Moment-based parameter estimation in binomial random intersection graph models

Joona Karjalainen and Lasse Leskelä
Department of Mathematics and Systems Analysis, Aalto University

Abstract

Random intersection graphs (RIG) can be used as parsimonious models of large and sparse networks. We derive moment-based parameter estimators for a class of RIG models and prove their consistency when only a subset of the data is used for estimation.

Random intersection graphs

RIGs are models of undirected and unweighted graphs, where a link is present between two nodes exactly when they share a common attribute (e.g., a hobby or an interest). The model $G(n, m_n, p_n)$ is specified as follows:

- n, the number of nodes
- m_n, the number of attributes
- $p_n \in (0,1)$, the probability that node i has attribute k
- $V_i \subset \{1,2,\dots,m_n\}$, the (random) set of attributes assigned to node i
- Node i links to node j if and only if $|V_i \cap V_j| \neq 0$.
- The attributes are assigned independently, so that $|V_i| \sim Bin(m_n, p_n)$.

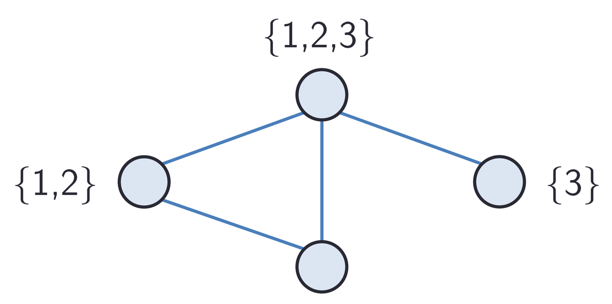


Figure 1: A realization of a random intersection graph with $n = 4$.

We consider a sequence of graphs $G(n, m_n, p_n)$ and its limiting behavior as $n \longrightarrow \infty$. In the limit, we wish to have

- a nontrivial average degree of the nodes, and
- a nontrivial *clustering coefficient* $\mathbb{P}(j \leftrightarrow k \mid i \leftrightarrow j, i \leftrightarrow k)$.

These are achieved when

$$p_n = \frac{\lambda}{\mu} n^{-1} \quad \text{and} \quad m_n = \frac{\mu^2}{\lambda} n, \qquad (1)$$

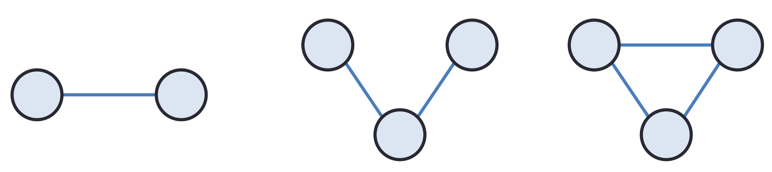
where

- λ is the limiting expected degree of a node, and
- μ is the limiting expected number of attributes of a node.

Question: How can we estimate λ and μ from a single observed network?

Asymptotic subgraph counts

Parameter estimates can be based on counting the numbers of certain subgraphs in the observed network. Consider the following subgraphs:



K_2 S_2 K_3

Let N_{K_2}, N_{S_2} and N_{K_3} be the empirical counts of links, 2-stars and triangles. Under model (1) it holds that

$$\mathbb{E}[N_{K_2}] \sim n^2 \mu^2 m_n^{-1},$$
$$\mathbb{E}[N_{S_2}] \sim n^3 (1+\mu) \mu^3 m_n^{-2},$$
$$\mathbb{E}[N_{K_3}] \sim n^3 \mu^3 m_n^{-2}.$$

Parameter estimators are found by solving for λ and μ and replacing $\mathbb{E}[N_*]$ with N_*. The variances of N_* can be bounded by using the following lemma.

Lemma 1 *The probability that a random intersection graph $G(|G|, m_n, p_n)$ contains a connected subgraph S with $|S|$ nodes satisfies*

$$\mathbb{P}(S \subset G) = O\left(|G|^{|S|} m_n p_n^{|S|}\right).$$

Induced subgraph sampling and consistency

Counting the triangles in the graph with a naïve method requires $O(n^3)$ operations. One may reduce the computation time by only using an *induced subgraph* $G^{(n_0)}$ of the data, i.e., a subset of $n_0 < n$ nodes and the links between them.

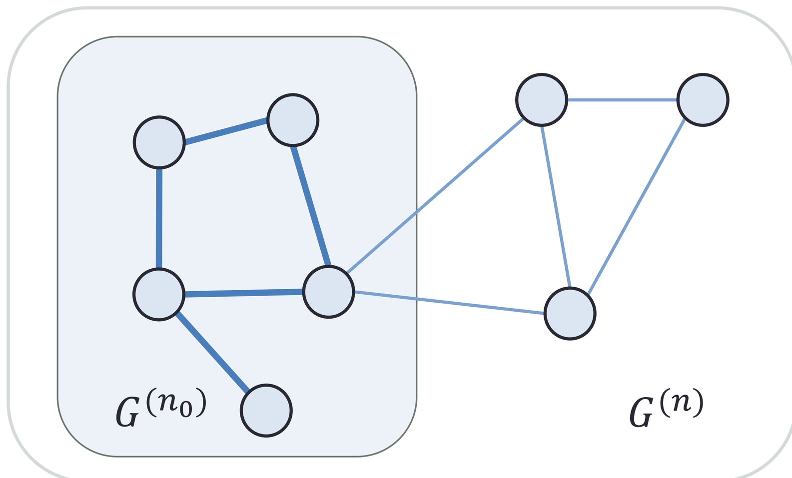


Figure 2: A data set $G^{(n)}$ with $n = 8$ and an induced subgraph $G^{(n_0)}$ with $n_0 = 5$.

Using the asymptotic subgraph counts we obtain the following estimators for λ and μ.

$$\hat{\lambda} = \frac{n}{n_0^2} \sum_{i \in G^{(n_0)}} \deg_{G^{(n_0)}}(i)$$

$$\hat{\mu}_1 = \frac{N_{S_2}(G^{(n_0)})}{3N_{K_3}(G^{(n_0)})} - 1$$

$$\hat{\mu}_2 = \left(\frac{n_0 N_{S_2}(G^{(n_0)})}{2N_{K_2}(G^{(n_0)})^2} - 1\right)^{-1}$$

Question: Given a sufficiently large n and n_0, are these estimators close to the true parameter values?

The following theorems confirm that this is the case, in a suitable sense:

Theorem 1 *Estimator $\hat{\lambda}$ is consistent, i.e., $\hat{\lambda} \xrightarrow{p} \lambda$, when $n_0 \gg n^{1/2}$. Moreover,*

$$\hat{\lambda}(G^{(n_0)}) = \lambda + O_p\left(\frac{n^{1/2}}{n_0}\right).$$

Theorem 2 *Estimators $\hat{\mu}_1$ and $\hat{\mu}_2$ are consistent when $n_0 \gg n^{2/3}$.*

The proofs are based on the second moment method, Lemma 1 and the continuous mapping theorem.

Simulations

- Parameters are estimated once for each $n = 50, 70, \dots, 1000$.
- Two sets of parameters, $(\lambda = 9, \mu = 3)$ and $(\lambda = 2, \mu = 0.5)$.
- The biases decrease rapidly as the size of the graph grows.
- $\hat{\mu}_1$ seems to be better than $\hat{\mu}_2$, but counting the triangles takes time.
- The data can be much larger ($n \approx 10^5$ with a simple implementation).

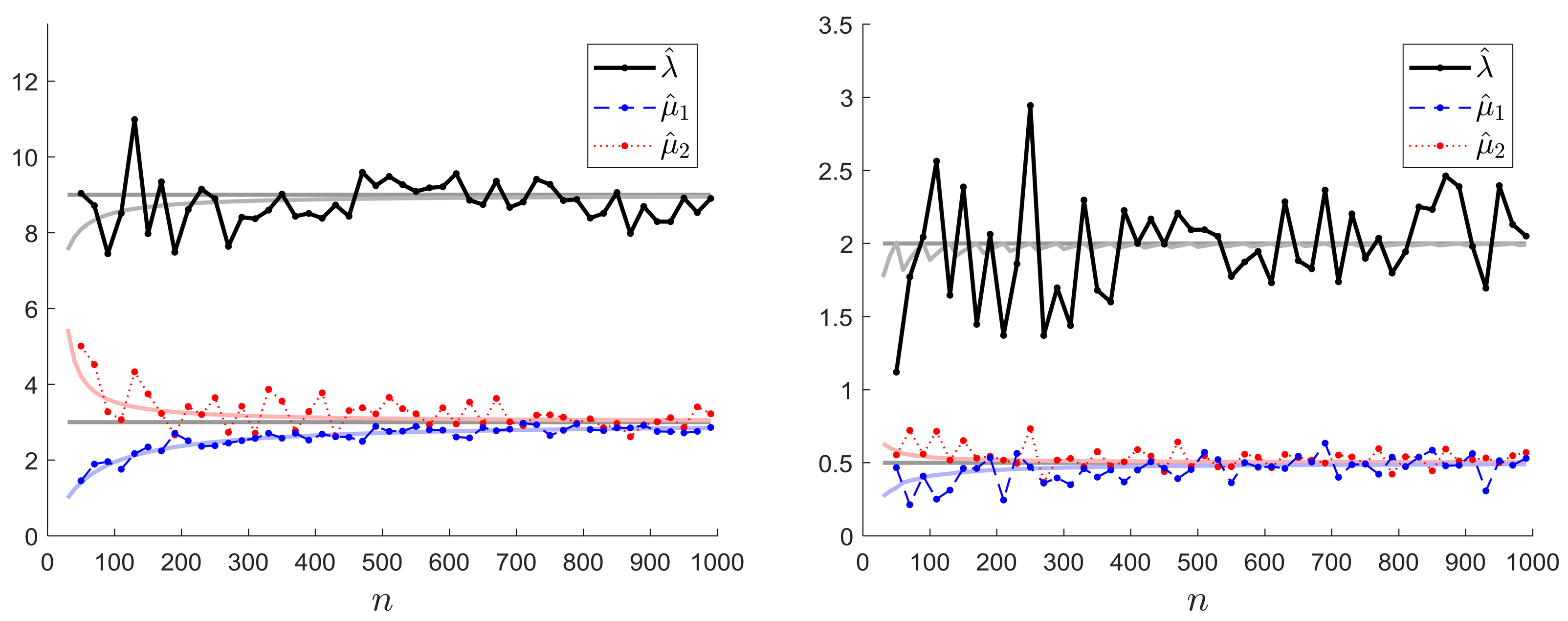


Figure 3: Estimated values for parameters (λ, μ) with $n_0 = n$ for simulated graphs of sizes $n = 50, \dots, 1000$.

References:

[1] J. Karjalainen and L. Leskelä: Moment-based parameter estimation in binomial random intersection graph models. 14th Workshop on Algorithms and Models for the Web Graph, 2017. arXiv:1704.04278

[2] M. Bloznelis: Degree and clustering coefficient in sparse random intersection graphs. Ann. Appl. Probab. 23(3), 1254–1289, 2013. http://dx.doi.org/10.1214/12-AAP874

[3] M. Karonski, E.R. Scheinerman and K.B. Singer-Cohen, On random intersection graphs: The subgraph problem, Combin. Probab. Comput. 8, 131-159, 1999.