end of the process, a.a.s. the total size of all but the largest special component is

$$X(n) = O(x'n/y) = O(k^3\omega^5 \log(n^{1/3}/k)) = o(n).$$

This proves Theorem 2.1(a), provided that $\omega^2 \le k \le n^{1/3}/\omega^2$. To extend the result to smaller values of $k$, we apply Observation 4.2. Fix any $k < k' := \omega^2$. Our bound above yields that a.a.s. the largest special component at the end of the $k'$-process has size $1-O(k'^3\omega^4 \log(n^{1/3}/k')) = 1-O(k\omega^{10} \log(n^{1/3}/k))$. Observation 4.2 implies that the same bound holds for all $2 \le k < k'$, thus proving Theorem 2.1(a). (Note that in the statement of the theorem, we replaced $\omega^{10}$ by $\omega$ which is allowed as in the statement $\omega$ is any function tending to infinity, regardless how slowly it does so.)

# 7 No component has a chance to become giant: $k \gg n^{1/3}(\log n)^{4/3}(\log\log n)^{1/3}$

Suppose now that $k \gg n^{1/3}$. As mentioned earlier, for this range of parameter $k$, collisions will start affecting the process much earlier, namely, when each component has size smaller than the total size of all special components—see Lemma 7.1. Intuitively, this results in no one component dominating the process, and so no component will be able to grow to linear size—see Theorem 2.1 (b). In order to prove that this happens we require a stronger bound on $k$, namely:

$$k \gg n^{1/3}(\log n)^{4/3}(\log\log n)^{1/3}. \tag{7}$$

For technical reasons, we also require the following upper bound

$$k \ll n/\log n. \tag{8}$$

Thus we have the range of $k$ for Theorem 2.1(b).

## 7.1 Early phase

For this section, we define:

$$\begin{aligned} \lambda_3 &= \left(\frac{k}{n^{1/3}}\log(k/n^{1/3})\right)^{1/2} \\ m_3 &= (n/2)(1-\lambda_3 n^{-1/3}). \end{aligned}$$

Note that, since $n^{1/3} \ll k \ll n/\log n$, we have $\lambda_3 \to \infty$ and $m_3 = (n/2)(1-o(1))$. So we are in the subcritical phase.

**Lemma 7.1.** *Suppose that $k \gg n^{1/3}$ and $k \ll n/\log n$. Let $\lambda_3$ and $m_3$ be defined as above. Then a.a.s. the following properties hold*

*(a)* $\hat{L}_1(\hat{\mathbb{G}}_k(n,m_3)) \le L_1(\hat{\mathbb{G}}_k(n,m_3)) = \Theta(n/k) \ll n^{2/3}$;

*(b)* $\sum_{i=1}^{k} \hat{L}_i(\hat{\mathbb{G}}_k(n,m_3)) \ge \frac{1}{4}\left(\frac{nk}{\log(k/n^{1/3})}\right)^{1/2} \gg n^{2/3}$.

*Moreover, for any $m \ge m_3$:*

*(c) the size of any non-special component in $\hat{\mathbb{G}}_k(n,m_3)$ is at most $n\log n/k$.*

Since the proof of this lemma is long, we split it into a few parts.

*Proof of Lemma 7.1(a).* From Lemma 4.4 we get that a.a.s. the size of the largest component in $\mathbb{G}(n, m_3)$ is equal to

$$x = \Theta(n^{2/3}\lambda_3^{-2}\log\lambda_3) = \Theta(n/k).$$

The size of the largest component in $\hat{\mathbb{G}}_k(n, m_3)$ is at most the size of the largest component in $\mathbb{G}(n, m_3)$, so property (a) holds a.a.s. □

Before we move on to the proof of property (b), let us note that Lemma 4.5 implies that a.a.s. the susceptibility of $\mathbb{G}(n, m_3)$ is

$$\chi \sim \frac{n/2}{n/2 - (n/2 - \lambda_3 n^{2/3}/2)} = \frac{n^{1/3}}{\lambda_3} = \left(\frac{n}{k\log(k/n^{1/3})}\right)^{1/2}.$$

Hence, one might expect the total size of all special components to be close to

$$k\chi \sim \left(\frac{nk}{\log(k/n^{1/3})}\right)^{1/2}.$$

As explained in the proof of Lemma 6.2(b), this is only an upper bound for the expected total size so more work would be needed to get a lower bound. More importantly, there is a problem with showing a concentration around the expectation. Indeed, if we were only concerned with $k \gg \sqrt{n\log n}$, the concentration of this total size would follow easily from the Hoeffding-Azuma inequality for martingales. Since we need a more sophisticated argument for $k$ near $n^{1/3}$, we provide another proof covering the whole range of $k$, namely, $n^{1/3} \ll k \ll n/\log n$.

*Proof of Lemma 7.1(b).* It will be simpler to work with the binomial random graph $\mathcal{G}(n, p)$ and then translate the results back to $\mathbb{G}(n, m_3)$. Let $\gamma = \gamma(n) = \sqrt{n}\log n \gg \sqrt{n}$, and let

$$p = \frac{m_3 - \gamma}{\binom{n}{2}} = \frac{1 - \lambda_3 n^{-1/3} - 2\gamma/n}{n-1} = \frac{1 - (1+o(1))\left(k\log(k/n^{1/3})/n\right)^{1/2}}{n-1}.$$

We start with $n$ isolated vertices, $k$ of them are special and form set $K$. We will find a lower bound for the sum of the sizes of all components in $\mathcal{G}(n, p)$ containing special vertices. Lemma 4.3 will then imply that the same bound holds in $\mathbb{G}(n, m_3)$ and so also in $\hat{\mathbb{G}}_k(n, m_3)$ (as this random variable is exactly the same in both models).

Consider the *breadth-first-search* process starting from $K$. Put all vertices of $K$ into a *queue* $Q$ (first-in first-out list); in any order. Call all vertices of $K$ *saturated*, and then do the following as long as $Q$ is not empty: remove $w$ from $Q$, expose all edges from $w$ to non-saturated vertices, put all new neighbours of $w$ into $Q$ and call them saturated. Note that all saturated vertices lie in special components.

Let $t$ be the random step at which this process halts; i.e. reaches $Q = \emptyset$. For all $i \le t$ we let $S_i$ denote the number of saturated vertices at step $i$ of the process. Set

$$s = \frac{1}{4}\left(\frac{nk}{\log(k/n^{1/3})}\right)^{1/2} = \frac{1}{4}k\chi.$$

It suffices to prove that a.a.s. we will reach a step $i$ for which $S_i \ge s$. Note that at any step $i \le t$, $|S_i| \ge i$. So it suffices to prove that a.a.s. we do not have:

$$t < s \qquad \text{and} \qquad S_i \le s \qquad \forall i \le t. \tag{9}$$

Let $Z_i$ denote the random variable counting the number of vertices added into $Q$ at the $i$th step of the process. Since we remove one vertex from $Q$ at each step, the size of $Q$ at the end of step $i$ is

$$k - i + \sum_{j=1}^{i} Z_j.$$

Note that $Z_i$ has binomial distribution $\mathrm{Bin}(n-S_i, p)$. Indeed, if we let $X_1, \ldots$ be a sequence of independent Bernoulli($p$) variables, then we can couple so that for all $i \le t$ we have

$$\sum_{j=1}^{i} Z_j = \sum_{j'=1}^{\sum_{\ell=1}^{i} n-S_\ell} X_{j'}.$$

So the probability that (9) holds is at most the probability that

$$\exists i < s \text{ such that } \sum_{j'=1}^{i(n-s)} X_{j'} \le i - k. \tag{10}$$

However, since

$$\begin{aligned}
\mathbb{E}\left[\sum_{j'=1}^{i(n-s)} X_{j'}\right] &= i(n-s)p \\
&= i\left(1 - \frac{1}{4}\left(\frac{k}{n\log(k/n^{1/3})}\right)^{1/2}\right)\left(1 - (1+o(1))\left(\frac{k\log(k/n^{1/3})}{n}\right)^{1/2}\right) \\
&= i\left(1 - (1+o(1))\left(\frac{k\log(k/n^{1/3})}{n}\right)^{1/2}\right),
\end{aligned}$$

(10) would imply that

$$\begin{aligned}
\mathbb{E}\left[\sum_{j'=1}^{i(n-s)} X_{j'}\right] - \sum_{j'=1}^{i(n-s)} X_{j'} &> k - (1+o(1))i\left(\frac{k\log(k/n^{1/3})}{n}\right)^{1/2} \\
&> k - (1+o(1))s\left(\frac{k\log(k/n^{1/3})}{n}\right)^{1/2} = \left(\frac{3}{4} - o(1)\right)k \quad > \quad \frac{1}{2}k.
\end{aligned}$$

We note that $\mathbb{E}\left[\sum_{j'=1}^{s(n-s)} X_{j'}\right] < s < \sqrt{nk}$ and apply Lemma 5.1 with $c = 1$ to obtain that the probability of (10) is at most

$$\exp\left(-\frac{(k/2)^2}{2\sqrt{nk}}\right) = \exp\left(-\frac{(k^3/n)^{1/2}}{8}\right) = o(1),$$

since $k \gg n^{1/3}$. This proves part (b). □

*Proof of Lemma 7.1(c).* Set $c = n\log n/k$. Note first that any non-special component in $\hat{\mathbb{G}}_k(n,m)$ is a component in $\mathbb{G}(n,m)$. We will run the $\mathbb{G}(n,m)$ process and say that round $m$ is *dangerous* if an edge added during this round connects two components of corresponding sizes $c_1$ and $c_2$, such that $c_1 \le c$, $c_2 \le c$, but $c_1 + c_2 > c$. We say that a dangerous round is *deadly* if the component formed contains no special vertex. We need to show that a.a.s. there are no deadly rounds. Clearly, the number of dangerous rounds is at most $n/c$. To bound the probability that the $i$th dangerous round is deadly, we run the $\mathbb{G}(n,m)$ process until the $i$th dangerous round; note that the process up to this point is independent of

the choice of special vertices, so we can choose them after the $i$th dangerous round. The probability that none of the $k$ vertices are in the component of size at least $c$ formed in this round is at most

$$\frac{\binom{n-c}{k}}{\binom{n}{k}} = \frac{(n-c)_k}{(n)_k} \le \left(1-\frac{c}{n}\right)^k \le \exp\left(-\frac{ck}{n}\right) = \frac{1}{n}.$$

So the expected number of deadly rounds is at most $\frac{n}{c}\cdot\frac{1}{n} = o(1)$ and so a.a.s. no dangerous round is deadly which finishes the proof of this property. □

### 7.2 Late phase

We still assume that $k \ll n/\log n$ so that Lemma 7.1 can be applied. We continue the $\hat{\mathbb{G}}_k(n,m)$ process from time $m_3$ on. We model it with a $(\mathcal{C},x,y)$-process as in Subsection 6.2. Again, we define $c_1,\dots,c_\ell$ to be the sizes of the non-special components that are joined to special components after step $m_3$. By Lemma 7.1(c), every $c_i \le c := n\log n/k$.

Let $v$ be any of the $k$ special vertices. We will let $X$ count the size of the component containing $v$, and we let $Y$ count the total size of the other $k-1$ special components. By Lemma 7.1(b,c), initially (i.e., at step $m_3$) we have

$$X \le x = \Theta(n/k) \le x' := 30n(\log n)^2/k; \qquad Y \ge y := \Theta((nk/\log(k/n^{1/3}))^{1/2}).$$

(Note that we used a loose upper bound for $X$ to make some room for an argument below that gives the desired upper bound for the failure probability.) Using the fact that $k \gg n^{1/3}(\log n)^{4/3}(\log\log n)^{1/3}$ we get $c \ll x' \ll y$. So Lemma 5.3 implies that at the end of the process, with probability at least $1-\exp(-x/(20c)) = 1-o(n^{-1})$, we have

$$\begin{aligned} X &= O(xt_\ell/y) = O(xn/y) = O\left(\frac{n(\log n)^2/k}{(nk/\log(k/n^{1/3}))^{1/2}}n\right) \\ &= O\left(\left(\frac{n(\log n)^4\log(k/n^{1/3})}{k^3}\right)^{1/2}n\right) = o(n). \end{aligned}$$

Multiplying by the $k$ choices for $v$, with probability at least $1-o(1)$, at the end of the process every special component has size $o(n)$. This completes the proof of Theorem 2.1(b).

### 7.3 Extending the argument for large values of $k$

Until now, we have assumed that $k \ll n/\log n$. To extend to higher values of $k$, we apply Observation 4.2. For any $\omega\to\infty$ with $n$, set $k' = n/(\omega\log n)$. Our bound above yields that a.a.s. the largest special component at the end of the process has size $O(\omega^{3/2}\log^4 n)$. Observation 4.2 says that the same bound holds for all $k \ge k'$, thus proving Theorem 2.1(c).

## 8 Concluding Remarks

Note that in Theorem 2.1(b) we needed to assume that $k \gg n^{1/3}(\log n)^{4/3}(\log\log n)^{1/3}$. This seems to be an artifact of the proof technique we use (the union bound over all special components) rather than the lower bound that is needed. It is natural to conjecture that a.a.s. $L_1(\mathbb{P}_k(n,M)) = o(n)$ even for $k \gg n^{1/3}$. Indeed, if $k = n^{1/3}\omega$ for any $\omega = \omega(n)\to\infty$, one can show (for example, using the argument as in the proof of Lemma 7.1(b)) that in $\hat{\mathbb{G}}_k(n,n/2)$, a.a.s. the total size of all special components is of order $n^{2/3}\sqrt{\omega}$. From the observations in Section 4.2 we know that a.a.s. the giant component has size at most $n^{2/3}\omega^{1/4}$ (in fact, of order $n^{2/3}$; as usual, we make some room for the argument to work),

and the largest component that appears after time $n/2$ is of order $n^{2/3}$. By Lemma 5.3, we get that a.a.s. the largest special component at time $m = n/2$ grows only to size $o(n)$. This supports the conjecture but it is not clear how to avoid using the union bound and so it remains an open problem.

# 9 Acknowledgments

We would like to thank Megan Dewar and John Proos from the Tutte Institute for Mathematics and Computing for stimulating discussions on the problem and its applications. We would also like to thank anonymous reviewers for their comments that greatly improved the quality of the paper.

# References

[1] D. Aldous. A random tree model associated with random graphs. *Random Structures and Algorithms*, 1:383–402, 1990.

[2] Tom Bohman. The triangle-free process. *Advances in Mathematics*, 221:1653–1677, 2009.

[3] Tom Bohman and Peter Keevash. Dynamic concentration of the triangle-free process. *Random Structures & Algorithms*, 58(2):221–293, 2021.

[4] B. Bollobás and O. Riordan. Constrained graph processes. *Electronic Journal of Combinatorics*, 7(1):R18, 2000.

[5] Béla Bollobás. *Random Graphs*. Cambridge University Press, 2001.

[6] R.W.R. Darling, David G. Harris, Dev R. Phulara, and John A. Proos. The combinatorial data fusion problem in conflicted-supervised learning. *Arxiv preprint*, 2019.

[7] P. Erdős and A. Rényi. On random graphs. i. *Publicationes Mathematicae*, 6:290–297, 1959.

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

[9] G. Fiz Pontiveros, S. Griffiths, and R. Morris. The triangle-free process and the ramsey number $r(3, k)$. *Mem. Amer. Math. Soc., to appear*, 2019.

[10] Alan Frieze and Michał Karoński. *Introduction to random graphs*. Cambridge University Press, 2015.

[11] M. Garey and D. Johnson. *Computers and Intractability: A Guide to the Theory of NP-Completeness*. W.H. Freeman and Company, 1979.

[12] Stefanie Gerke, Dirk Schlatter, Angelika Steger, and Anusch Taraz. The random planar graph process. *Random Structures & Algorithms*, 32(2):236–261, 2008.

[13] Svante Janson and Malwina Luczak. Susceptibility in subcritical random graphs. *Journal of Mathematical Physics*, 49(12):125207, 2008.

[14] Svante Janson, Tomasz Łuczak, and Andrzej Ruciński. *Random graphs*. John Wiley & Sons, 2000.

[15] Tomasz Łuczak. Component behavior near the critical point of the random graph process. *Random Structures and Algorithms*, 1:287–310, 1990.

[16] Colin McDiarmid. *Concentration*, pages 195–248. Springer Berlin, 1998.

[17] D. Osthus and A. Taraz. Random maximal $h$-free graphs. *Random Structures and Algorithms*, 18:61–82, 2001.

[18] Andrzej Ruciński and Nick Wormald. Random graph processes with degree restrictions. *Combinatorics, Probability and Computing*, 221:169–180, 1992.

[19] Vijay Vazirani. *Approximation Algorithms*. Springer-Verlag, Berlin and Heidelberg, 2003.