Stein's Method for Erdös-Rényi Mixture Graph Models

Anum Fatima [1] Gesine Reinert [2]

[1]University of Oxford & Lahore College for Women University, Lahore
[2]University of Oxford & Alan Turing Institute, London

COMMONWEALTH SCHOLARSHIPS

Background

Reinert & Ross [1] provide a general bound on the distance between two arbitrary distributions of sequences of Bernoulli random variable using Stein's method and used it to give the bound on the distance between Exponential Random Graph Model (ERGM) and a corresponding Erdös Rényi (ER) model. We extend this result to Erdös-Rényi Mixture Graph (ERMG) models, a special case of Stochastic Block Model (SBM) where an SBM is a graph whose actors or vertices are partitioned into subgroups and these subgroups are called blocks. The probability of an edge between two vertices say, u and v depends only on the block(s), the two vertices belong to, where all the edges are created independently. In ERMG the group identification of all the vertices is known and so is the probability of an edge between any two vertices.

Erdös-Rényi Mixture Graph

Let $\mathbf{X} = \{X_{u,v} = \mathbb{I}((u,v) \in \mathcal{E}) \forall u, v \in \mathcal{V}\}$ denote the adjacency matrix of a graph $\mathcal{G}_n$ with vertex set $\mathcal{V}$ and edge set $\mathcal{E}$, then the $ERMG(n, L, \mathbf{p})$ model for graph $\mathcal{G}$ is

$$\mathbb{P}(\mathbf{X} = \mathbf{x}) = \prod_{i \leq j} (1 - p_{i,j})^{N_{i,j}} \left(\frac{p_{i,j}}{1 - p_{i,j}} \right)^{t_{i,j}(\mathbf{x})}.$$

Here $t_{i,j}$ is the total number of edges between any two blocks i and j, $N_{i,j} = \binom{n_i}{2}$ for $i = j$ and $N_{i,j} = n_i n_j$ when $i \neq j$ where n_i is the number of vertices in block i and $\sum_{i=1}^{L} n_i = n$. Also $N = \sum_{i \leq j}^{L} N_{i,j} = \binom{n}{2}$. The vector $g \in \{1, \ldots, L\}^n$ indicates the types of vertices. We denote $\mathbf{p} : \{p_{i,j}; i, j = 1, \ldots, L\}$ where $p_{i,j} = \mathbb{P}((u,v) \in \mathcal{E} | u \text{ is of type } i, v \text{ is of type } j) = \mathbb{P}((u,v) \in \mathcal{E} | g_u = i, g_v = j)$.

Stein's Method

Stein's method for probability distributions explicitly bounds the distance between a probability distribution $\mathcal{L}(X)$ of interest and a usually well-understood approximating distribution $\mathcal{L}_0(Z)$, often called the *target* distribution. Here we use distances between probability distributions of the form

$$d_{\mathcal{H}}(\mathcal{L}(X), \mathcal{L}_0(Z)) := \sup_{h \in \mathcal{H}} |\mathbb{E}h(X) - \mathbb{E}h(Z)| \leq \sup_{f \in \mathcal{F}(\mathcal{H})} |\mathbb{E}\mathcal{T}f(X)|$$

where $\mathcal{H}$ is a set of test functions. In this equation $\mathcal{T}f(x)$ is a *Stein operator* for the distribution $\mathcal{L}_0$, with an associated *Stein class* $\mathcal{F}(\mathcal{T})$ of functions such that

$$\mathbb{E}[\mathcal{T}f(Z)] = 0 \text{ for all } f \in \mathcal{F}(\mathcal{T}) \iff Z \sim \mathcal{L}_0.$$

where $\mathcal{F}(\mathcal{H}) = \{f_h | h \in \mathcal{H}\}$ is the set of solutions of the Stein equation $h(x) - \mathbb{E}[h(Z)] = \mathcal{T}f(x)$ for the test functions $h \in \mathcal{H}$, with $Z \sim \mathcal{L}_0$.

Stein method for Erdös-Rényi Mixture Graph Models

Reinert & Ross [1] used the generator of a continuous time Glauber dynamics Markov chain, $(X(t))_{t \geq 0}$ for $\mathcal{L}(\mathbf{X})$ where $\mathbf{X} \in \{0,1\}^N$ is a random vector, to get a Stein operator for Exponential random graph models. The generator is

$$\mathcal{A}f(x) = \frac{1}{N} \sum_{s \in [N]} \left[q(\mathbf{x}^{(s,1)} | \mathbf{x}^{(s)}) \Delta_s f(\mathbf{x}) + \left(f(\mathbf{x}^{(s,0)}) - f(\mathbf{x}) \right) \right].$$

Here $\mathbf{x}^{(s,1)}$ is a vector with 1 in sth coordinate and is otherwise same as $\mathbf{x}$ and $q_{\mathbf{X}}(\mathbf{x}^{(s,1)} | \mathbf{x}) := \mathbb{P}\left(X_s = 1 | (X_u)_{u \neq s} = (x_u)_{u \neq s}\right)$. Also $\Delta_s f(\mathbf{x}) = f(\mathbf{x}^{(s,1)}) - f(\mathbf{x}^{(s,0)})$.

Building on the results from Reinert & Ross (2019) we use this operator to get a Stein equation for an $ERMG(n, L, \mathbf{p})$ model for any $h : \{0,1\}^N \to \mathbb{R}$ given as

$$\frac{1}{N} \sum_{s \in [N]} \left[(p_{i,j}) \Delta_s f(\mathbf{x}) + \left(f(\mathbf{x}^{(s,0)}) - f(\mathbf{x}) \right) \right] = h(\mathbf{x}) - \mathbb{E}h(\mathbf{X}),$$

with the solution of this Stein equation, $f_h(\mathbf{x}) := -\int_0^\infty \mathbb{E}\left[h(\mathbf{X}(t)) - \mathbb{E}h(\mathbf{X}) | \mathbf{X}(0) = \mathbf{x} \right] dt$.

Further for this solution of the ERMG Stein equation, we bound

$$|\Delta_s f_h(\mathbf{x})| \leq \|\Delta_s h\| N.$$

Comparison with other Graph Models

Let $X, Y \in \{0,1\}^N$ be random vectors and $h : \{0,1\}^N \to \mathbb{R}$, then Lemma 2.4 in Reinert & Ross (2019) suggests using the following to bound the distance between the distributions of X and Y.

$$|\mathbb{E}h(\mathbf{X}) - \mathbb{E}h(\mathbf{Y})| \leq \frac{1}{N} \sum_{s \in [N]} \mathbb{E}\left[|q_X(\mathbf{Y}^{(s,1)} | \mathbf{Y}) - q_Y(\mathbf{Y}^{(s,1)} | \mathbf{Y})| |\Delta_s f_h(\mathbf{Y})| \right],$$

where f_h is the solution of ERMG Stein equation.

Erdös-Rényi Models

Let $\mathbf{Z} \sim ER(n, a^*)$ and $\mathbf{X} \sim ERMG(n, L, \mathbf{p})$ represent the adjacency matrices of undirected simple random graphs $\mathcal{G}_n^*$ and $\mathcal{G}_n$ respectively defined on the same vertex set $\mathcal{V}$, for a test function h such that $\|\Delta h\| = O(n^{-2})$, we get

$$|\mathbb{E}h(\mathbf{X}) - \mathbb{E}h(\mathbf{Z})| \leq \|\Delta h\| \sum_{i \leq j}^{L} N_{i,j} |p_{i,j} - a^*|$$

For example, if we choose h to be the proportion of a fixed type of sub-graphs then $\|\Delta h\| = O(n^{-2})$ and the bound is small when the weighted average of the difference between edge probabilities under the two graph models is small. We then see what choice of a^* gives the smallest bound.

Exponential Random Graph Models

Let $\mathbf{X} \sim ERMG(n, L, \mathbf{p})$ and $\mathbf{Y} \sim ERGM(\beta)$. For a test function h such that $\|\Delta h\| = O(n^{-2})$, using the bound from [1] for $\frac{1}{2}|\Phi|'(1) < 1$ and the bound above, we bound the distance as

$$d_{\mathcal{H}}(\mathcal{L}(\mathbf{X}), \mathcal{L}(\mathbf{Y})) \leq \|\Delta h\| \left\{ \sum_{i \leq j}^{L} N_{i,j} |p_{i,j} - a^*| + N \left(4 \left(1 - \frac{1}{2} |\Phi|'(1) \right) \right)^{-1} \sum_{l=2}^{k} \beta_l \sqrt{\text{Var}(\Delta_{12} t_l(\mathbf{Z}))} \right\}.$$

where a^* is the solution of the equation $\phi(a^*) = a^*$ satisfying $\phi'(a^*) < 1$ where

$$\phi(a^*) = \frac{1 + \tanh(\Phi(a^*))}{2} \quad , \quad \Phi(a^*) := \sum_{l=1}^{k} \beta_l e_l (a^*)^{e_l - 1} \quad , \quad t_l(\mathbf{x}) = \frac{t(\mathbf{H}_l, \mathbf{x})}{n(n-1)...(n - v_l + 3)}$$

and for $n \in \mathbb{N}$, $\mathbf{H}_1, \mathbf{H}_2, ..., \mathbf{H}_k$ are k connected graphs with $\mathbf{H}_1$ a single edge. Also $\beta = (\beta_1, \beta_2, ..., \beta_k)$ with $\beta_l \in \mathbb{R}$, $v_l := |V(\mathbf{H}_l)|$ for $l = 1, ..., k$ and e_l is the number of edges in $\mathbf{H}_l$. For more details on the results on ERGM see Reinert & Ross (2019).

Other Erdös-Rényi Mixture Models

If $\mathbf{X} \sim ERMG(n, L, \mathbf{p})$ and $\mathbf{Y} \sim ERMG(n, L^*, \mathbf{p}^*)$ are the adjacency matrices of graphs $\mathcal{G}_n$ and $\mathcal{G}_n^*$ defined on the same vertex set $\mathcal{V}$ with block identifications $g_\cdot$ and $l_\cdot$ respectively, for a test function h, our bound on the distance between two graph models is

$$|\mathbb{E}h(\mathbf{X}) - \mathbb{E}h(\mathbf{Y})| \leq \|\Delta h\| \sum_{s \in [N]} |p_{g_s} - p^*_{\ell_s}|.$$

We choose our test function h such that $\|\Delta h\| = O(n^{-2})$.

Special Cases

Case: 1

Here we assume that the two graphs $\mathcal{G}_n$ and $\mathcal{G}_n^*$ have the same blocks schemes so that $g_u = g_u^*$ for any $u \in \mathcal{V}$, $L^* = L$, $n_i^* = n_i$ for all $i = 1, \ldots, L$, but with different edge probabilities $\mathbf{p}$ and $\mathbf{p}^*$. For such $\mathbf{x} \in \mathcal{G}_n$ and $\mathbf{y} \in \mathcal{G}_n^*$, the bound simplifies to

$$|\mathbb{E}h(\mathbf{X}) - \mathbb{E}h(\mathbf{Y})| \leq \|\Delta h\| \sum_{i \leq j}^{L} N_{i,j} |p_{i,j} - p^*_{i,j}|,$$

where $N_{i,j} = \binom{n_i}{2}$ for $i = j$ and $N_{i,j} = n_i n_j$ for $i \neq j$ is the possible number of edges of type (i,j).

Case: 2

Here we consider a case where some or all of L blocks of vertices in $\mathcal{G}_n^*$ are divided into further subgroups with a corresponding nested structure of the group assignments. If $G_1, G_2, \cdots, G_L$ denote the blocks of vertices, such that $\cup_{i=1}^{L} G_i = \mathcal{V}$, we say the subgroup $G_i^{(a)}$ is the subset of block G_i i.e. $G_i^{(a)} \subseteq G_i$ such that $\cup_{a=1}^{c_i} G_i^{(a)} = G_i$ where c_i is the number of subgroups of block G_i. For a partitioned block G_i we denote $|G_i^{(a)}| = n_i^{(a)}$ and the number of possible edges between $G_i^{(a)}$ and $G_j^{(b)}$ by $N_{i,j}^{(a,b)}$, so that $N_{i,j}^{(a,b)} = n_i^{(a)} n_j^{(b)}$ for $a \neq b$ and $N_{i,j}^{(a,b)} = \binom{n_i^{(a)}}{2}$ for $i = j, a = b$ for $i, j = 1, \ldots, L$, $a = 1, \ldots, c_i$ and $b = 1, \ldots, c_j$.

Also the probability of an edge between a vertex pair $s = (u,v)$ is

$$\mathbb{P}(X_s = 1) = \sum_{i \leq j}^{L} \mathbb{P}(X_s = 1) \mathbb{I}(u \in G_i^{(a)}, v \in G_j^{(b)}) = p_{i,j}^{*(a,b)}.$$

As $p_{i,j}^{*(a,b)}$ are known and $\mathbf{x} \in \mathcal{G}_n$ and $\mathbf{y} \in \mathcal{G}_n^*$, the bound refines to

$$|\mathbb{E}h(\mathbf{X}) - \mathbb{E}h(\mathbf{Y})| \leq \|\Delta h\| \left\{ \sum_{i=1}^{L} \sum_{a \leq b}^{c_i} N_{i,i}^{(a,b)} |p_{i,i} - p_{i,i}^{*(a,b)}| + \sum_{i<j}^{L} \sum_{a=1}^{c_i} \sum_{b=1}^{c_j} N_{i,j}^{(a,b)} |p_{i,j} - p_{i,j}^{*(a,b)}| \right\}.$$

Note that for a non-partitioned group we have $c_i = 1$, $G_i = G_i^{(1)}$, and $n_i = n_i^{(1)}$.

Graphon Models

The graphon model of Lovász & Szegedy [2], called the W-graph, is a model based on the symmetric and measurable *graphon* function $W : [0,1]^2 \to [0,1]$ that is important in the study of dense graphs. A W-graph on n vertices is created taking an independent random variable $U_i \sim U[0,1]$ and assigning it to the vertex i, for all $i = 1, \ldots, n$ and than placing an edge between a vertex pair $s = (i,j)$ with probability $p_{i,j} = W(U_i, U_j)$.

An SBM is a W-graph for which the unit square is divided into $L \times L$ blocks and the graphon function is piecewise constant for these blocks setting W equals to $p_{i,j}$ on $(i,j)^{th}$ block.

We are currently working to bound the distance between such a W-graph model and a $ERMG(n, L, \mathbf{p})$ model, in particular the case when the W-graph has one additional vertex compared to the $ERMG(n, L, \mathbf{p})$ graph.

Example

As an example, we consider the Political blog network data from [3]. The data consists of $n = 1490$ blogs of which $n_1 = 758$ are labeled *Liberal* and $n_2 = 732$ are labeled *Conservative*. After removing self-loops and multiple edges there are 16715 edges in the whole network. The estimated edge probabilities for this network are $\hat{p}_{1,1} = 0.02545$; $\hat{p}_{1,2} = 0.002839$ and $\hat{p}_{2,2} = 0.02930$. We propose using $a^* = 16715/\binom{1490}{2} = 0.015068$. Using these estimates and $\|\Delta h\| = \frac{1}{N}$ our bound on the distance between $ER(1490, 0.015068)$ and the corresponding ERMG model gives

$$\frac{1}{\binom{1490}{2}} \left\{ \binom{758}{2} |0.02545 - 0.015068| + (758)(732) |0.002839 - 0.015068| + \binom{732}{2} |0.02930 - 0.015068| \right\} = 0.0122339.$$

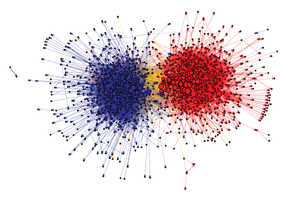

Figure 1. Political Blog network. Figure from [3].

References

[1] G. Reinert and N. Ross.
Approximating stationary distributions of fast mixing Glauber dynamics, with applications to exponential random graphs.
The Annals of Applied Probability, 29(5):3201–3229, 2019.

[2] Limits of dense graph sequences.
Journal of Combinatorial Theory, Series B, 96(6):933–957, 2006.

[3] L. A. Adamic and N. Glance.
The political blogosphere and the 2004 U.S. election: Divided they blog.
In *Proceedings of the 3rd International Workshop on Link Discovery*, LinkKDD '05, page 36–43, New York, USA, 2005. Association for Computing Machinery.