

Community Detection for the Generalized Random Dot Product Graph

Dissertation Prospectus

John Koo

Graph and network data, in which samples are represented not as a collection of feature vectors but as relationships between pairs of observations, are increasingly widespread in various fields ranging from sociology to computer vision. One common goal of analyzing graph data is community detection or graph clustering, in which the graph is partitioned into disconnected subgraphs in an unsupervised yet meaningful manner (e.g., by optimizing an objective function or recovering unobserved labels). Because traditional clustering techniques were developed for data that can be represented as vectors, they cannot be applied directly to graphs. Common examples of such clustering techniques include K -means clustering [5] and Gaussian Mixture Models [3]. In addition, these techniques assume that the vectors are sampled from a specific type of distribution (whether implicitly in the case of K -means clustering or explicitly in the case of GMM). In this research, we investigate the use of a family of spectral decomposition based approaches for community detection in block models (random graph models with inherent community structure). First, we demonstrate that under the Generalized Random Dot Product Graph [8] framework, all graphs generated by Block Models (generative graph models with inherent community structure) can be represented as collections of vectors in Euclidean space via the Adjacency Spectral Embedding [1, 7]. Then, noting the particular structure or configuration that the vectors take, we select appropriate clustering techniques to apply on the vector representation of the graph. Finally we derive the asymptotic properties of these methods to show that our methods are consistent or achieve desirable properties with high probability. To illustrate this approach, we primarily focus on the Popularity Adjusted Block Model [9].

Our work is essentially about connecting two well-known families of generative graph models, the Block Model and the (Generalized) Random Dot Product Graph [8] in order to perform community detection. It is straightforward to show that all Block Models are various cases of the GRDPG. This fact has been leveraged to develop community detection methods [1, 7, 8] for the Stochastic Block Model [6] and Degree Corrected Block Model [4]. Our work begins by extending this to the Popularity Adjusted Block Model. We show that under the GRDPG framework, the latent configuration that generates the PABM consists of orthogonal subspaces. Then we show that two community detection algorithms arise naturally from this latent configuration: Orthogonal Spectral Clustering, which is an algorithm of our own design, and an existing algorithm, Sparse Subspace Clustering [2]. We then extend this work to more general configurations in the latent space that imply community structure. The overall goal of our research is to develop a general framework or approach to community detection for a wide range of Block Models.

The overall generative model for inducing community structure can be described by the following.

Let $(A, X) \sim \text{GRDPG}_{p,q}(F, n)$ such that:

1. Define functions $\gamma_1, \dots, \gamma_K$ such that each $\gamma_k : [0, 1]^r \mapsto \mathbb{R}^d$ and $\gamma_k(t) \neq \gamma_l(t)$ when $k \neq l$.
2. Sample labels $Z_1, \dots, Z_n \stackrel{\text{iid}}{\sim} \text{Categorical}(\pi_1, \dots, \pi_K)$.
3. Sample $T_1, \dots, T_n \stackrel{\text{iid}}{\sim} D$ with support $[0, 1]^r$.
4. Set latent positions $X_i = \gamma_{Z_i}(T_i)$ and $X = \begin{bmatrix} X_1 & \cdots & X_n \end{bmatrix}^\top$.
5. $A \sim \text{BernoulliGraph}(X I_{p,q} X^\top)$

The SBM, DCBM, and PABM are all special cases of this generative model. Based on the structure of the γ_k functions that correspond to each of the SBM, DCBM, and PABM, there are clear choices for clustering algorithms on the embeddings of each of these Block Models. Extensions of these results may consider which clustering algorithms result in consistent estimators given restrictions on each of the γ_k , and how to approach community detection when γ_k are unknown altogether.

The estimated timeline of completion is as follows:

1. Literature review and proofs of main theorems: December 2022
2. Simulations, real data analyses, R package: March 2022
3. Dissertation Completion: April 2022

References

- [1] Avanti Athreya, Donniell E. Fishkind, Minh Tang, Carey E. Priebe, Youngser Park, Joshua T. Vogelstein, Keith Levin, Vince Lyzinski, Yichen Qin, and Daniel L Sussman. Statistical inference on random dot product graphs: a survey. *Journal of Machine Learning Research*, 18(226):1–92, 2018. URL <http://jmlr.org/papers/v18/17-448.html>.
- [2] Ehsan Elhamifar and Rene Vidal. Sparse subspace clustering. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2790–2797, 2009. doi: 10.1109/CVPR.2009.5206547.
- [3] Chris Fraley and Adrian E Raftery. Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association*, 97(458):611–631, 2002. doi: 10.1198/016214502760047131. URL <https://doi.org/10.1198/016214502760047131>.
- [4] Brian Karrer and M. E. J. Newman. Stochastic blockmodels and community structure in networks. *Physical Review E*, 83(1), Jan 2011. ISSN 1550-2376. doi: 10.1103/physreve.83.016107. URL <http://dx.doi.org/10.1103/PhysRevE.83.016107>.
- [5] S. Lloyd. Least squares quantization in pcm. *IEEE Transactions on Information Theory*, 28(2): 129–137, 1982. doi: 10.1109/TIT.1982.1056489.
- [6] François Lorrain and Harrison C. White. Structural equivalence of individuals in social networks. *The Journal of Mathematical Sociology*, 1(1):49–80, 1971. doi: 10.1080/0022250X.1971.9989788. URL <https://doi.org/10.1080/0022250X.1971.9989788>.
- [7] Vince Lyzinski, Daniel L. Sussman, Minh Tang, Avanti Athreya, and Carey E. Priebe. Perfect clustering for stochastic blockmodel graphs via adjacency spectral embedding. *Electron. J. Statist.*, 8(2):2905–2922, 2014. doi: 10.1214/14-EJS978. URL <https://doi.org/10.1214/14-EJS978>.

- [8] Patrick Rubin-Delanchy, Joshua Cape, Minh Tang, and Carey E. Priebe. A statistical interpretation of spectral embedding: the generalised random dot product graph, 2017.
- [9] Srijan Sengupta and Yuguo Chen. A block model for node popularity in networks with community structure. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 80(2): 365–386, March 2018. ISSN 1369-7412. doi: 10.1111/rssb.12245.