The phase transition in inhomogeneous random intersection graphs

Milan Bradonjić* Aric Hagberg† Nicolas W. Hengartner‡
Nathan Lemons§ Allon G. Percus¶

January 31, 2013

Abstract

We analyze the component evolution in inhomogeneous random intersection graphs when the average degree is close to 1. As the average degree increases, the size of the largest component in the random intersection graph goes through a phase transition. We give bounds on the size of the largest components before and after this transition. We also prove that the largest component after the transition is unique. These results are similar to the phase transition in Erdős-Rényi random graphs; one notable difference is that the jump in the size of the largest component varies in size depending on the parameters of the random intersection graph.

Keywords: Random intersection graphs, random graphs, giant component, phase transition, branching process.

*Mathematics of Networks and Communications, Bell Labs, Alcatel-Lucent, 600 Mountain Avenue, Murray Hill, New Jersey 07974, USA; milan@research.bell-labs.com. Research supported in part by NIST grant 60NANB10D128. Part of this work was done at Los Alamos National Laboratory.

†Center for Nonlinear Studies and Theoretical Division, Los Alamos National Laboratory, Los Alamos, NM 87545, USA; hagberg@lanl.gov.

‡Information Sciences Group, Los Alamos National Laboratory, Los Alamos, NM 87545, USA; nickh@lanl.gov.

§Center for Nonlinear Studies and Theoretical Division, Los Alamos National Laboratory, Los Alamos, NM 87545, USA; nlemons@lanl.gov.

¶School of Mathematical Sciences, Claremont Graduate University, Claremont, CA 91711, USA; allon.percus@cgu.edu.

1 Introduction

The well-studied Erdős-Rényi graph, $G(n,p)$, is a basic model for random networks that is amenable to structural analysis. However, $G(n,p)$ is not suited as a model for real-world networks; perhaps the most common criticism is that sparse realizations of $G(n,p)$ do not exhibit clustering [11]. Thus $G(n,p)$ is not a good model for most social networks which are usually sparse and have nontrivial clustering. In many cases this phenomenon (sparsity together with clustering) is a result of the graph originating as the intersection graph of a larger bipartite graph. For example, the well-known collaboration graphs of scientists (or of movie actors) is derived from the bipartite graph of scientists and papers (respectively, actors and movies) [24, 18].

A simple natural model for such networks is the random intersection graph. Random intersection graphs were introduced by Karoński, Scheinerman and Singer-Cohen [23, 15] and have recently attracted much attention [4, 8, 14, 10, 22]. We study the phase transition for components in the inhomogeneous random intersection graph model defined by Nikoletseas, Raptopoulos and Spirakis [19, 20]. Let $\mathbf{p} = (p_i)_{i=1}^m$ be a sequence of m probabilities, V a set of n vertices and $A = \{a_1, a_2, \ldots, a_m\}$ a set of m attributes. For $(v, a_i) \in V \times A$, define independent indicator random variables $\mathcal{I}_{v,a_i} \equiv$ Bernoulli(p_i). A random bipartite graph B is defined on the vertices V and attributes to contain exactly those edges, (v, a_i) for which $\mathcal{I}_{v,a_i} = 1$. Finally, the random intersection graph G is obtained from the bipartite graph B by projecting onto the vertices V: two vertices are connected in G if they share at least one common attribute in B.

This paper is concerned with asymptotic results; for each n let $\mathbf{p}^{(n)}$ be a vector of $m = m(n)$ probabilities. This defines a sequence of random intersection graphs indexed by n. We say an event E_n holds *with high probability* if $\mathbb{P}[E_n] \to 1$ as $n \to \infty$. We show that depending on the sequences $\mathbf{p}^{(n)}$ one may observe larger or smaller jumps in the phase transition.

In previous work [1, 16], the phase transition was located for random intersection graphs defined with uniform probabilities. The component evolution of inhomogeneous random intersection graphs has been studied for a different model of random intersection graphs [3, 2]. The results in these papers used tools developed by Bollobás, Janson and Riordan[5] and are exact, though they only consider those cases when the giant component is linear. In [6] another general model of sparse random graphs is introduced and analyzed. Behrisch [1] studied the uniform homogeneous case when all $p_i \equiv p = c/\sqrt{nm}$

and noted that if $p = \omega(1/n)$, then the largest component jumps from size $O(np \log n)$ to $\Theta(p^{-1})$; a smaller jump than observed in Erdős Rényi random graphs. On the other hand for $p = O(1/n)$ the largest component jumps from size $O(\log n)$ to $\Theta(n)$; a jump similar to that in Erdős-Rényi random graphs. Indeed for m large enough and $p_i \equiv c/\sqrt{mn}$, the random intersection model is equivalent to $G(n, p)$[9, 21]. Our theorems show these phenomena occurring in the more general setting of inhomogeneous random intersection graphs as well.

Theorem 1.1. *If $n \sum p_i^2 < 1$ then with high probability all components in G will have size at most $O(\max\{np \log n, \log n\})$ where $p = \max p_i$.*

Note that each attribute a_i contributes a clique of expected size np_i to the random intersection graph; thus Theorem 1.1 is very close to best possible.

Theorem 1.2. *If $n \sum p_i^2 = c > 1$ and there exists a $\gamma > 1/2$ such that $\max \mathbf{p}^{(n)} = o(n^{-\gamma})$, then with high probability there exists a unique largest component. This component will have size $(1 - \rho)n$ where ρ is the unique solution in $[0, 1)$ to the equation*

$$x = \prod_{i=1}^{m} \left[1 - p_i(1 - (1 - p_i(1 - x))^n)\right]. \tag{1.1}$$

All other components will have size of order $O(\max\{np \log n, \log n\})$.

Importantly, the unique largest component guaranteed by Theorem 1.2 is not necessarily linear in n. Under the conditions of the Theorems 1.1 and 1.2, there is, however, a jump in the size of the largest component when transitioning from the subcritical phase (when $n \sum p_i^2 < 1$) to the supercritical phase (when $n \sum p_i^2 > 1$). Thus a phase transition is observed. The phase transition is made apparent in comparing Theorems 1.1 and 1.3 though the later is not necessarily best possible.

As our model is quite general, our theorems do not always give the best possible bounds. What is perhaps surprising is that despite the generality of the model, we can locate the phase transition exactly. No assumptions of uniformity nor convergence of the sequences $\mathbf{p}^{(n)}$ are necessary; we only require that $n \sum p_i^2$ be a constant. Because of this generality, there are many cases where the solutions to Equation(1.1) do not converge as $n \to \infty$. In such cases, even the order of magnitude of the unique largest component may

fluctuate. To compensate for this, we state a weaker version of Theorem 1.2 which gives a lower bound on the order of magnitude of the unique largest component. We also show, in Proposition 3.3, how to use Theorem 1.2 to derive the exact size of the largest component in the uniform (homogeneous) case. In this way we recover exactly previously proved results using our more general method [1, 16].

Theorem 1.3. *If $n\sum p_i^2 = c > 1$ and there exists a $\gamma > 1/2$ such that $p = \max p_i = o(n^{-\gamma})$, then with high probability there exists a unique giant component. If $p^{-1} = o(n)$, this component will be of size at least $\Omega(p^{-1})$. Otherwise it will be of size $\Omega(n)$. All other components will have size at most $O(\max\{np\log n, \log n\})$.*

Example. Let $m = n^\alpha$, $\alpha < 1$, and set $p_i \equiv c/\sqrt{mn}$ for some constant c. If $c < 1$ then by Theorem 1.1, each component has size at most $O(\sqrt{n/m}\log n)$. On the other hand, if $c > 1$, there exists a unique largest component whose size is $\Omega(np)$ by Theorem 1.2. See Proposition 3.3 for the derivation of the exact bound. These bounds are the same as obtained by Behrisch [1].

Example. Let $m = \beta n$, and set $p_i \equiv c/\sqrt{mn}$ for some constant c. If $c < 1$, then each component has size at most $O(\log n)$. This is the same bound as obtained in [16]. On the other hand, if $c > 1$ Theorem 1.2 implies the existence of a unique largest linear component; in Proposition 3.3 the exact size is derived. Here our bounds are the same as previously derived [16].

As is standard in the analysis of the phase transition of random graphs we will use both concentration results and the theory of branching processes, specifically Galton-Watson processes. In the next two sections, we collect the results we will use from these two areas. We do not provide proofs for results which are either well known or easily derived from well know results.

2 Concentration of Measure

Recall the Chernoff inequality on $X \sim \mathrm{Bin}(n,p)$ with $t > 0$ (for a proof see [13] Theorem 2.1)

$$\mathbb{P}\left[X \geq \mathbb{E}X + t\right] \leq \exp\left(-\frac{t^2}{2(np + t/3)}\right). \tag{2.1}$$

Given a subset of the attributes $\{a_{i_1}, a_{i_2}, \ldots, a_{i_k}\}$, it will be useful to approximate the number of vertices likely to be connected to at least one of the given attributes. To find an upper bound on the number of vertices, we first estimate

$$W = \sum_{i \in \{i_1, i_2, \ldots, i_k\}} p_i$$

and then use Equation (2.1). To find good approximations of W as above, we need the following useful generalization of Equation (2.1) due to McDiarmid [17] and further generalized by Chung and Lu [7].

Theorem 2.1 ([7]). *Suppose Y_i are independent random variables with $M_1 \leq Y_i \leq M_2$, for $1 \leq i \leq n$. Let $Y = \sum_{i=1}^{n} Y_i$ and $||Y|| = \sqrt{\sum_{i=1}^{n} \mathbb{E}(Y_i^2)}$. Then*

$$\mathbb{P}[Y \geq \mathbb{E}Y + \lambda] \leq \exp\left(-\frac{\lambda^2}{2(||Y||^2 + M_2\lambda/3)}\right), \tag{2.2}$$

$$\mathbb{P}[Y \leq \mathbb{E}Y - \lambda] \leq \exp\left(-\frac{\lambda^2}{2(||Y||^2 - M_1\lambda/3)}\right). \tag{2.3}$$

3 Branching Processes

We shall make use of the theory of Galton-Watson branching processes. In a single-type Galton-Watson branching process, each individual has descendants given by a common distribution, Z. Standard results (see [12], Chapter 1) show that if the mean of Z is less than 1, the process dies out eventually while if the mean is greater than 1, there is a positive probability given by $1 - \rho$, that the process survives indefinitely. In this case, ρ is the unique solution in $[0, 1)$ to the equation

$$x = \sum_{i=0}^{\infty} \mathbb{P}[Z = i] x^i. \tag{3.1}$$

For a random intersection graph G with parameters n and $\mathbf{p} = (p_i)_{i=1}^{m}$, we will associate the Galton-Watson process where descendants are taken from the probability distribution of the degree of a random vertex, v, in G (i.e. $\mathbb{P}[Z = k] = \mathbb{P}[d(v) = k]$ for each k.) Lemma 3.1 elucidates the relationship between $n, \mathbf{p}$ and the associated Galton-Watson process.

Lemma 3.1. *Let n, $\mathbf{p} = (p_1, p_2, \ldots, p_{m_1})$ and $\mathbf{q} = (q_1, q_2, \ldots, q_{m_2})$ be given such that there exists a $S \subset [m_2]$ and a one-to-one map $\pi : [m_2]\backslash S \to [m_1]$ with the properties that*

(i) $\forall j \in [m_2]\backslash S,\ p_{\pi(j)} = q_j$,

(ii) $\forall j \in S, \forall i \in [m_1]\backslash\pi\big([m_2]\backslash S\big),\ p_i \geq q_j$,

(iii) $n\sum_{i=1}^{m_1} p_i^2 = n\sum_{j=1}^{m_2} q_j^2$.

If $\mathcal{X}$ and $\mathcal{Y}$ are the Galton-Watson processes associated with $\mathbf{p}, \mathbf{q}$ respectively, then the probability that $\mathcal{X}$ dies out is at least as large as the probability that $\mathcal{Y}$ dies out.

Proof. Let X and Y be the degree distributions for the degrees of an arbitrary vertex in the random intersection graphs with parameters $n, \mathbf{p}$ and $n, \mathbf{q}$ respectively. Then $\mathcal{X}$ and $\mathcal{Y}$ are the Galton-Watson processes where each generation is chosen independently from X, respectively Y. Writing $x_i = \mathbb{P}[X = i]$ and $y_i = \mathbb{P}[Y = i]$ it is easy to see that

$$\sum_{i=1}^{n} i x_i = \sum_{j=1}^{n} j y_j\,, \tag{3.2}$$

which is equivalent to condition (iii). Moreover, conditions (i)-(iii) imply that $x_0 \geq y_0$ and that there exists an $l > 0$ with

$$\forall i, 0 < i \leq l,\ \ x_i \leq y_i \text{ and } \forall i, i > l,\ \ x_i \geq y_i. \tag{3.3}$$

It is now easy to show that $f(z) = \sum_{i=0}^{\infty} x_i z^i$ dominates $g(z) = \sum_{j=0}^{\infty} y_j z^j$ on the interval $[0, 1]$. Indeed writing $h(z) = f(z) - g(z)$ we have $h(0) = x_0 - y_0 \geq 0$ while $h(1) = 0$. Then Equation (3.2) and Statement (3.3) imply that for $0 < z < 1$, $h'(z) \leq h'(1) = 0$. That is, h is decreasing on $[0, 1]$, hence $h(z) \geq 0$ for $z \in [0, 1]$. In particular if the expected values $f'(1) = g'(1)$ are greater than 1, then there is a non-zero probability $1 - \rho$ that the process $\mathcal{Y}$ survives. In this case, ρ satisfies the equation $\rho = g(\rho)$. Then $f(\rho) \geq \rho$ which implies that the solution to the equation $z = f(z)$ is at least ρ. Then the probability that $\mathcal{X}$ dies is at least as large as the probability that $\mathcal{Y}$ dies out. □

3.1 Multi-type Galton Watson Processes

It will be convenient to consider multi-type Galton-Watson processes as well. For a random intersection graph with parameters n and $\mathbf{p} = (p_i)_{i=1}^m$, we associate the following $m+1$ type Galton-Watson process. Individuals of type 0 relate to the vertices in the associated random bipartite graph B, while all other individuals in this process relate to attributes of B. Moreover, individuals of type 0 can have offspring of each of the types $1, 2, \ldots, m$; the amount is taken from Bernoulli(p_i) respectively. Individuals of types $i = 1, 2, \ldots, m$ only have offspring of type 0, where the amount is taken from the distribution $\mathrm{Bin}(n, p_i)$. The process starts with one individual of type 0. We review standard results [12] which imply that if $n\sum_i p_i^2 = c > 1$ then this multi-type process survives with positive probability 1-ρ where ρ is given by the unique solution in $(0, 1)$ to Equation (1.1). Note that for a given parameter set, the associated single-type Galton-Watson process and multi-type Galton-Watson process have the same probabilities of survival and extinction.

Consider a general multi-type Galton-Watson process with $m+1$ types labeled $0, 1, \ldots, m$. For each positive integer N and each type i, define $f_N^i(x_0, x_1, \ldots, x_m)$ to be the generating functions for the descendants at time N given that the process started with exactly one individual of type i. That is, let $p_N^i(r_0, r_1, \ldots, r_m)$ represent the probability that the process starting with one individual of type i will in the N^{th} generation have $r_0, r_1, \ldots, r_m$ offspring of types $0, 1, \ldots, m$ respectively. Then the generating functions can be expressed as

$$f_N^i(x_0, x_1, \ldots, x_m) = \sum_{r_0=0}^{\infty} \sum_{r_1=0}^{\infty} \cdots \sum_{r_0=m}^{\infty} p_N^i(r_0, r_1, \ldots, r_m) x_1^{r_1} x_2^{r_2} \cdots x_m^{r_m}.$$

Writing $\mathbf{x} = (x_0, x_1, \ldots, x_m)$ and $\mathbf{f}_N(\mathbf{x}) = (f_N^0(\mathbf{x}), f_N^1(\mathbf{x}), \ldots, f_N^m(\mathbf{x}))$, we have

$$f_{N+1}^i(\mathbf{x}) = f_N^i(\mathbf{f}_1(\mathbf{x})). \tag{3.4}$$

From the definition, $f_N^i(\mathbf{0})$ is exactly the probability of extinction by the N^{th} generation if the process starts with one individual of type i. Thus $f_{N+1}^i(\mathbf{0}) \geq f_N^i(\mathbf{0})$ and in particular, $\lim f_N^i(\mathbf{0})$ exists and is less than or equal to 1. Writing $q_i = \lim f_N^i(0)$ we see that $\mathbf{q} = (q_0, q_1, \ldots, q_m)$ is a solution to the equation

$$\mathbf{f}_1(\mathbf{q}) = \mathbf{q}. \tag{3.5}$$

Let m_{ij} be the expected number of offspring of type j from an individual of type i and let $\mathbf{M} = (m_{ij})$ be the matrix of these first moments. Suppose $\mathbf{s}$ is a vector with $|1 - s_i| \leq 1$ for each i. Then from Taylor's theorem with remainder we have

$$\mathbf{f}_N(\mathbf{1} - \mathbf{s}) = \mathbf{1} - \mathbf{M}^N\mathbf{s} + o(|\mathbf{s}|) \qquad |\mathbf{s}| \to \mathbf{0}$$

Assume that for any such vector $\mathbf{s}$, there exists an N_0 with $|\mathbf{M}^{N_0}\mathbf{s}| > 2|\mathbf{s}|$. Then it follows that there exists a nonnegative solution different from $\mathbf{1}$ to Equation (3.5). Indeed fix $\epsilon > 0$. If $\mathbf{q} = \mathbf{1}$, then there exists a sufficiently large N such that $|\mathbf{1} - \mathbf{f}_N(\mathbf{0})| < \epsilon$. Using $\mathbf{s} = \mathbf{1} - \mathbf{f}_N(\mathbf{0})$ we conclude that

$$|\mathbf{1} - \mathbf{f}_{N+N_0}(\mathbf{0})| = |\mathbf{1} - \mathbf{f}_{N_0}(\mathbf{f}_N(\mathbf{0}))| > |\mathbf{1} - \mathbf{f}_N(\mathbf{0})|,$$

a contradiction to the fact that $\mathbf{q} - \mathbf{f}_N(0) \to \mathbf{0}$ monotonically as $N \to \infty$.

If, in addition to the above assumption, we also assume that $q_i < 1$ for all i, it follows that if $\mathbf{q}_1$ is any vector in the unit cube not equal to $\mathbf{1}$, we have $\mathbf{f}_N(\mathbf{q}_1) \to \mathbf{q}$ as $N \to \infty$. (For a proof, see II.7.2 in [12].) On the other hand, if there exist two types, i and j with $q_i = 1$ and $q_j \neq 1$, then Equations (3.4) and (3.5) imply that for all all N, $1 = f_N^i(\mathbf{q})$. In particular $f_N^i(\mathbf{x})$ is thus independent of x_j which implies that $(\mathbf{M}^N)_{ij} = 0$ for all N.

Corollary 3.2. *If n and $\mathbf{p}$ are given with $n\sum p_i^2 = c > 1$ then the associated multi-type Galton-Watson process, as defined above, survives with probability $1 - \rho$ where ρ is the unique solution in $[0, 1)$ to Equation (1.1).*

Proof. Without loss of generality, we suppose the p_i are all nonzero. Note that the probability generating functions associated with the multi-type Galton-Watson process are

$$f_1^i(\mathbf{x}) = \begin{cases} (1 - p_i + x_0 p_i)^n & \text{if } i > 0 \\ \prod_{i=1}^m (1 - p_i + x_i p_i) & \text{if } i = 0. \end{cases} \tag{3.6}$$

Thus the solution to Equation (3.5) is exactly given by Equation (1.1).

To show that this gives the extinction probability of the branching process, it remains to verify the following two assumptions. First, that for any vector $\mathbf{s}$ sufficiently close to $\mathbf{1}$, there exists an N with $|\mathbf{M}^N\mathbf{s}| > 2|\mathbf{s}|$. Secondly, that for each pair i, j there exists an N such that $(\mathbf{M}^N)_{ij} \neq 0$.

Note that for $i, j > 0$, $m_{ij} = 0$, while for $i > 0$, $m_{0i} = p_i$, $m_{i0} = np_i$ and for all i, $m_{ii} = 0$. As $n\sum p_i^2 = c > 1$ then we have $(\mathbf{M}^{2N})_{00} = c^N$ which

clearly implies the first statement. Secondly, it is not hard to check that $(\mathbf{M}^{2k+1})_{ij} \neq 0$ when exactly one of i, j are equal to 0. On the other hand, $(\mathbf{M}^{2k})_{ij} \neq 0$ when $i, j > 0$ and when $i = j = 0$. Thus the second assumption above also holds and we can conclude that the unique solution in $(0, 1)$ to Equation (1.1) is indeed the probability of extinction when the process starts with one individual of type 0. □

Finally, we show here how to use Equation (1.1) to derive the size of the largest component in the supercritical phase for the uniform homogeneous case.

Proposition 3.3. *Let $m = m(n)$ be a sequence of integers indexed by n and let $c > 1$ be given. Define $p = \sqrt{c/mn}$. Then the associated $m + 1$ type Galton-Watson process eventually dies out with probability given by*

$$\rho = \begin{cases} 1 - (1 - \zeta)mp & \text{if } m = o(n) \\ \zeta & \text{if } n = o(m) \\ \zeta^* & \text{if } m = \Theta(n), \end{cases} \tag{3.7}$$

where ζ and ζ^* are the unique solutions in $(0, 1)$ to the equations $\exp(c(x-1)) = x$ and $\exp(mp \exp(np(x-1)) - 1) = x$, respectively.

Proof. In each case we will use the fact that $1 - p = \exp(-p - o(p))$.

When $mp = o(1)$, letting $\rho = 1 - (1 - \zeta)mp$, we have

$$\begin{aligned} \left[1 - p(1 - (1 - p(1 - \rho))^n)\right]^m &= \left[1 - p\left(1 - e^{-np(1-\rho)(1+o(1))}\right)\right]^m \\ &= \left[1 - p\left(1 - e^{-c(1-\zeta)(1+o(1))}\right)\right]^m \\ &= \left[1 - p\big(1 - \zeta(1 - o(1))\big)\right]^m \\ &= \exp\left[-mp\big(1 - \zeta(1 - o(1))\big)(1 + o(1))\right] \\ &= 1 - \big(1 - \zeta(1 - o(1))\big)mp(1 + o(1)) \to \rho. \end{aligned} \tag{3.8}$$

Secondly, if $np = o(1)$ then we have

$$\begin{aligned} \left[1 - p(1 - (1 - p(1 - \zeta))^n)\right]^m &= \left[1 - p\left(1 - e^{-np(1-\zeta)(1+o(1))}\right)\right]^m \\ &= [1 - np^2(1 - \zeta)(1 + o(1))]^m \\ &= \exp\big(-c(1 - \zeta)(1 + o(1))\big) \to \zeta\,. \end{aligned} \tag{3.9}$$

Finally, if $n = \Theta(m)$, then np and mp are constants. We have

$$\begin{aligned}
&\prod_{i=1}^{m} \left[1 - p(1 - (1 - p(1 - \zeta^*))^n)\right] \\
&\leq \left[1 - p\left(1 - e^{-np(1-\zeta^*)(1+o(1))}\right)\right]^m \\
&= \exp\left(mp\left(e^{np(\zeta^*-1)(1+o(1))} - 1\right)(1 - o(1))\right). \qquad (3.10)
\end{aligned}$$

□

Note if n is replaced with $n(1-o(1))$ and m with $m(1-o(1))$, the asymptotic results are the same.

4 Proofs of Main Theorems

4.1 Discovery Process

For a random intersection graph G, we define the following discovery process. Let B be the bipartite graph associated to G and let v_1 be a vertex (as opposed to an attribute) of B. For $i = 0, 1, 2, \ldots$ inductively define sets of unsaturated vertices, discovered vertices and discovered attributes, denoted by U_i, V_i, A_i, respectively. Initially set $A_0 = \emptyset$, and $V_0 = U_0 = \{v_1\}$. At step i, if U_{i-1} is empty the process terminates. Otherwise pick $v_i \in U_{i-1}$. Let A_i' denote the set of attributes connected to v_i in $A \backslash A_{i-1}$. Thus A_i' is the set of newly discovered attributes. Discover next the vertices, V_i' of $V \backslash V_{i-1}$ connected to at least one attribute in A_i'. Let X_i denote the cardinality of V_i'. Note again that X_i is the number of newly discovered vertices. Define the sets

$$\begin{aligned}
A_i &= A_{i-1} \cup A_i' \\
V_i &= V_{i-1} \cup V_i' \\
U_i &= (U_{i-1} \backslash \{v_i\}) \cup V_i'.
\end{aligned}$$

A vertex or an attribute can only be discovered once. Crucially, the event that the vertex v_i is connected to an attribute $a \in A \backslash A_{i-1}$ is independent of the history of the discovery process. Similarly, the event that an attribute $a \in A_i'$ is connected to a vertex $v \in V \backslash V_{i-1}$ is independent of the history of the discovery process.

4.2 Subcritical phase

Proof of Theorem 1.1. Let $\mathbf{p} = (p_i)_{i=1}^m$ be given such that $n\sum p_i^2 < 1$. Let G be a random intersection graph obtained from $\mathbf{p}$ and set $p = \max p_i$. Consider the discovery process of G: if A_i' is known but V_i has not yet been discovered, define the random variables $W_i = \sum_{j\in A_i'} p_j$ and $X_i^+ \sim \mathrm{Bin}(n, W_i)$. W_i can be thought of as the weight of the attributes associated to the vertex v_i. Note that X_i^+ stochastically dominate X_i, as

$$1 - \prod_{j\in A_i'}(1-p_j) \leq \sum_{j\in A_i'} p_j.$$

The proof now follows from the following three claims.

Claim 4.1. *$\sum_{i=1}^k X_i$ is stochastically dominated by $X_{(k)}^+ \sim Bin(n, \sum_{i=1}^k W_i)$.*

Claim 4.2. *Let $k > \frac{10}{(1-c)^2} np\log n$. Then $n\mathbb{P}[\sum_{i=1}^k W_i > \frac{k+kc}{2n}] = o(1)$.*

Claim 4.3. *Let $k > (15\log n)/(1-c)^2$ and $X_{(k)}^+ \sim Bin(n, \sum_{i=1}^k W_i)$. If $\sum_{i=1}^k W_i \leq (k-1+kc)/2n$, then $n\mathbb{P}[X_{(k)}^+ \geq k-1] = o(1)$.*

Before we prove the claims, we show that they imply the theorem. First, note that the probability that the component in G containing v_1 has size at least k is bounded by $\mathbb{P}[\sum_{i=1}^k X_i \geq k-1]$. Claims 4.1 and 4.3 imply that if $\sum_{i=1}^k W_i$ is small enough then all components have size $O(\log n)$. However, to prove $\sum_{i=1}^k W_i$ is indeed small enough in Claim 4.2 we need $k = \Theta(np\log n)$. As k is the upper bound on the component sizes, we conclude that all components in G have size at most $O(\max\{np\log n, \log n\})$ as desired. □

We now prove Claims 4.1, 4.2, 4.3.

Proof of 4.1. It is clear that for each i, X_i is stochastically dominated by X_i^+. Similarly $\sum_{i=1}^k X_i^+$ is stochastically dominated by $X_{(k)}^+ \sim \mathrm{Bin}(n, \sum_{i=1}^k W_i)$. □

Proof of 4.2. Recall that W_i is the weight of the attributes associated to v_i in the discovery process. As attributes can only be discovered once during the process, $W_i \leq \sum_{j=1}^m p_j \mathcal{I}_{v_i,a_j}$. In particular $\sum_{i=1}^k W_i \leq \sum_{i=1}^k \sum_{j=1}^m p_j \mathcal{I}_{v_i,a_j}$.

The last sum consists of km summands each of which is no greater than p. Applying Theorem 2.1 with $M_2 = p$ it follows that

$$\left\| \sum_{i=1}^{k} \sum_{j=1}^{m} p_j \mathcal{I}_{v_i,a_j} \right\|^2 = k\mathbb{E}\left[\sum_{j=1}^{m} p_j^2 \mathcal{I}_{v_i,a_j} \right] = k \sum_{j=1}^{m} p_j^3 \leq pk \sum_{j=1}^{m} p_j^2 = cpk/n\,.$$

Applying Theorem 2.1 with $\lambda = (1-c)k/(2n)$, we obtain

$$\begin{aligned} n\mathbb{P}\left[\sum_{i=1}^{k} W_i > \frac{1+c}{2}\frac{k}{n} \right] &\leq n\mathbb{P}\left[\sum_{i=1}^{k} \sum_{j=1}^{m} p_j \mathcal{I}_{v_i,a_j} > \frac{1+c}{2}\frac{k}{n} \right] \\ &\leq n \exp\left(-\frac{(1-c)^2 k^2}{(2n)^2 2\left(cpk/n + p(1-c)k/(6n)\right)} \right) \\ &= n \exp\left(-\frac{(1-c)^2 k}{2np(4c+1)} \right) = o(1)\,, \end{aligned}$$

where the last equality follows for $k > \frac{10}{(1-c)^2} np \log n$. □

Proof of 4.3. As $X^+_{(k)} \sim \mathrm{Bin}(n, \sum_{i=1}^{k} W_i)$ and $\sum_{i=1}^{k} W_i \leq (k+kc)/2n$, it follows that $X^+_{(k)}$ is stochastically dominated by $X^{++}_{(k)} \sim \mathrm{Bin}(n, (k+kc)/2n)$. By Chernoff's inequality,

$$\begin{aligned} \mathbb{P}[X^{++}_{(k)} \geq k-1] &\leq \mathbb{P}\left[X^{++}_{(k)} \geq \frac{(1+c)k}{2} + \frac{(1-c)k}{2} - 1 \right] \\ &\leq \exp\left(-\frac{\left(\frac{1-c}{2}k - 1\right)^2}{(1+c)k + 2\left(\frac{1-c}{2}k - 1\right)/3} \right) \\ &\leq \exp\left(-\frac{(1-c)^2 k}{5(2+c)} \right) = o(1)\,, \end{aligned}$$

where the last equality follows by letting $k > (15 \log n)/(1-c)^2$. □

4.3 Supercritical phase

Proof of Theorem 1.2. Let $\gamma \in (1/2, 2/3)$ and $\mathbf{p} = (p_i)_{i=1}^m$ be given such that $p = \max p_i = o(n^{-\gamma})$ and $\sum p_i^2 = c/n$ with constant $c > 1$. Let G be the random intersection graph obtained. Consider the same discovery

process as defined in Section 4.1 on the associated random bipartite graph B. In particular, recall that A'_i is the set of newly discovered attributes at step i and that $W_i = \sum_{j \in A'_i} p_j$. Let $k_- = \max\{\frac{5npc}{(1-c)^2} \log n, \frac{125c}{(1-c)^2} \log n\}$ and $k_+ = n^\gamma$. The following is an adaptation of standard results for the phase transition in Erdős-Rényi random graphs (Theorem 5.4, [13]).

First note that for each $k \in [k_-, k_+]$, the following holds

$$\mathbb{E}\left[\sum_{i=1}^{k} W_i\right] \geq \frac{kc}{n}(1 - o(1)). \tag{4.1}$$

To see this, note that the probability that attribute j is discovered by the kth step is $1 - (1 - p_j)^k$. Thus

$$\begin{aligned}
\mathbb{E}\left[\sum_{i=1}^{k} W_i\right] &= \sum_{j=1}^{m} p_j(1 - (1 - p_j)^k) \\
&= \sum_j \left(kp_j^2 - \binom{k}{2} p_j^3 + \cdots + \binom{k}{k}(-p_j)^{k+1}\right) \\
&\geq k \sum_j \left(p_j^2 - \frac{k-1}{2} p_j^3\right) \\
&= k\frac{c}{n}(1 - o(1))\,.
\end{aligned}$$

The last equality follows from $p_j = o(n^{-\gamma})$ which implies $pk = o(1)$.

We now use Equation (4.1) to show that for each $k \in [k_-, k_+]$,

$$\mathbb{P}\left[\sum_{i=1}^{k} W_i \leq \frac{ck}{n} - \frac{(c-1)k}{3n}\right] = o(n^{-5/3})\,. \tag{4.2}$$

This follows from (2.3) by writing $\sum_{i=1}^{k} W_i = \sum_{j=1}^{m} p_j I_j$, with I_j the indicator random variable equal to 1 with probability $1 - (1 - p_j)^k$ and 0 otherwise. Clearly, $p_j I_j \geq 0$ for each j and $||\sum_{i=1}^{k} W_i||^2 = \sum_j \mathbb{E}[(p_j I_j)^2] = \sum_j p_j^2[1 -$

$(1-p_j)^k]$. Thus

$$\begin{aligned}\mathbb{P}\left[\sum_{i=1}^{k} W_i \leq \frac{ck}{n} - \frac{(c-1)k}{3n}\right] &\leq \exp\left[-\frac{(\frac{c-1}{3})^2 k^2/n^2}{2\sum_j p_j^2[1-(1-p_j)^k]}\right] \\ &\leq \exp\left(-\frac{(c-1)^2k^2}{18n^2\sum_j kp_j^3}\right) \\ &\leq \exp\left(-\frac{(c-1)^2 k}{18cpn}\right) = o(n^{-5/3}).\end{aligned}$$

Indeed, it follows from (2.2) and a similar derivation that for $k \in [k_-, k_+]$

$$\mathbb{P}\left[\sum_{i=1}^{k} W_i > (2c-1)\frac{k}{n}\right] = o(n^{-5/3}). \tag{4.3}$$

We now show that with high probability there are no components with $k \in [k_-, k_+]$ vertices. In particular, we show that either the discovery process terminates after k_- steps, or that for each $k \in [k_-, k_+]$, there are at least $(c-1)k/2$ unsaturated vertices. As $\sum_{i=1}^{k} X_i = |\mathcal{U}_k| + k - 1$, it will be enough to show that for each $k \in [k_-, k_+]$, with high probability, $\sum_{i=1}^{k} X_i$ is at least $(c+1)k_+/2$. From Equation (4.2), with high probability the weight of the discovered attributes after k steps will be at least $(2ck+k)/3n$.

As we only need to find $(c+1)k_+/2$ vertices, we can bound each X_i from below by $X_i^- \sim \text{Bin}(n-(c+1)k_+/2, W_i)$. We further bound from below $\sum_{i=1}^{k} X_i^-$ by $X_{(k)}^-$ where $X_{(k)}^- \sim \text{Bin}(n-(c+1)/2, p_k^-)$, where p_k^- is defined as

$$p_k^- = \sum_{i=1}^{k} W_i - \frac{1}{2}\left(\sum_{i=1}^{k} W_i\right)^2.$$

Equation (4.3) implies that with high probability $p_k^- = \sum_{i=1}^{k} W_i(1-o(1))$. In turn, Equation (4.2) then implies that

$$p_k^- \geq \frac{(2c+1)k}{3n}(1-o(1))\,.$$

The probability there is a component of size between k_- and k_+ is thus bounded above in the following manner:

$$n\sum_{k=k_-}^{k_+}\mathbb{P}\left[\sum_{i=1}^{k}X_i\le k-1+\frac{(c-1)k}{2}\right]\le n\sum_{k=k_-}^{k_+}\mathbb{P}\left[X^-_{(k)}\le k-1+\frac{(c-1)k}{2}\right]$$
$$\le n\sum_{k=k_-}^{k_+}\exp\left(-\frac{(c-1)^2k^2}{25(2c+1)k}\right)$$
$$\le nk_+\exp\left(-\frac{(c-1)^2k_-}{25(2c+1)}\right)=o(1).$$

We now show that if there is a component of size at least k_+, then it is unique with high probability. Suppose that there are two vertices v' and v'' which belong to components of size at least k_+. Consider our discovery process starting at v'. At the end of the k_+ step, there are at least $(c-1)k_+/2$ unsaturated vertices which belong to the component containing v'. Similarly, if we consider the discovery process starting with v'', again there are at least $(c-1)k_+/2$ unsaturated vertices in the component containing v''. Denote by A' and A'' the sets of attributes discovered by the k_+ step for each of the two discovery processes, respectively. If the two components are distinct, then in particular none of the unsaturated vertices in V' are connected to any of the unsaturated vertices in V''. With high probability this will not occur. Indeed,

$$\mathbb{P}[V'\not\sim V'']\le\prod_{i\notin(A'\cup A'')}(1-p_i)^{(c-1)k_+}\le\left[e^{-\sum_{i\notin(A'\cup A'')}p_i}\right]^{k_+(c-1)}$$
$$\le\exp\left(-\frac{(c-1)k_+}{\sqrt{n}}\right)=\exp\left(-(c-1)n^{\gamma-\frac{1}{2}}\right)=o\left(\frac{1}{n^2}\right).$$

The last inequality follows from the fact that $\sum_{i\notin(A'\cup A'')}p_i^2\ge\frac{1}{n}$ implies $\sum_{i\notin(A'\cup A'')}p_i\ge\frac{1}{\sqrt{n}}$. Note that Equation (4.3) implies that $n\sum_{i\notin(A'\cup A'')}p_i^2=c-o(1)>1$.

We have yet to show that G contains a component of size at least k_+. Denote by $\rho=\rho(n,\mathbf{p})$ the probability that a given vertex of the random intersection graph will be in a small (i.e. of size at most k_-) component. Now ρ is bounded from below by the extinction probability $\rho_-=\rho_-(n,\mathbf{p})$ of the associated multi-type branching process.

To bound ρ from above, recall Equation (4.3) which implies that after k_- steps of the discovery process, with high probability $W_{k_-}\le(2c-1)k_-/n$.

Thus we bound ρ by $\rho_+ = \rho_+(n - k_-, \mathbf{p}') + o(1)$ where $\mathbf{p}' = \{p_i | i \in A \backslash A'\}$ for a suitable set of attributes A' such that $\sum_{i \in A'} p_i \leq (2c-1)k_-/n$. Lemma 3.1 implies that ρ_+ is largest when A' consists of the smallest (by weight) elements of A. Thus assuming without loss of generality that the sequence $\mathbf{p}$ is monotone increasing, let l be maximal such that $\sum_{i=1}^{l} p_i < (2c-1)k_-/n$. Then if A' consists of the first l attributes, and $\mathbf{p}' = (p_i)_{i>k}^m$, ρ_+ will be largest given that $\sum_{i \in A'} p_i \leq (2c-1)k_-/n$.

The definition of l implies that $n \sum_{i=1}^{l} p_i^2 = o(1)$ and thus $\sum_{i=1}^{l} \max\{p_i, np_i^2\} = o(1)$. It follows from Equation (1.1) that in the limit as $n \to \infty$ $\rho_+ = \rho_-(1 + o(1))$. Thus Y, the expected number of vertices in small components of G is $\rho(1 + o(1))n$. To see that Y is strongly concentrated about its mean, note that

$$\mathbb{E}[Y^2] \leq n\rho(n, \mathbf{p})k_- + n\rho(n, \mathbf{p})n\rho(n - k_-, \mathbf{p}) = (1 + o(1))\mathbb{E}[Y]\,.$$

Thus by Chebyshev's inequality the variance of Y is $o(1)$ as desired. □

Proof of Theorem 1.3. The proof is exactly the same as above except that we give a weaker upper bound for ρ. Let G_1 be the random intersection graph with parameters $n - k_-, \mathbf{p}' = (p_i)_{i>k}$ as above. Let Y be the degree distribution of G_1 and $\mathcal{Y}$ the associated single-type Galton-Watson branching process. Let G_2 be the random intersection graph on $n - k_-$ vertices and $m = \lfloor \frac{c}{np^2} \rfloor$ attributes, each assigned the probability p. Let X be the degree distribution of G_2 and $\mathcal{X}$ the corresponding Galton-Watson process. The probability generating function for $\mathcal{X}$ and $\mathcal{Y}$ will satisfy the conditions of Lemma 3.1 and thus the extinction probability for $\mathcal{X}$ gives an upper bound on the extinction probability for $\mathcal{Y}$ and thus an upper bound on the probability a vertex in G_1 is in a small component. Applying Proposition 3.3 it follows that the expected size of the largest component in G is at least $\Omega(\min\{p^{-1}, n\})$. □

References

[1] M. Behrisch, *Component evolution in random intersection graphs*, Electronic Journal of Combinatorics **14** (2007).

[2] M. Bloznelis, *Component evolution in general random intersection graphs*, SIAM J. Discrete Math. **24** (2010), 639–654.

[3] ———, *The largest component in an inhomogeneous random intersection graph with clustering*, Electronic Journal of Combinatorics **17** (2010).

[4] Mindaugas Bloznelis, Jerzy Jaworski, and Valentas Kurauskas, *Assortativity and clustering of sparse random intersection graphs.*

[5] B. Bollobás, S. Janson, and O. Riordan, *The phase transition in inhomogeneous random graphs*, Random Structures and Algorithms **31** (2007), 3–122.

[6] Béla Bollobás, Svante Janson, and Oliver Riordan, *Sparse random graphs with clustering*, Random Structures and Algorithms **38** (2011), no. 3, 269–323 (en).

[7] F. Chung and L. Lu, *Complex graphs and networks*, American Mathematical Society, 2006.

[8] M. Deijfen and W. Kets, *Random intersection graphs with tunable distribution and clustering*, Probability in the Engineering and Informational Sciences **23** (2009), 661–674.

[9] J. A. Fill, E. R. Scheinerman, and K. V. Singer-Cohen, *Random intersection graphs when $m = \omega(n)$: An equivalence theorem relating the evolution of the $G(n, m, p)$ and $G(n, p)$ models*, Random Structures and Algorithms **16** (2000), 156–176.

[10] E. Godehardt, J. Jarowski, and K. Rybarczyk, *Challenges at the interface of data analysis computer science and optimization*, ch. Clustering Coefficients of Random Intersection Graphs, pp. 243–253, Springer, 2012.

[11] J. L. Guillaume and M. Latapy, *Bipartite structure of all complex networks*, Information Processing Letters **90** (2004), 215–221.

[12] T. E. Harris, *The theory of branching processes*, Dover, 1989.

[13] S. Janson, T. Łuczak, and A. Ruciński, *Random graphs*, Wiley Interscience, 2000.

[14] J. Jaworski, M. Karoński, and d. Stark, *The degree of a typical vertex in generalized random intersection graph models*, Discrete Mathematics **306** (2006), 2152–2165.

[15] M. Karoński, E.R. Scheinerman, and K.B. Singer-Cohen, *On random intersection graphs: the subgraph problem*, Combinatorics, Probability and Computing **8** (1999).

[16] A. N. Lageråas and M. Lindholm, *A note on the component structure in random intersection graphs with tunable clustering*, Electronic Journal of Combinatorics **15** (2008).

[17] C. McDiarmid, *Probalistic methods for algrothmic discrete mathematics*, ch. Concentration, pp. 195–249, Springer, 1998.

[18] M. E. J. Newman, *Scientific collaboration networks. I. Network construction and fundamental results*, Phyical Review E **64** (2001), no. 1, 016131.

[19] S. Nikoletseas, C. Raptopoulos, and P. Spirakis, *Large independent sets in general random intersection graphs*, Theoretical Computer Science **406** (2008), 215–224.

[20] S.E. Nikoletseas, C. Raptopoulos, and P.G. Spirakis, *Expander properties nd the cover time of random intersection grpahs*, Theoretical Computer Science **410** (2009), 5261–5272.

[21] K. Rybarczyk, *Equivalence of a random intersection graph and $g(n, p)$*, Random Structures and Algorithms **38** (2011), 205–234.

[22] Katarzyna Rybarczyk, *The coupling method for inhomogeneous random intersection graphs*, http://arxiv.org/abs/1301.0466.

[23] K.B. Singer-Cohen, *Random intersection graphs*, PhD thesis, Johns Hopkins University, 1995.

[24] D.J. Watts and S.H. Strogatz, *Collective dynamics of Small-World networks*, Nature **393** (1998), no. 6684, 440–442.