Project Number 001907

DELIS

Dynamically Evolving, Large-scale Information Systems

Integrated Project

Member of the FET Proactive Initiative **Complex Systems**

DELIS-TR-0063

The Existence and Efficient Construction of Large Independent Sets in General Random Intersection Graphs

Christoforos Raptopoulos and Sotiris Nikoletseas and Paul Spirakis

2004

Dynamically Evolving, Large-scale Information Systems

The Existence and Efficient Construction of Large Independent Sets in General Random Intersection Graphs*

S. Nikoletseas [(1,2)], C. Raptopoulos [(1,2)] and P. Spirakis [(1,2)]

(1) Computer Technology Institute (CTI), Greece
(2) Computer Engineering and Informatics Department, Patras University, Greece

Abstract

We investigate here the existence and efficient algorithmic construction of close to optimal independent sets in random models of intersection graphs. In particular:

- We propose *a new model* for random intersection graphs ($G_{n,m,\vec{p}}$) which includes the model provided by [10] (the "uniform" random intersection graphs model) as an important special case. We also define an interesting variation of the model of random intersection graphs, similar in spirit to random regular graphs.
- For this model we derive *exact formulae* for the mean and variance of the number of independent sets of size k (for any k) in the graph.
- We then propose and analyse *three algorithms* for the efficient construction of large independent sets in this model. The first two are variations of the greedy technique while the third is a totally new algorithm. Our algorithms are analysed for the special case of uniform random intersection graphs.

  Our analyses show that these algorithms succeed in finding *close to optimal* independent sets for an interesting range of graph parameters.

Due to lack of space, a clearly marked Appendix is added, to be read at the discreetion of the Program Committee members.
Track A submission

1 Introduction

Random graphs, introduced by P. Erdös and A. Rényi in their celebrated work [8] in 1959, still continue to attract a huge amount of research and interest in the communities of Theoretical Computer Science, Graph Theory and Discrete Mathematics. This growing interest is due to (at least) the following reasons: a) the combinatorial interest of these structures themselves, as well as their mathematical beauty and conceptual challenges posed by their investigation, b) their motivation by real world aspects and in paticular the issue of "reliable network computing", in the sense that the random elements in these graphs may capture faults or unavailabilities in corresponding networks, c) the random graph is a "somewhat" typical instance and is thus heavily used in the average case analysis of graph algorithms.

There exist various models of random graphs. The most famous is the $G_{n,p}$ random graph, a sample space whose points are graphs produced by randomly sampling the edges of a graph on n vertices independently, with the same probability p. Because of the (technically convenient) independence properties of $G_{n,p}$ graphs, this model has been extensively studied, both from the combinatorial as well as from the algorithmic point of view. Other models have also been quite a lot investigated: $G_{n,r}$ (the "random regular graphs", produced by randomly and equiprobably sampling a graph from all regular graphs of n vertices and vertex degree r), $G_{n,M}$ (produced by randomly and equiprobably selecting an element of the class of graphs on n vertices having M edges). For an excellent survey of these models, see [3, 1].

*This work has been partially supported by the IST Programme of the European Union under contract numbers IST-2001-33116 (FLAGS) and 001907 (DELIS).

In this work we investigate, both combinatorially and algorithmically, a *new model* of random graphs. We nontrivially extend the $G_{n,m,p}$ model ("random intersection graphs") introduced by M. Karoński, E.R. Sheinerman and K.B. Singer-Cohen [10] and K.B. Singer-Cohen [16]. In their model, to each of the n vertices of the graph, a random subset of a universal set of m elements is assigned, by independently choosing elements with the same probability p. Two vertices u, v are then adjacent in the $G_{n,m,p}$ graph if and only if their assigned sets of elements have at least one element in common. We extend this model (which we call hereafter "uniform", because of the same probability of selecting elements) by proposing *two new models* which we define below.

Definition 1 (General random intersection graph) *Let us consider a universe $M = \{1, 2, \ldots, m\}$ of elements and a set of vertices $V = \{v_1, v_2, \ldots, v_n\}$. If we assign independently to each vertex v_j, $j = 1, 2, \ldots, n$, a subset S_{v_j} of M by choosing each element $i \in M$ independently with probability p_i, $i =$ 1, 2, $\ldots, m$, and put an edge between two vertices v_{j_1}, v_{j_2} if and only if $S_{v_{j_1}} \cap S_{v_{j_2}} \neq \emptyset$, then the resulting graph is an instance of the general random intersection graph $G_{n,m,\vec{p}}$, where $\vec{p} = [p_1, p_2, \ldots, p_m]$.*

Definition 2 (Regular random intersection graph) *Let us consider a universe $M = \{1, 2, \ldots, m\}$ of elements and a set of vertices $V = \{v_1, v_2, \ldots, v_n\}$. If we assign independently to each vertex v_j, $j = 1, 2, \ldots, n$, a subset S_{v_j} consisting of λ different elements of M, randomly and uniformly chosen, and draw an edge between two vertices v_{j_1}, v_{j_2} if and only if $S_{v_{j_1}} \cap S_{v_{j_2}} \neq \emptyset$, then the resulting graph is an instance of the regular random intersection graph $G_{n,m,\lambda}$.*

The latter model may abstract λ-SAT random formulae. We note the following:

Note 1: When $p_1 = p_2 = \cdots = p_m = p$ the general random intersection graph $G_{n,m,\vec{p}}$ reduces to the $G_{n,m,p}$ as in [10] and we call it the *uniform* random intersection graph.

Note 2: When in the uniform case $mp \geq \alpha \log n$ for some constant $\alpha > 1$ then the model $G_{n,m,p}$ and the model $G_{n,m,\lambda}$ for $\lambda \in (1 \pm \epsilon)mp$, $\epsilon \in (0, 1)$, are essentially *equivalent*, i.e. they assign *almost* the same probability to edge monotone graph events. This follows from degree concentration via Chernoff bounds. Thus, all our results proved here for $G_{n,m,p}$ translate to $G_{n,m,\lambda}$.

1.1 Importance and Motivation

Let us now emphasize the importance of random intersection graphs. First of all, we note that (as proved in [11]) any graph is a random intersection graph. Thus, the $G_{n,m,p}$ model is very general. Furthermore, for some ranges of the parameters m, p ($m = n^\alpha, \alpha > 6$) the spaces $G_{n,m,p}$ and $G_{n,p}$ are equivalent (as proved by Fill, Sheinerman and Singer-Cohen [9], showing that in this range the total variation distance between the graph random variables has limit 0).

Second, random intersection graphs (and in particular our new, non- uniform model) may model real-life applications more accurately (compared to the $G_{n,p}$ case). This is because in many cases the independence of edges is not well-justified. In fact, objects that are closer (like moving hosts in mobile networks or sensors in smart dust networks) are more probable to interact with each other. Even epidemiological phenomena (like spread of disease) tend to be more accurately captured by this "proximity-sensitive" random intersection graphs model. Other applications may include oblivious resource sharing in a distributed setting, interactions of mobile agents traversing the WEB etc.

1.2 Our Contribution

In this work:

1. We first introduce *two new models*, as explained above: the $G_{n,m,\vec{p}}$ model and the $G_{n,m,\lambda}$ model. We feel that our models are important, in the sense that $G_{n,m,\vec{p}}$ is a very general model and $G_{n,m,\lambda}$ is very focused (so it is particularly precise in abstracting several phenomena).

2. We show interesting *relations between the models* we introduce, i.e. we prove that when $mp = \alpha \log n$ then $G_{n,m,p}$ is almost equivalent to $G_{n,m,\lambda}$ (see Note 2 above).

3. Under these models we study the well known and fundamental problem of *finding a maximum independent set of vertices*. In particular, in the most general $G_{n,m,\vec{p}}$ model we estimate *exactly* the mean and the variance of the number of independent sets of size k. To get exact formulas

for the variance, we introduce and use a *"vertex contraction technique"* to evaluate the covariance of random indicator variables of non-disjoint sets of vertices. This technique, we believe, has its own combinatorial interest and may be used in investigating other combinatorial problems as well. Using the exact formulae derived, one then can use the Markov inequality and the well known second moment method (see [1]) to get thresholds for existence (with high probability) of independent sets of size k.

4. Finally, we provide and analyse *three efficient algorithms* for finding large independent sets:

   - Algorithm I is the classic greedy algorithm (for example see [2]) for maximum independent set approximation.
   - Algorithm II is a variation of the above where a random new vertex is tried each time instead of that of current minimum degree.
   - Algorithm III is a totally new algorithm (that we propose) pertinent to the model $G_{n,m,\vec{p}}$.

   For clarity, all our algorithms are analysed for the uniform random intersection graphs model.

   Our algorithms are analysed for the interesting case where $mp \geq \alpha \log n$, (for some constant $\alpha > 1$), in which no isolated vertices exist in $G_{n,m,p}$ and also the results translate to $G_{n,m,\lambda}$ (see Note 2).

   To our knowledge, it is the first time that algorithms for random intersection graphs are proposed and analysed. Our analyses show that in many interesting ranges of p, m, the sizes of the independent sets obtained by the algorithms are quite large.

1.3 Related Work

The model of uniform random intersection graphs was introduced by M. Karoński, E.R. Sheinerman and K.B. Singer-Cohen [10]. In the same work the authors study the evolution of such graphs with respect to the existence of small induced subgraphs. Other important properties, including connectivity, cliques and independent sets are investigated in [16]. The question of how close $G_{n,m,p}$ and $G_{n,p}$ are for various values of m, p has been studied by Fill, Sheinerman and Singer-Cohen in [9].

Also, geometric proximity between randomly placed objects is nicely captured by the model of random geometric graphs (see e.g. [4, 7]) and important variations (like random scaled sector graphs, [6]).

We feel that this work is also similar in flavor (particularly with respect to extending random graph models and designing and analysing efficient algorithms) to our previous works on random regular graphs with edge faults [12, 14] where we investigate multiconnectivity, giant connected component and expansion properties of such graphs. Also, to our work in [13] where we introduced the model of stochastic graph processes (i.e. dynamically changing random graphs with random edge insertions and deletions), generalizing the Erdös - Rényi model of random graph processes.

2 The Size of Independent Sets - Exact Formulae

2.1 Mean Value

Theorem 1 *Let $X^{(k)}$ denote the number of independent sets of size k in a random intersection graph $G(n, m, \vec{p})$, where $\vec{p} = [p_1, p_2, \ldots, p_m]$. Then*

$$E\left[X^{(k)}\right] = \binom{n}{k} \prod_{i=1}^{m} \left((1-p_i)^k + kp_i(1-p_i)^{k-1}\right).$$

Proof. Let V' be any set of k vertices and let

$$X_{V'} = \begin{cases} 1 & \text{if } V' \text{ is an independent set} \\ 0 & \text{otherwise.} \end{cases}$$

Clearly,

$$X^{(k)} = \sum_{V', |V'|=k} X_{V'}$$

and by the linearity of expectation

$$E\left[X^{(k)}\right] = \binom{n}{k} E\left[X_{V'}\right] = \binom{n}{k} P\{V' \text{ is an independent set}\}.$$

In order to determine $E\left[X_{V'}\right]$, let us look at the $G(n, m, p)$ graph from the point of view of the elements of $M = \{1, 2, \ldots, m\}$. The set V' will be an independent set if and only if every element of U is chosen by at most one of the k vertices in V'. Since the elements of U are chosen independently, it follows that

$$E\left[X_{V'}\right] = \prod_{i=1}^{m} P\{\text{element } i \text{ is chosen at most once by the vertices in } V'\}.$$

However, it is obvious that when a specific element i of M is chosen at most once by the vertices in V', then it is either chosen by exactly one vertex in V' or it is not chosen by any of them. Hence,

$$E\left[X_{V'}\right] = \prod_{i=1}^{m} (P\{\text{no vertex in } V' \text{ chooses i}\} + P\{\text{exactly one vertex chooses i}\}).$$

The probability that no vertex in V' chooses element i is exactly $(1 - p_i)^k$, which follows from the observation that each of the k vertices of V' chooses i with probability p_i and independently of the choices of other vertices.

Furthermore, the probability that exactly one vertex in V' chooses element i is exactly $kp_i(1-p_i)^{k-1}$, since there are k different vertices in V' and the probability that only one particular vertex chooses i is $p_i(1 - p_i)^{k-1}$. We have therefore proven that

$$E\left[X^{(k)}\right] = \binom{n}{k} \prod_{i=1}^{m} \left((1 - p_i)^k + kp_i(1 - p_i)^{k-1}\right).$$

□

2.2 Exact Variance Calculation

Theorem 2 *Let $X^{(k)}$ denote the number of independent sets of size k in a random intersection graph $G(n, m, \vec{p})$, where $\vec{p} = [p_1, p_2, \ldots, p_m]$. Then*

$$Var\left(X^{(k)}\right) = \sum_{s=1}^{k} \binom{n}{2k-s} \binom{2k-s}{s} \left(\gamma(k,s) E\left[X^{(k)}\right] - E^2\left[X^{(k)}\right]\right)$$

where $E\left[X^{(k)}\right]$ is the mean number of independent sets of size k and

$$\gamma(k, s) = \prod_{i=1}^{m} \left((1 - p_i)^{k-s} + (k - s)p_i(1 - p_i)^{k-s-1} \left(1 - \frac{sp_i}{1 + (k-1)p_i}\right)\right).$$

Proof. Let V' be any set of k vertices and let

$$X_{V'} = \begin{cases} 1 & \text{if } V' \text{ is an independent set} \\ 0 & \text{otherwise.} \end{cases}$$

Clearly,

$$X^{(k)} = \sum_{V', |V'|=k} X_{V'}$$

and if V_1', V_2' are any two sets of k vertices, then

$$
\begin{aligned}
Var\left(X^{(k)}\right) &= \sum_{V_1',V_2',|V_1'|=|V_2'|=k} Cov(X_{V_1'}, X_{V_2'}) \\
&= \sum_{s=1}^{k} \sum_{\substack{V_1',V_2',|V_1'|=|V_2'|=k \\ |V_1'\cap V_2'|=s}} Cov(X_{V_1'}, X_{V_2'}) \\
&= \sum_{s=1}^{k} \sum_{\substack{V_1',V_2',|V_1'|=|V_2'|=k \\ |V_1'\cap V_2'|=s}} E\left[X_{V_1'}X_{V_2'}\right] - E\left[X_{V_1'}\right]E\left[X_{V_2'}\right] \\
&= \sum_{s=1}^{k} \sum_{\substack{V_1',V_2',|V_1'|=|V_2'|=k \\ |V_1'\cap V_2'|=s}} P\{X_{V_1'}X_{V_2'}=1\} - E^2\left[X^{(k)}\right]. \qquad (1)
\end{aligned}
$$

Since

$$
\begin{aligned}
P\{X_{V_1'}X_{V_2'}=1\} &= P\{X_{V_1'}=1|X_{V_2'}=1\}P\{X_{V_2'}=1\} \\
&= P\{X_{V_1'}=1|X_{V_2'}=1\}E\left[X^{(k)}\right] \qquad (2)
\end{aligned}
$$

the problem of computing the variance of $X^{(k)}$ is reduced to computing the conditional probability $P\{X_{V_1'}=1|X_{V_2'}=1\}$, i.e. the probability that V_1' is an independent set given that V_2' is an independent set, where V_1', V_2' are any two sets of k vertices that have s vertices in common. In order to compute $P\{X_{V_1'}=1|X_{V_2'}=1\}$, we will try to merge several vertices into one supervertex and study its probabilistic behaviour.

Towards this goal, let us fix an element i of $M=\{1,2,\ldots,m\}$ and let us consider two (super)vertices v_1, v_2 of the $G(n,m,p)$ graph that choose element i independently with probability $p_i^{(1)}$ and $p_i^{(2)}$ respectively. Let also S_{v_1}, S_{v_2} denote the sets of elements of M assigned to v_1 and v_2 respectively. Then,

$$
\begin{aligned}
P\{i\in S_{v_1}|\nexists(v_1,v_2)\} &= P\{i\in S_{v_1}, i\notin S_{v_2}|\nexists(v_1,v_2)\} \\
&= \frac{P\{i\in S_{v_1}, i\notin S_{v_2}, \nexists(v_1,v_2)\}}{P\{\nexists(v_1,v_2)\}} \\
&= \frac{p_i^{(1)}(1-p_i^{(2)})(1-p_i^{(1)}p_i^{(2)})^{m-1}}{(1-p_i^{(1)}p_i^{(2)})^m} \\
&= \frac{p_i^{(1)}(1-p_i^{(2)})}{1-p_i^{(1)}p_i^{(2)}} \qquad (3)
\end{aligned}
$$

where (v_1,v_2) denotes an edge between v_1 and v_2. From this equation we may conclude the following:

- Conditional on the fact that the edge (v_1,v_2) does not exist, the probabilistic behaviour of vertex v_1 is identical to the probabilistic behaviour of a single vertex that chooses element i of M independently with probability $\frac{p_i^{(1)}(1-p_i^{(2)})}{1-p_i^{(1)}p_i^{(2)}}$.
- Conditional on the fact that the edge (v_1,v_2) does not exist, the probabilistic behaviour of the vertices v_1 and v_2 considered as a unit is identical to the probabilistic behaviour of a single vertex that chooses element i of M independently with probability

$$
\begin{aligned}
P\{i\in S_{v_1}\cup S_{v_2}|\nexists(v_1,v_2)\} &= P\{i\in S_{v_1}|\nexists(v_1,v_2)\} + P\{i\in S_{v_2}|\nexists(v_1,v_2)\} \\
&= \frac{p_i^{(1)}(1-p_i^{(2)})}{1-p_i^{(1)}p_i^{(2)}} + \frac{p_i^{(2)}(1-p_i^{(1)})}{1-p_i^{(1)}p_i^{(2)}} \\
&= \frac{p_i^{(1)}+p_i^{(2)}-2p_i^{(1)}p_i^{(2)}}{1-p_i^{(1)}p_i^{(2)}} \qquad (4)
\end{aligned}
$$

where i is a fixed element of M. The first of the above equations follows from the observation that if there is no edge between v_1 and v_2 then the sets S_{v_1} and S_{v_2} are disjoint, meaning that element i cannot belong to both of them. The second equation follows from symmetry.

Let us now consider merging one by one the vertices of the $G(n,m,p)$ graph into one supervertex. Let w_j denote a supervertex containing j simple vertices that form an independent set, i.e. there are no edges between them. It is obvious that the probabilistic behaviour of w_j is irrelevant to how partial mergings are made. Moreover, if w_{j_1}, w_{j_2} are two supervertices representing two disjoint sets of simple vertices, we say that an edge (w_{j_1}, w_{j_2}) exists if and only if any edge connecting a simple vertex in w_{j_1} and a simple vertex in w_{j_2} exists. Thus, the event $\{\nexists(w_{j_1}, w_{j_2})\}$ is equivalent to the event $\{$the vertices in w_{j_1} together with those in w_{j_2} form an independent set$\}$.

Using equation (4) one can show that $P\{i \in S_{w_2}\} = \frac{2p_i}{1+p_i}$, $P\{i \in S_{w_3}\} = \frac{3p_i}{1+2p_i}$ and by induction

$$P\{i \in S_{w_j}\} = \frac{jp_i}{1+(j-1)p_i} \tag{5}$$

where i is a fixed element of M and S_{w_j} represents the union of all the sets of elements of M that are assigned to each simple vertex contained in w_j. More formally,

$$S_{w_j} = \bigcup_{v \in w_j} S_v$$

where v denotes a simple vertex and S_v denotes the set of elements of M assigned to v. Note that because of the definition of w_j, the subsets S_v in the above union are disjoint.

In the light of the above, let V_1' be any set of k (simple) vertices and let V_2' be an independent set of k vertices that has s vertices in common with V_1'. Since there is no edge between any two vertices in V_2', we can treat the $k-s$ vertices of V_2' not belonging to V_1' and the s vertices belonging to both V_1' and V_2' as two seperate supervertices w_{k-s} and w_s respectively that do not communicate by an edge. Hence, by equations (3), (4) and (5), the probabilistic behaviour of w_s is identical to the probabilistic behaviour of a single vertex w_s' that chooses the elements of M independently with probabilities $\{p_i^{(w_s')}, i = 1, \ldots, m\}$ respectively, where

$$p_i^{(w_s')} = \frac{p_i^{(w_s)}(1-p_i^{(w_{k-s})})}{1-p_i^{(w_s)}p_i^{(w_{k-s})}} = \frac{\frac{sp_i}{1+(s-1)p_i}\left(1-\frac{(k-s)p_i}{1+(k-s-1)p_i}\right)}{1-\frac{sp_i}{1+(s-1)p_i}\frac{(k-s)p_i}{1+(k-s-1)p_i}} = \frac{sp_i}{1+(k-1)p_i}. \tag{6}$$

Let now V'' denote a set of $k-s$ simple vertices and a vertex identical to w_s'. Then, for a fixed element i of M, each of the $k-s$ simple vertices chooses i independently with probability p_i, while the supervertex w_s' chooses i independently with probability $p_i^{(w_s')}$. Using a similar proof to that of Theorem 1, one can easily verify that

$$\begin{aligned} P\{X_{V_1'} = 1 | X_{V_2'} = 1\} &= P\{V'' \text{ is an independent set}\} \\ &= \prod_{i=1}^{m}\left((1-p_i)^{k-s} + (k-s)p_i(1-p_i)^{k-s-1}(1-p_i^{(w_s')})\right) \stackrel{def}{=} \gamma(k,s). \end{aligned}$$

Hence, by equations (1) and (2), we have

$$Var\left(X^{(k)}\right) = \sum_{s=1}^{k}\binom{n}{2k-s}\binom{2k-s}{s}\left(\gamma(k,s)E\left[X^{(k)}\right] - E^2\left[X^{(k)}\right]\right)$$

where $E\left[X^{(k)}\right]$ is given by Theorem 1. □

3 Finding Large Independent Sets in $G_{n,m,p}$

We start from the classic greedy approach.

3.1 Algorithm I

Algorithm I:

Input: An instance $G(V, E)$ of $G_{n,m,p}$.
Output: An independent set V' of G.

1. set $V' := \emptyset$;
2. set $U := V$;
3. **while** $U \neq \emptyset$ **do**
4. **begin**
5.   let $x :=$ vertex of *minimum* degree in the graph induced by U;
6.   set $V' := V' \cup \{x\}$;
7.   eliminate x and all its neighbours from U;
8. **end**
9. **output** V';

3.1.1 Analysis of the Expected Size of the Independent Set constructed

As can be seen in e.g. Ausiello et al in [2], if $r = |V'|$ eventually, and $\delta = \frac{|E|}{n}$, i.e. δ is the *density* of G), then

$$r(2\delta + 1) \geq n. \tag{7}$$

This holds for any input graph G (for a proof of (7) see Appendix). Taking expectations we get

$$E\left[r(2\delta + 1)\right] \geq n.$$

where the expectation is taken over all instances of the distribution $G_{n,m,p}$ (notice that both r, δ are random variables).

Now note that the property "$\exists$ independent set of size r" is monotone decreasing on the number of edges (i.e. if we add edges to G the size of the independent sets cannot increase), while the property "the density of G is δ" is monotone increasing. Hence, by the FKG inequality (see [1]) we get

$$E\left[r\delta\right] \leq E\left[r\right] E\left[\delta\right]$$

or equivalently

$$E\left[r(2\delta+1)\right] = 2E\left[r\delta\right] + E\left[r\right] \leq 2E\left[r\right] E\left[\delta\right] + E\left[r\right] = E\left[r\right] \left(2E\left[\delta\right] + 1\right).$$

Using the fact that $E\left[r(2\delta + 1)\right] \geq n$, we conclude that

$$E\left[r\right] \geq \frac{n}{2E\left[\delta\right] + 1} = \frac{n}{2\frac{E(|E|)}{n} + 1}. \tag{8}$$

In order to compute $E(|E|)$, let us define the indicator random variables

$$X_{u,v} = \begin{cases} 1 & \text{if there is an edge } (u, v) \text{ in } G \\ 0 & \text{otherwise.} \end{cases}$$

Clearly,

$$|E| = \sum_{u,v \in V, u \neq v} X_{u,v}$$

and by the linearity of expectation

$$E(|E|) = \binom{n}{2} E\left[X_{u,v}\right] = \binom{n}{2} P\{X_{u,v} = 1\}.$$

But

$$P\{X_{u,v} = 1\} = P\{\exists i \in M : i \in S_u \cap i \in S_v\} = 1 - (1 - p^2)^m$$

where S_u, S_v are the sets of elements of $M = \{1, 2, \ldots, m\}$ assigned to u, v respectively.

Hence,

$$E(|E|) = \binom{n}{2} \left(1 - (1 - p^2)^m\right).$$

Applying the above result to inequality (8), we conclude the following

Lemma 1 *The expected cardinality of the independent set constructed by Algorithm I is at least*

$$\frac{n^2}{2\binom{n}{2}\left(1 - (1 - p^2)^m\right) + n} = \frac{n^2}{2E(|E|) + n}.$$

The next result is easily derived from Lemma 1.

Corollary 1 (Sparse $G_{n,m,p}$ theorem) *When p is such that $E(|E|) = \Theta(n)$, then the expected size of the independent set provided by Algorithm I is also $\Theta(n)$.*

Remark: The above analysis carries out in an almost similar way to the general random intersection graphs model.

For example, if $p = \frac{\alpha}{\sqrt{nm}}$, where $0 < \alpha < 1$, then

$$E[r] \geq \frac{1}{\alpha m p^2} = \frac{n}{\alpha}$$

which is a quite large independent set indeed.

3.1.2 A Concentration Result for Sparce Graphs

We are interested in intersection graphs $G_{n,m,p}$ for p satisfying

$$\frac{\omega(n)}{n\sqrt{m}} \leq p \leq \sqrt{\frac{2\log n - \omega(n)}{m}}$$

for the smallest possible function $\omega(n) \to \infty$, as $n \to \infty$. This is the range for nontrivial graphs (see [9]).

We consider here the case $p < \sqrt{\frac{1}{8nm}}$ which is in the range of nontrivial graphs.

In the sequel we assume that

$$p(n) = \frac{c(n)}{m}$$

where $c(n) \to \infty$, as $n \to \infty$. For example, since $c(n) = mp$, if we take p in the range of nontrivial graphs, then

$$\frac{\sqrt{m}}{n}\omega(n) \leq c(n) \leq \sqrt{2m\log n - \omega(n)m}. \tag{9}$$

A choice of $c(n)$ satisfying this is $c(n) = \alpha \log n$, where $\alpha > 1$, since, from (9), $\omega(n)$ must be less than $2\log n$.

Notice that our assumption $p < \sqrt{\frac{1}{8nm}}$ implies $c(n) = mp < \sqrt{\frac{m}{8n}}$ which is satisfied by $c(n) = \alpha \log n$, i.e. for $m \geq 8\alpha^2 n \log^2 n$.

Let us now consider a vertex v and let S_v denote the set of elements assigned to it. Using Chernoff bounds (see e.g. [5]), for $mp = \alpha \log n$ and $\epsilon \in (0,1)$, we get

$$P\{||S_v| - mp| \leq \epsilon mp\} \geq 1 - e^{-\frac{\epsilon^2}{2}mp} = 1 - n^{-\frac{\alpha\epsilon^2}{2}}.$$

Then,

$$P\{\exists v : ||S_v| - mp| \geq \epsilon mp\} \leq \sum_{v \in V} P\{||S_v| - mp| \geq \epsilon mp\} = n^{-\frac{\alpha\epsilon^2}{2}+1}.$$

If we choose the parameter α so that $-\frac{\alpha\epsilon^2}{2} + 1 > 2$, then all vertices have each a number of chosen elements "around" mp with probability at least $1 - \frac{1}{n^2}$.

Let us condition $G_{n,m,p}$ on this event. Because of symmetry, the elements chosen by each vertex are otherwise uniform in $\{1, 2, \ldots, m\}$.

Let us now consider the following algorithm.

Algorithm II:

Input: An instance $G(V, E)$ of $G_{n,m,p}$.
Output: An independent set V' of G.

1. set $V' := \emptyset$;
2. set $U := V$;
3. **while** $U \neq \emptyset$ **do**
4. **begin**
5. let $u :=$ a random vertex of U;
6. $U := U - \{u\}$;
7. let $S(V') := \cup_{u \in V'} S_u$;
8. **if** (u intersects with any vertex in V') **then** reject u
9. **else** $V' := V' \cup \{u\}$;
10. **end**

The difference between Algorithm I and Algorithm II is that in the latter we do not use the (useful) heuristic urging us to choose at each iteration the vertex with the current minimum degree. It is clear that if r_1, r_2 are the sizes of the independent sets provided by Algorithm I and Algorithm II respectively, then

$$P\{r_1 \geq x\} \geq P\{r_2 \geq x\}$$

for all $x > 0$, i.e. the random variable r_1 stochastically dominates r_2.

We now concentrate on estimation of r_2 with high probability (whp). Clearly, after i successful node insertions into V' the following are true:

- $|S(V')| \in (1 \pm \epsilon)imp = (1 \pm \epsilon)ic(n)$.
- The next tried node u is rejected with probability

$$P_{rej} = 1 - \left(1 - \frac{|S(V')|}{m}\right)^{|S_u|}$$

since each element $l \in S_u$ belongs also in $S(V')$ with probability $\frac{|S(V')|}{m}$, which in turn follows from independence and uniformity.

Combining these two observations we conclude that the probability that a vertex u is rejected after i successful insertions is

$$P_{rej} \leq \frac{|S(V')||S_u|}{m}$$

which is at most $\frac{(1+\epsilon)^2 ic^2(n)}{m}$, for any $\epsilon \in (0,1)$, provided that $\frac{ic^2(n)}{m} \to 0$, i.e. provided that $i = o\left(\frac{m}{c^2(n)}\right)$. (Note also that, by the Bernoulli inequality, we have

$$P_{acc} = 1 - P_{rej} = \left(1 - \frac{|S(V')|}{m}\right)^{|S_u|} \geq 1 - \frac{|S(V')||S_u|}{m}$$

and when $\frac{|S(V')||S_u|}{m} \to 0$, then $P_{acc} \to 1$.)

Since $i \leq n$ and $1+\epsilon < 2$, for any $\epsilon \in (0,1)$, we have that

$$P_{rej} < \frac{4nc^2(n)}{m}.$$

Moreover, since $mp < \sqrt{\frac{m}{8n}}$ by assumption, we get

$$P_{rej} < \frac{1}{2}.$$

Thus, the number r_2 of nodes that are successfuly inserted into V' is at least the number of successes of the Bernoulli $\mathcal{B}(n, \frac{1}{2})$.

¿¿¿From Chernoff boundes then, for any $\beta > 0$ we have $r_2 \geq (1-\beta)\frac{n}{2}$ with probability at least $1 - \exp\{-\frac{\beta^2}{2}\frac{n}{2}\}$.

Since r_1 stochastically dominates r_2 we eventually have (set $\beta = \frac{1}{2}$), by combining events

Theorem 3 *Consider a random intersection graph $G_{n,m,p}$ with $p < \sqrt{\frac{1}{8nm}}$ and $mp = \alpha \log n$, ($\alpha > 1$ a constant). Then Algorithm I constructs an independent set of size at least $\frac{n}{4}$ with probability at least $1 - \frac{1}{2n^2}$.*

3.2 Algorithm III

Algorithm III:

Input: A random intersection graph $G_{n,m,p}$.
Output: An independent set of verices A_m.

1. set $A_0 := V$; set $L := M$;
2. **for** $i = 1$ **to** m **do**
3. **begin**
4. select a random label $l_i \in L$; set $L := L - \{l_i\}$;
5. set $D_i := \{v \in V : l_i \in S_v\}$;
6. **if** $(|D_i| \geq 1)$ **then** select a random vertex $u \in D_i$ and set $D_i := D_i - \{u\}$;
7. set $A_i := A_{i-1} - D_i$;
8. **end**
9. **output** A_m;

Firstly, we concern ourselves with the correctness of the above algorithm.

Theorem 4 *Algorithm III correctly finds an independent set of vertices.*

Proof. In order to prove the correctness of Algorithm III let us consider any two vertices v_1 and v_2 that are connected via an edge, i.e. there is at least one element $i \in M$ that belongs to both S_{v_1} and S_{v_2}. It is easy to verify that at most one of these vertices can belong to A_m, since after the examination of element i of M, the algorithm will choose at most one of v_1 and v_2 (another possible scenario is that by the time the algorithm starts the examination of i one of the vertices v_1 and v_2 or both have been excluded from the independent set in previous steps). □

The following result concerns the efficiency of Algorithm III.

Theorem 5 *For the case $mp = \alpha \log n$ for some constant $\alpha > 1$ and $m \geq n$ with high probability we have for some constant $\beta > 0$:*

1. *If $np \to \infty$ then $|A_m| \geq (1-\beta)\frac{n}{\log n}$.*
2. *If $np \to b$ where $b > 0$ is a constant then $|A_m| \geq (1-\beta)n(1-e^{-b})$.*
3. *If $np \to \infty$ then $|A_m| \geq (1-\beta)n$.*

Proof. Let us define the indicator random variables

$$X_v^{(i)} = \begin{cases} 1 & \text{if vertex } v \text{ of } A_{i-1} \text{ does not contain } l_i \\ 0 & \text{otherwise} \end{cases}$$

and

$$I_{D_i} = \begin{cases} 1 & \text{if } |D_i| \geq 1 \\ 0 & \text{otherwise.} \end{cases}$$

Clearly,

$$|A_i| = \sum_{v \in A_{i-1}} X_v^{(i)} + I_{D_i} \quad , \text{for } i = 1, 2, \ldots, m.$$

Since the elements of M are chosen independently, the random variables $X_v^{(i)}$ are independent of the set A_{i-1}. Hence, by Wald's equation (for the expectation of the sum of a random number of independent random variables, see [15]) and the linearity of expectation we have

$$E(|A_i|) = E(|A_{i-1}|)(1-p) + P\{|D_i| \geq 1\} \quad , \text{for } i = 1, 2, \ldots, m.$$

Using the above equation we can prove inductively that

$$E(|A_m|) = n(1-p)^m + \sum_{i=1}^{m} (1-p)^{m-i} P\{|D_i| \geq 1\}. \tag{10}$$

(*Note:* By a similar proof one can verify that the term $n(1-p)^m$ is the mean number of isolated vertices in the graph. By choosing $mp \geq \alpha logn$ for some constant $\alpha > 1$ the mean number of isolated vertices tends to 0.)

Now let $L_i = \{v \in V : l_i \in S_v\}$, i.e. L_i is the set of vertices having l_i (before examining them for other elements of M). Then

$$P\{|D_i| \geq 1\} = 1 - P\{|D_i| = 0\} = 1 - (P\{v \notin D_i\})^n \tag{11}$$

where v is any specific vertex. The second equality follows from symmetry. But

$$\begin{aligned} P\{v \notin D_i\} &= P\{v \notin L_i \cap v \notin D_i\} + P\{v \in L_i \cap v \notin D_i\} \\ &= P\{v \notin L_i\} + P\{v \in L_i \cap v \notin D_i\} \\ &= 1 - p + P\{v \in L_i \cap \{v \in L_1 \cup L_2 \cup \cdots \cup L_{i-1}\}\}. \end{aligned}$$

Since the choices of the elements of M are independent, the events $\{v \in L_i\}$ and $\{v \in L_1 \cup L_2 \cup \cdots \cup L_{i-1}\}$ are also independent. Hence

$$\begin{aligned} P\{v \notin D_i\} &= 1 - p + P\{v \in L_i\} P\{v \in L_1 \cup L_2 \cup \cdots \cup L_{i-1}\} \\ &= 1 - p + p\,(1 - P\{l_j \notin S_v, j = 1, 2, \ldots, i-1\}) \\ &= 1 - p + p\left(1 - (1-p)^{i-1}\right) = 1 - p(1-p)^{i-1}. \end{aligned}$$

So, by equation (11) we get

$$P\{|D_i| \geq 1\} = 1 - \left(1 - p(1-p)^{i-1}\right)^n.$$

Now by equation (10) we have

$$\begin{aligned} E(|A_m|) &= n(1-p)^m + \sum_{i=1}^{m} (1-p)^{m-i} \left(1 - \left(1 - p(1-p)^{i-1}\right)^n\right) \\ &= n(1-p)^m + \frac{1}{p}\left(1 - (1-p)^m\right) - \sum_{i=1}^{m} (1-p)^{m-i} \left(1 - p(1-p)^{i-1}\right)^n. \end{aligned}$$

In the interesting case where $mp \geq \alpha \log n$ for some constant $\alpha > 1$ and $m \geq n$ (implying that $p \to 0$) we get

$$\begin{aligned} E(|A_m|) &\sim n(1-p)^m + \frac{1}{p}\left(1 - (1-p)^m\right) - \sum_{i=1}^{m} (1-p)^{m-i} (1-p)^n \\ &= n(1-p)^m + \frac{1}{p}\left(1 - (1-p)^m\right)\left(1 - (1-p)^n\right) \\ &\sim \frac{1}{p}\left(1 - (1-p)^n\right). \end{aligned}$$

We now distinguish three cases, covering all possible values of np.

1. If $np \to \infty$ then $E(|A_m|) \sim \frac{1}{p}$. The largest p to have $np \to \infty$, $mp \geq \alpha \log n$ and $m \geq n$ is $p = \frac{\log n}{n}$. So, we conclude that $E(|A_m|) = \Omega(\frac{n}{\log n})$.
2. If $np \to b$ where $b > 0$ is a constant then $E(|A_m|) \sim \frac{n}{b}(1 - e^{-b}) = \Theta(n)$.
3. If $np \to 0$ then $E(|A_m|) \sim \frac{1}{p}(1 - 1 + np) = \Theta(n)$.

The proof ends with the observation that since $E(|A_m|) \to \infty$ in all tree cases, then one can use chernoff bounds to prove that $|A_m| \geq (1-\beta)E(|A_m|)$ for any constant $\beta > 0$ with very high probability. □

4 Conclusions and Further Work

We proposed here a very general, yet tractable, model for random intersection graphs. We believe that it can be useful in many technological applications. We also did the first step in analysing algorithms for random intersection graphs, and we did that for the problem of construction of large independent sets of vertices.

The finding of efficient algorithms for construction of other interesting graph objects (for example hamiltonian lines, long paths, giant components, dominating sets etc) is a subject of our future work.

5 Acknowledgement

We wish to thank M. Karoński and K.B. Singer-Cohen for informing us about their work in random intersection graphs and providing useful material.

References

[1] N. Alon and J.H. Spencer, *"The Probabilistic Method"*, Second Edition, John Wiley & Sons, Inc, 2000.

[2] G. Ausiello, P. Crescenzi, G. Gambosi, V. Kann, A. Marchetti-Spaccamela and M. Protasi, *"Complexity and Approximation"*, Springer-Verlag Berlin Heidelberg, 1999.

[3] B. Bollobás, *"Random Graphs"*, Second Edition, Cambridge University Press, 2001.

[4] J. Díaz, M.D. Penrose, J.Petit and M. Serna, *"Approximating Layout Problems on Random Geometric Graphs"*, Journal of Algorithms, 39:78-116, 2001.

[5] J. Díaz, J.Petit and M. Serna, Chapter titled *"A Guide to Concentration Bounds"*, in the "Handbook of Randomized Computing - Volumes I & II (Combinatorial Optimization 9)", pp.457-507, Kluwer Academic Publishers, Volume I, 2001.

[6] J. Díaz, J.Petit and M. Serna, *"A Random Graph Model for Optical Networks of Sensors"*, in the 1st International Workshop on Efficient and Experimental Algorithms, (WEA), 2003.

[7] J. Díaz, J.Petit and M. Serna, *"Random Geometric Problems on* $[0,1]^2$*"*, in J. Rolim, M. Luby and M. Serna, editors, Randomization and Approximation Techniques in Computer Science, volume 1518 of Lecture Notes in Computer Science, pages 294-306, Springer Verlag, Berlin, 1998.

[8] P. Erdös and A. Rényi, *"On the Evolution of Random Graphs"*, Publ. Math. Inst. Hungar. Acad. Sci. 5, 17-61.

[9] J.A. Fill, E.R. Sheinerman and K.B 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"*, `http://citeseer.nj.nec.com/fill98random.html`

[10] M. Karoński, E.R. Scheinerman and K.B. Singer- Cohen, *"On Random Intersection Graphs: The Subgraph Problem"*, Combinatorics, Probability and Computing journal (1999) 8, 131-159.

[11] E. Marczewski, *"Sur deux propriétés des classes d' ensembles"*, Fund. Math. 33, 303- 307 (1945).

[12] S. Nikoletseas, K. Palem, P. Spirakis and M. Yung, *"Short Vertex Disjoint Paths and Multiconnectivity in Random Graphs: Reliable Network Computing"*, in the Proceedings of the 21st International Colloquium on Automata, Languages and Programming (ICALP), Lecture Notes in Computer Science Vol. 820 (Springer Verlag), pp. 508-519, 1994. Also, in the Special Issue on Randomized Computing of the International Journal of Foundations of Computer Science (IJFCS), Vol. 11 No. 2 (2000), pp. 247-262, World Scientific Publishing Company, 2000.

[13] S. Nikoletseas, J. Reif, P. Spirakis and M. Yung, *"Stochastic Graphs Have Short Memory: Fully Dynamic Connectivity in Polylogarithmic Expected Time"*, in the Proceedings of the 22nd International Colloquium on Automata, Languages and Programming (ICALP), Lecture Notes in Computer Science Vol. 944 (Springer Verlag), pp. 159-170, 1995.

[14] S. Nikoletseas and P. Spirakis, *"Expander Properties in Random Regular Graphs with Edge Faults"*, in the Proceedings of the 12th Annual Symposium on Theoretical Aspects of Computer Science (STACS), Lecture Notes in Computer Science Vol. 900 (Springer Verlag), pp. 421-432,1995.

[15] S.M. Ross, *"Stochastic Processes"*, Second Edition, John Wiley & Sons, Inc., 1996.

[16] K.B. Singer-Cohen, *"Random Intersection Graphs"*, PhD thesis, John Hopkins University, 1995.

APPENDIX

Lemma 2 *Running Algorithm I for any graph $G(V,E)$, we have*

$$r(2\delta + 1) \geq n$$

where $n = |V|$, r is the cardinality of the independent set V' produced by Algorithm I and $\delta = \frac{|E|}{n}$ is the "density" of G.

Proof. Let x_i be the vertex chosen at iteration i and let d_i be the degree of x_i in the graph induced by U. Since Algorithm I deletes x_i together with all its neighbours and since d_i is the minimum vertex degree currently induced by U, then at least $\frac{d_i(d_i+1)}{2}$ edges are deleted at iteration i.

Summing over all iterations we have

$$\sum_{i=1}^{r} \frac{d_i(d_i + 1)}{2} \leq m = \delta n.$$

Since Algorithm I stops when all vertices are eliminated we get

$$\sum_{i=1}^{r} (d_i + 1) = n.$$

Hence,

$$\sum_{i=1}^{r} \frac{d_i(d_i + 1)}{2} + 2\sum_{i=1}^{r} (d_i + 1) = \sum_{i=1}^{r} (d_i + 1)^2 \leq n(2\delta + 1).$$

However, the Cauchy-Schwartz inequality states that $\sum_{i=1}^{r} (d_i + 1)^2$ is minimum when $d_{i+1} = \frac{n}{r}$, for all i. Thus,

$$\sum_{i=1}^{r} (d_i + 1)^2 \geq \frac{n^2}{r}$$

which implies that

$$\frac{n^2}{r} \leq n(2\delta + 1)$$

or equivalently

$$n \leq r(2\delta + 1).$$

□