Two Models of Random Intersection Graphs and their Applications

Erhard Godehardt

Clinic of Thoracic and Cardiovascular Surgery, Heinrich Heine University, D-40225 Düsseldorf, Germany

Jerzy Jaworski

Faculty of Mathematics and Computer Science, Adam Mickiewicz University, PL-60769 Poznań, Poland

---

Let $\mathcal{B}G_{n,m}\left(n,\mathcal{P}_{(m)}\right)$ be a random bipartite graph with the 2-partition $(\mathcal{V},\mathcal{W})$ of the vertex set $\mathcal{V}\cup\mathcal{W}=\{v_1,v_2,\ldots,v_n\}\cup\{w_1,w_2,\ldots,w_m\}$, such that

- each vertex $v\in\mathcal{V}$ chooses its degree and then its neighbors from $\mathcal{W}$ independently of all other vertices from $\mathcal{V}$;
- the vertex v chooses its degree according to the probability distribution $\mathcal{P}_{(m)}=(P_0,P_1,\ldots,P_m)$, i.e., $\Pr\{|\Gamma(v)|=k\}=P_k,\ k=0,1,\ldots,m$, where $\Gamma(v)$ denotes the set of neighbors of a vertex v;
- for every $\mathcal{S}\subseteq\mathcal{W},\ |\mathcal{S}|=k,$ the probability that $\mathcal{S}$ coincides with the set of neighbors of a vertex v is given by $\Pr\{\mathcal{S}=\Gamma(v)\}=P_k\Big/\binom{m}{k}$.

Let X^* be a discrete random variable with probability distribution $\mathcal{P}_{(m)}$, i.e., $\Pr\{X^*=k\}=P_k,\ k=0,1,\ldots,m$. Due to the homogeneous structure of the random graph $\mathcal{B}G_{n,m}\left(n,\mathcal{P}_{(m)}\right)$, X^* defines the degree of a given vertex from $\mathcal{V}$ (we have n independent, identically distributed random variables $X^*(v_1),\ldots,X^*(v_n)$). In contrast to the degrees of vertices from $\mathcal{V}$, the degrees of vertices from $\mathcal{W}$, denoted by $Y^*(w_j),\ j=1,2,\ldots,m$, generally are not independent. Only if the X^* are binomially distributed, then the Y^* are independent. In this case the model $\mathcal{B}G_{n,m}\left(n,B(m,p)\right)$ is equivalent to the well known random bipartite graph $\mathcal{G}_{n,m,p}$ on $n+m$ labeled vertices where each of all $n\,m$ possible edges between the sets $\mathcal{V}$ and $\mathcal{W}$ appears independently with a given probability p. Moreover

$$\mathcal{B}G_{n,m}\left(n,\mathcal{P}_{(m)}\right)=\mathcal{B}G_{n,m}\left(n,B(m,p)\right)=\mathcal{B}G_{m,n}\left(m,B(n,p)\right),$$

and $\mathcal{B}G_{n,m}\left(n,\mathcal{P}_{(m)}\right)$ can be considered as a bipartite version of the general digraph model introduced in [5], and many properties of the bipartite model can be proved in the same manner (see [4], [5]).

Each bipartite graph $\mathcal{B}G$ with the 2-partition $(\mathcal{V}, \mathcal{W})$ of the vertex set generates two intersection graphs. The first one with vertex set $\mathcal{V}$ has two vertices joined by an edge if and only if the sets of neighbors of these vertices in $\mathcal{B}G$ have a non-empty intersection or, more generally, this intersection has more than s vertices. The second intersection graph generated by $\mathcal{B}G$ is defined on the vertex set $\mathcal{W}$ analogously. The main purpose of the paper is to study some properties of the two models of random intersection graphs generated by $\mathcal{B}G_{n,m}\left(n, \mathcal{P}_{(m)}\right)$. The first model, $\mathcal{I}G_n^{active}\left(n, \mathcal{P}_{(m)}\right)$, will be called the "active" one, since the vertices from $\mathcal{V}$ choose their neighbors in the original bipartite graph. The second intersection graph model will be treated as the "passive" one since its vertices were chosen in the bipartite graph, and therefore is denoted by $\mathcal{I}G_m^{passive}\left(n, \mathcal{P}_{(m)}\right)$. If $\mathcal{P}_{(m)}$ is a binomial distribution, then $\mathcal{I}G_m^{passive}(n, B(m,p)) = \mathcal{I}G_m^{active}(m, B(n,p))$, and this is the model as introduced in [6].

We can interpret certain types of subgraphs of the intersection graph on the set $\mathcal{V}$ of objects, sharing some properties which are given in $\mathcal{W}$, as object clusters. Similarly, we can define clusters of properties as subgraphs of the intersection graph on $\mathcal{W}$. The respective probability models can be applied in classification theory if we want to test the validity of groups detected by some numerical cluster uncovering algorithm. We also can test possible correlations between the different attributes of the objects in a data sample. For more information on how graph-theoretical concepts can be used in defining a cluster concepts, outlining possible clusters in a data set, and testing the randomness of such clusters, we refer the reader to [1], [2], [3].

As an example of our results, we give the probability distributions of the number of isolated vertices for both models. From these general formulas, the limit distributions for the number of isolated vertices can be derived for special cases of $\mathcal{P}_{(m)}$. As usually, $\mathrm{E}_k(\cdot)$ denotes the k-th factorial moment of a random variable and $(n)_k = n(n-1)\cdots(n-k+1)$.

Theorem 1 *Let* $C^a = C^a\left(\mathcal{I}G_n^{active}\right)$ *denote the number of isolated vertices in* $\mathcal{I}G_n^{active}\left(n, \mathcal{P}_{(m)}\right)$. *Then*

$$\mathrm{E}_k(C^a) = (n)_k \sum_{j=0}^{m} \sum_{(j_1,\ldots,j_k)} \frac{(m)_j}{(m)_{j_1}\cdots(m)_{j_k}} P_{j_1}\cdots P_{j_k}\left[\frac{\mathrm{E}_j(m-X^*)}{(m)_j}\right]^{n-k},$$

where the second sum is over all k*-tuples* $(j_1, j_2, \ldots, j_k)$, *with* $0 \le j_i \le m$ *for* $i = 1, 2, \ldots, k$, $\ j_1 + j_2 + \cdots + j_k = j \le m$, *assuming* $(m)_0 = 1$, $\mathrm{E}_0(\cdot) = 1$.

Theorem 2 *Let* $C^p = C^p\left(\mathcal{I}G_m^{passive}\right)$ *denote the number of isolated vertices in* $\mathcal{I}G_m^{passive}\left(n, \mathcal{P}_{(m)}\right)$. *Then*

$$\mathrm{E}_k(C^p) = (m)_k \left(\frac{k}{m} P_1 + \frac{\mathrm{E}_k(m - X^*)}{(m)_k}\right)^n .$$

For an application of this general model, we will focus on the special case where $\mathcal{P}_{(m)} = (P_d)$ is the degenerate distribution, i.e., where each active vertex in the original bipartite graph chooses exactly d neighbors from $\mathcal{W}$. As we will see, active random intersection graphs for large n and appropriately chosen m and d, behave almost like the classical $\mathcal{G}_{n,p}$-model. This holds, since the edges "appear" almost independently. Let $m = m(n)$, $d = d(n)$, and let $nd^2/m - \log n \to c$ as $n \to \infty$ for any constant c. Then, for sequences $\left(\mathcal{I}G_n^{active}(n, (P_d))\right)_{n\to\infty}$ of active random intersection graphs, the distribution of the number C^a of isolated vertices tends to a Poisson distribution with the parameter $\lambda = e^{-c}$. Theorem 2 implies that if $m = m(n)$, $d = d(n)$, and let $nd/m - \log m \to c$ as $n \to \infty$ for any constant c, then, for sequences $\left(\mathcal{I}G_m^{passive}(n, (P_d))\right)_{n\to\infty}$ of passive random intersection graphs, the distribution of the number C^p of isolated vertices tends to a Poisson distribution with parameter $\lambda = e^{-c}$. From our results one can see that in the case of a degenerate distribution, $\mathcal{P}_{(m)} = (P_d)$, both intersection graph models $\mathcal{I}G_n^{active}(n, (P_d))$ and $\mathcal{I}G_m^{passive}(n, (P_d))$ differ much. This difference becomes even more evident, when the degree distribution is studied. One can easily show that the degree distribution of any vertex in this active random intersection graph is binomially distributed. This is not the case for the passive model. For example, it is evident that in the passive model with degenerate distribution, a vertex either is isolated or it has to belong to a complete subgraph on d vertices.

References

[1] Bock, H.H. (1996): Probabilistic models in cluster analysis. *Computational Statistics and Data Analysis*, **23**, 5–28.

[2] Godehardt, E. (1990): *Graphs as structural models.* Vieweg, Braunschweig.

[3] Godehardt, E., Jaworski, J. (1996): On the connectivity of a random graph. *Random Structures and Algorithms*, **9**, 137–161.

[4] Jaworski, J., Palka, Z. (2001, to appear): Remarks on a General Model of a Random Digraph. *Ars Combinatoria.*

[5] Jaworski, J., Smit, I. (1987): On a Random Digraph. *Annals of Discrete Mathematics*, **33**, 111–127.

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