Clustering Bipartite Graphs with the Generalized Power Method

Guillaume Braun* Inria Lille

Hemant Tyagi* Inria Lille

1 Introduction

The interactions between objects of two different types can be naturally encoded as a bipartite graph where nodes correspond to objects and edges to the links between the objects of different type. One can find examples of such data in various fields, e.g., interactions between customers and products in e-commerce, interactions between plants and pollinators, investors and assets networks, judges vote predictions and constraint satisfaction problems.

Clustering is one of the most important analysis tasks on bipartite graphs aimed at gathering nodes that have similar connectivity profiles. To this end, several methods have been proposed in the literature, e.g., convex optimization approaches, spectral methods, modularity function maximization and variational approaches. The performance of the algorithms are generally evaluated under the Bipartite Stochastic Block Model (BiSBM), a variant of the Stochastic Block Model (SBM), where the partitions of the rows and the columns are decoupled. In particular, edges are independent Bernoulli random variables with parameters depending only on the communities of the nodes.

In the setting where the number of type II nodes (n_2) is of a different order than the number of type I nodes (n_1), classical methods can fail. In particular, when the bipartite graph is very sparse, and assuming w.l.o.g $n_2 \gg n_1$, it becomes impossible to consistently estimate the latent partition of the type II nodes, whereas it is still possible to estimate the latent partition of the type I nodes.

Contributions. We extend the work of [4] that was specialized to a symmetric BiSBM with two type I and type II node communities to a general BiSBM by proposing a new algorithm also based on the power method, but which avoids estimating model parameters. We derive an upper bound for the misclustering rate (i.e., the fraction of misclassified nodes) and show that this rate is minimax optimal (up to a constant factor) under the setting of [4]. The details of our work are in [1].

2 The Generalized Power Method (GPM)

Let $A \in \{0,1\}^{n_1 \times n_2}$ denote the adjacency matrix of the observed bipartite graph with K (resp. L) row (resp. column) - clusters. Our aim is to cluster the rows of A. Let us introduce $B := \mathcal{H}(AA^\top)$ where $\mathcal{H}(X) = X - \mathrm{diag}(X)$. Given an initial rough estimate $Z_1^{(0)} \in \{0,1\}^{n_1 \times K}$ of the row-clusters, we can iteratively refine the partition by repeating the following steps for $0 \le t \le T-1$

- Form $W^{(t)} = (Z_1^{(t)})^\top (D^{(t)})^{-1}$ where $D^{(t)} = \mathrm{diag}((Z_1^{(t)})^\top \mathbf{1}_{n_1})$, and $\mathbf{1}_{n_1}$ is the all ones vector.
- Update $Z_1^{(t+1)} := \mathcal{P}(BW^{(t)})$ where $\mathcal{P}$ projects on to the extremal points of the unit simplex of $\mathbb{R}^K$.

3 Main results

One can show the following consistency guarantee for GPM.

Theorem 1 (Informal)**.** *Assume that $A \sim BiSBM$ with a latent row partition $Z_1 \in \{0,1\}^{n_1 \times K}$, and a edge-sparsity level p_{max} such that $n_1 n_2 p_{max}^2 \gtrsim \log n_1$. Then, if GPM is initialized with a partition estimate $Z_1^{(0)}$ that recovers a large enough portion of the clusters, the misclutering rate r of the output $Z_1^{(T)}$ of GPM satisfies w.h.p. for $T \gtrsim \log n_1$*

$$r(Z_1^{(T)}, Z_1) \le \exp(-\Omega(n_1 n_2 p_{max}^2)).$$

To obtain an initial estimate $Z_1^{(0)}$ that satisfies the theorem requirements, one can use a spectral method applied on B as shown in [1].

Next, we show that the rate of convergence of Theorem 1 is optimal when $K = L = 2$.

Theorem 2 (Informal)**.** *Suppose that $A \sim BiSBM$ with $K = L = 2$, $n_2 \gg n_1 \log n_1$, $n_1 n_2 p_{max}^2 \to \infty$ and $n_1 n_2 p_{max}^2 = O(\log n_1)$. Then there exists a constant $c_1 > 0$ such that*

$$\inf_{\hat{z}} \sup_{\theta \in \Theta} \mathbb{E}(r(\hat{z}, z)) \ge \exp(-c_1 n_1 n_2 p_{max}^2)$$

where the infimum is taken over all measurable functions $\hat{z}$ of A, and Θ is a set of admissible parameters corresponding to approximately equal sized clusters.

The proof of Theorem 1 relies on the framework developed by [2]. The proof of Theorem 2 is based on a Bayesian argument from [3] to reduce the problem to a two hypothesis testing problem. The error associated with the resulting complex two hypothesis testing problem is then lower bounded by the error associated with a simpler test.

References

[1] G. Braun and H. Tyagi. Minimax optimal clustering of bipartite graphs with a generalized power method. *arXiv:2205.12104*, 2022.

[2] C. Gao and A.Y. Zhang. Iterative algorithm for discrete structure recovery. *Ann. Stat.*, 50(2):1066 – 1094, 2022.

[3] M. Ndaoud. Sharp optimal recovery in the two component gaussian mixture model. *arXiv:1812.08078*, 2018.

[4] M. Ndaoud, S. Sigalla, and A.B. Tsybakov. Improved clustering algorithms for the bipartite stochastic block model. *IEEE Transactions on Information Theory*, 68(3):1960–1975, 2022.

*Emails: guillaume.braun@inria.fr, hemant.tyagi@inria.fr