Notes on "Network Representation Using Graph Root Distributions" by Jing Lei

Presenter: Jesús Arroyo

August 11th, 2020

This paper [2] studies the graph root distribution model (GRD), which is very similar generalized RDPG [3], in the sense that the edge probabilities are given by indefinite inner products of random vectors. The main difference is that here the latent positions can have infinite dimension, which can be used to represent some graphons. The paper first studies distributions that can be represented with this model, and then the estimation error of a truncated weighted spectral embedding (ASE).

1 Graphon model

- Let $\mathbf{A} = (A_{ij}, i \in \mathbb{N}, j \in \mathbb{N})$ be a two-dimensional infinite binary array of random variables with exchangeable rows and columns, that is, for any finite permutation of the nodes $\sigma : [N] \to [N]$, it holds that
$$(A_{ij}, i \in \mathbb{N}, j \in \mathbb{N}) \overset{d}{=} (A_{\sigma(i)\sigma(j)}, i \in \mathbb{N}, j \in \mathbb{N}).$$
That is, relabeling the rows and columns (nodes) doesn't change the distribution of the array. This node exchangeability formalizes the idea of a vertex sampling scheme, in which we assume that the n vertices that were observed in an adjacency matrix $A \in \{0,1\}^{n\times n}$ are representative from the population [1].

- According to the Aldous-Hoover theorem, the distribution of $\mathbf{A}$ can be obtained by first sampling i.i.d. random variables $s_1, s_2, \ldots, \sim U(0,1)$, and then then sample independently sample
$$A_{ij} \sim \text{Ber}(W(s_i, s_j))$$
for some symmetric function $W : [0,1] \times [0,1] \to [0,1]$, which is called a *graphon*. This function is not identifiable, but an equivalence relation can be defined by looking at all measure preserving transformations h such that $W(s,t) = W_2(h(s), h(t))\ \forall s, t \in [0,1]$.

- The *cut-distance* measures the dissimilarity between two graphons as
$$\delta_\square(W_1, W_2) = \inf_{h_1,h_2} \sup_{S\times S' \subset [0,1]^2} \left| \int_{S\times S'} W_1(h_1(s), h_1(t)) - W_2(h_2(s), h_2(t))\ ds\ dt \right|.$$
The infimum in this definition removes the non-identifiability of the graphons. The set $S \times S'$ can be thought as a subgraph, so this distance finds the subgraph in which the graphons are the most different. The cut-distance is zero if and only if the graphons are equivalent.

- Node exchangeable graphs are either dense (expected node degree $\approx n$) or empty. A parameter ρ_n is introduced to model the sparsity of the network, so that $A_n \sim \text{Ber}(\rho_n W)$, but I won't focus on this part.

2 Graph root distribution

-
- The graph root distribution (GRD) model defines a distribution over $\mathcal{K}$ such that if $X_1, X_2 \sim F$ then
$$\mathbb{P}_F(< X_1, X_2 >_{\mathcal{K}} \in [0,1]) = 1.$$
Here, $\mathcal{K}$ is a *Krein space* which is the direct sum of two Hilbert spaces $\mathcal{H}_+$ and $\mathcal{H}_-$, so any vector $x \in \mathcal{K}$ can be expressed as $x = (y, z)$, with $y \in \mathcal{H}_+, z \in \mathcal{H}_-$ and
$$< x_1, x_2 >_{\mathcal{K}} = < y_1, y_2 >_{\mathcal{H}_+} - < z_1, z_2 >_{\mathcal{H}_-} .$$
Thus, if these Hilbert spaces are finite dimensional, then GRDPG and GRD are the same model. Hence, we know that all the models listed in the paper (SBM, degree corrected SBM, mixed membership, etc.) are special cases of GRD.
- A graphon can be spectrally decomposed as
$$W(s,t) = \sum_{j=1}^{\infty} \lambda_j \phi_j(s)\phi_j(t) - \sum_{k=1}^{\infty} \gamma_k \psi_k(s)\psi_k(t),$$
where $\lambda_1 \geq \ldots$ and $\gamma_1 \geq \ldots$ are non-negative eigenvalues, and $\{\psi_j\} \cup \{\phi_k\}$ are mutually orthogonal functions. This is again analogous to the GRDPG for finite n, in which A_n can be described with the eigenvalues and eigenvectors of a symmetric matrix of bounded rank.
- One of the main results of this paper is to show that if W admits a *strong spectral decomposition*, which means that
$$\sum_{j=1}^{\infty} \lambda_j \phi_j^2(s) + \gamma_j \psi_j^2(s) < \infty \qquad \forall s,$$
then W can be represented as a GRD.
- Proposition 3.2 shows that all *trace-class operators* (which are the ones for which the sum of the absolute value of its eigenvalues converges) admit a strong spectral decomposition, and hence they are GRD. Some examples
  - Finite rank
  - Smooth graphons, in the sense that for some $a > 1/2$,
$$|W(x,y) - W(x,y')| \leq C|y - y'|^a.$$
This example is important because Proposition 3.3 shows that these graphons are dense in the space of continuous graphons (and hence can approximate any continuous graphon with arbitrary precision).
  - Continuous positive semidefinite graphons (or the difference between two positive semidefinite graphons).
- A GRD is identifiable up to some indefinite orthogonal transformation. This is again equivalent to GRDPGs. The author chooses the canonical representation as the one that diagonalizes the covariance operator.
- To study the space of graphons, the author defines the orthogonal Wasserstein distance as
$$d_{OW}(F_1, F_2) = \inf_{v \in \mathcal{V}(F_1, F_2)} \inf_{Q} \mathbb{E}_{(Z_1, Z_2) \sim v} \|Z_1 - QZ_2\|,$$

where Q is an indefinite orthogonal matrix, and $\mathcal{V}(F_1, F_2)$ is the space of joint distributions in $\mathcal{K} \times \mathcal{K}$ with marginal distributions F_1 and F_2 on the appropriate Hilbert spaces. The first infimum removes the non-identifiability issue, while the second one finds the joint distribution that minimizes the difference. When $F_1 = F_2$, one can choose a distribution such that $Z_1 = Z_2$, and hence $d_{OW} = 0$. Note that this distance depends on the "latent positions", so it is more directly linked to the GRD model than the cut-distance, and Theorem 3.8 shows that the OW distance is in fact an upper bound for the cut-distance.

3 Estimation

- To estimate the GRD model, the authors use an analogous method to the adjacency spectral embedding (ASE), i.e., given $\mathbf{A}_n$ (a graph with n nodes) from $\mathbf{A}$, the authors define a truncated embedding as

$$\hat{X} = (\hat{\lambda}_1^{1/2}\hat{U}_{\cdot 1}, \dots, \hat{\lambda}_p^{1/2}U_{\cdot p_1}),$$
$$\hat{Y} = (\hat{\gamma}_1^{1/2}\hat{V}_{\cdot 1}, \dots, \hat{\gamma}_p^{1/2}\hat{V}_{\cdot p_2}),$$

  where $\hat{U}, \hat{V}$ are eigenvectors of $\mathbf{A}_n$, and $\hat{\lambda}_1, -\hat{\gamma}_1, \dots$ are positive and negative eigenvalues respectively.

- To study the reconstruction error of the "ASE", there are three assumptions introduced by the author:

  1. A canonical representation in which the true latent positions are equal to the scaled eigenvectors, i.e., $\mathbb{E}_{X\sim F}(X_j^2) = \lambda_j$ and so on.
  2. Polynomial decay on the eigenvalues and eigengaps, i.e., $\lambda_j \sim \frac{1}{j^\alpha}$ and $\lambda_j - \lambda_{j+1} \gtrsim \frac{1}{j^\beta}$, for $1 < \alpha \leq \beta$ (similarly for γ). This is for convenience of the presentation, and the author claims that having zeros in the eigenvalues or eigengaps will be too cumbersome.
  3. Finite fourth moment of the latent positions $\mathbb{E}_{Z\sim F}\|Z\|^4$, required to estimate the covariance operator.

- There are three sources of error in the GRD estimation process: the error in estimating the full distribution of the latent positions F using a sample of n nodes ($\hat{F}$), the approximation error of using a truncation of the spectrum $F^{(p)}$ to estimate the full spectrum F, and the error introduced by using the spectral decomposition of A_n ($\hat{F}_A^{(p)}$)

  1. Theorem 4.1 calculates the estimation error between $\hat{F}_A^{(p)}$ and $\hat{F}^{(p)}$ (the empirical distribution of the estimated latent positions and the one of the p-truncation of the true latent positions, both for n nodes)
  2. Lemma 4.2 bounds the approximation error between $F^{(p)}$ and F
  3. Lemma 4.3 bounds the expected estimation error between $\hat{F}^{(p)}$ and $F^{(p)}$ (i.e., the error in the distribution of the truncated latent positions, using the empirical distribution with n nodes and the full distribution).
  4. Finally, Theorem 4.4 gives the GRD estimation error, which is the sum of the errors in the previous three results. That is,

$$\delta_W(\hat{F}_A^{(p)}, F) = O_P\left(\underbrace{n^{-\frac{\alpha-1}{4\beta}} + p^{\beta+\frac{1}{2}}n^{-1/2}}_{d_W(\hat{F}_A^{(p)}, \hat{F}^{(p)})} + \underbrace{n^{-\frac{1}{p}}}_{d_W(\hat{F}^{(p)}, F^{(p)})} + \underbrace{p^{-\frac{\alpha-1}{2}}}_{d_W(F^{(p)}, F)}\right).$$

  This error goes to zero if the embedding dimension satisfies $p \to \infty$ and $p = o(\log n)$. Theorem 4.1 actually requires $p = o(n^{1/(2\beta+\alpha)})$.

References

[1] Harry Crane. *Probabilistic foundations of statistical network analysis*. CRC Press, 2018.

[2] Jing Lei. Network Representation Using Graph Root Distributions. 2018.

[3] Patrick Rubin-Delanchy, Carey E Priebe, Minh Tang, and Joshua Cape. A statistical interpretation of spectral embedding: the generalised random dot product graph. *arXiv preprint arXiv:1709.05506*, 2017.