

# Random matrices, perturbations and their applications in statistics

Sebestyén Kovács

Supervisor: Ágnes Backhausz

Eötvös Loránd University

Thursday 9<sup>th</sup> January, 2025

- 1 Introduction and motivations.
- 2 Simulations for community detection, Stochastic Block Model.
- 3 The problem of  $\mathbb{Z}_2$  synchronization.
- 4 Application on real data.

- Our main task was to filter out data observed with noise.
- The problem of Stochastic Block Model (SBM): reconstructing homogeneous, dense communities in the vertex set of a random graph.
- Differences between SBM and  $\mathbb{Z}_2$  synchronization.
- Finally, I applied the Stochastic Block Model to a larger deterministic graph (obtained from real data) as well.

# Stochastic Block model

- Let us assume that  $x \in \{1, -1\}^n$ , where the  $i$ -th coordinate of  $x$  is 1 if the  $i$ -th vertex belongs to the first group ( $I$ ) and  $-1$  if to the second group ( $J = V/I$ ; for  $i = 1 \dots n = |V|$ ).
- The distribution of the entries of adjacency matrix of our random graphs looks like this: ( $0 < q < p < 1$ , we draw the edges independently from one another)

$$P(A_{ij} = 1) = \begin{cases} p = p_{in}, & \text{if } i \in I \text{ and } j \in I, \text{ or } i \in J \text{ and } j \in J, \\ q = p_{out}, & \text{otherwise} \end{cases}$$

- The algorithm written in the article by Abbe, Fan, Wan and Zhong ([1]) wants to estimate  $x$ , we need to calculate the second eigenvector of the random adjacency matrix  $A$ :
- Compute  $u_2$ , the eigenvector of  $A$  corresponding to its second largest eigenvalue  $\lambda_2$ .
- Set  $\hat{x}^i := \text{sgn}(u_2^i)$ .

# Some metrics for the quality of classification.

- The natural definition of the misclassification rate when estimating  $x$  with  $\hat{x}$ :

$$r(x, \hat{x}) := \frac{1}{n} \sum_{i=1}^n 1_{\{x_i \neq \hat{x}_i\}}.$$

- The  $L^2$  distance between the separating and estimated vector can be estimated by averaging such observations:

$$\|x - \hat{x}\|_2 = \sqrt{\sum_{i=1}^n (x_i - \hat{x}_i)^2}.$$

- $V = \{1, 2, 3, \dots, 600\}$  (vertex set),  $I = \{1, 2, 3, \dots, 300\}$  (first group),  $J = \{301, 302, 303, \dots, 600\}$  (second group).
- We generated 8 different graphs independently from one another for every  $p_{in}$  and  $p_{out}$  pair.
- We calculated the average of the metrics for every  $p_{in}$  and  $p_{out}$  pair.

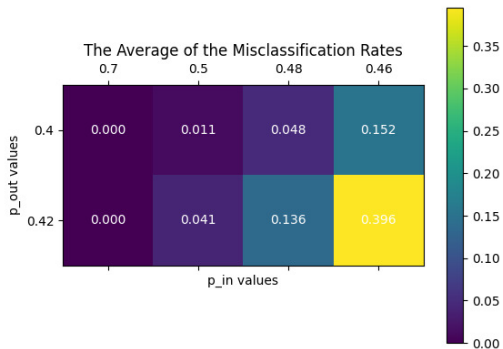


Figure: The average of the missclassification rates

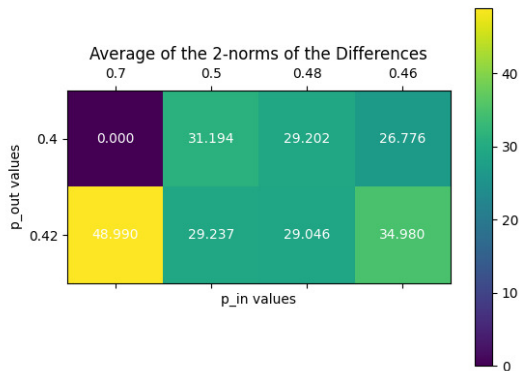


Figure: The average of the 2-norms of the differences



# First theorem (SBM)-Abbe, Fan, Wang, Zhong:

- Let us assume that the edge probabilities satisfy the following equations:

- 

$$p := a \cdot \frac{\ln(n)}{n} \quad \text{and} \quad q := b \cdot \frac{\ln(n)}{n}.$$

- If  $\sqrt{a} - \sqrt{b} > \sqrt{2}$  then there exist an  $\eta(a, b) > 0$  and  $s \in \{1, -1\}$  such that with probability  $1 - o(1)$

$$\lim_{n \rightarrow \infty} \left( \sqrt{n} \cdot \min_{i \in [n]} (s \cdot x_i \cdot u_2^i) \right) \geq \eta(a, b).$$

- And if  $0 < \sqrt{a} - \sqrt{b} \leq \sqrt{2}$ , the misclassification rate will not be too high on average:

$$\mathbb{E} \left[ \min_{s \in \{\pm 1\}} \frac{1}{n} \sum_{i=1}^n \mathbf{1}_{\{x_i \neq s \hat{x}_i\}} \right] \leq n^{-(1+o(1)) \frac{(a-b)^2}{2}}.$$

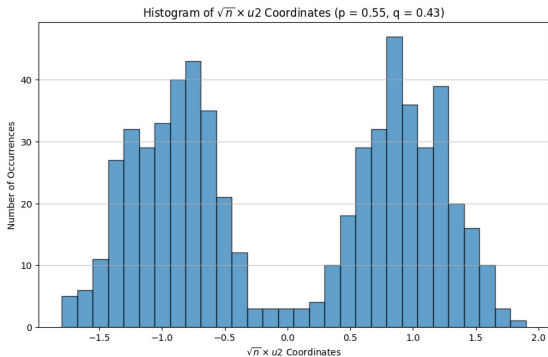


Figure: Histogram of  $\sqrt{n} \cdot u_2$  coordinates ( $p = 0.55, q = 0.43$ )

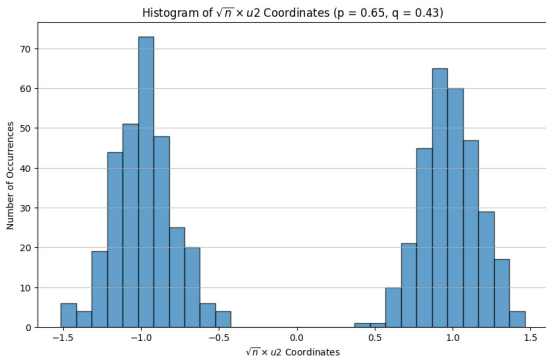


Figure: Histogram of  $\sqrt{n} \cdot u_2$  coordinates ( $p = 0.65, q = 0.43$ )

# The problem of $\mathbb{Z}_2$ synchronization

- In the  $\mathbb{Z}_2$  synchronization problem, we observe noisy versions of random  $\pm 1$  values and aim to filter out the noise, which is typically chosen from a normal distribution. We assume that we know the random matrix  $Y$ , generated as follows:

$$Y_{ij} = x_i \cdot x_j + \sigma \cdot W_{ij}, \quad \text{where } x \in \{\pm 1\}^n,$$

$$i < j \Rightarrow W_{ij} \sim N(0, 1), \quad \sigma > 0, \quad W_{ii} = 0, \quad W_{ij} = W_{ji}.$$

- Let us further assume that variables  $\{W_{ij} : i < j\}$  are independent from one another.
- Our aim is to recover  $x$  from  $Y$ .**
- The algorithm of Abbe, Fan, Wan, and Zhong estimates the values as follows:
- Compute the leading eigenvector of  $Y$ , denoted by  $u$ ;
- Take the estimate  $\hat{x}_i := \text{sgn}(u_i)$ .

## Second theorem ( $\mathbb{Z}_2$ synchronization)-Abbe, Fan, Wang, Zhong:

- *A*: We suppose that  $\sigma \leq \sqrt{\frac{n}{(2 + \varepsilon) \log n}}$  for some  $\varepsilon > 0$ .
- *E*: With probability  $1 - o(1)$ , the leading eigenvector  $u$  of  $Y$  with unit  $\ell_2$  norm satisfies

$$\sqrt{n} \cdot \min_{i \in [n]} \{s \cdot x_i \cdot u_i\} \geq 1 - \sqrt{\frac{2}{2 + \varepsilon}} + \frac{C}{\sqrt{\log n}},$$

- for a suitable  $s \in \{\pm 1\}$ , where  $C > 0$  is an absolute constant.

# Own simulation-combining the SBM and the $\mathbb{Z}_2$ synchronization problem

- We fixed two edge probabilities:  $p = 0.5$  and  $q = 0.4$ , and then independently generated ten random graphs, each with 600 vertices, using these probabilities.
- To the adjacency matrix of each graph, we added a scaled version of a  $600 \times 600$  symmetric matrix sampled from a standard normal multivariate distribution with the scaling factor  $\sigma$  varying according to the noise levels.
- Difference from  $\mathbb{Z}_2$  synchronization: we added noise not to  $x \cdot x^\top$  but to the adjacency matrices.

# Own simulation-combining the SBM and the $\mathbb{Z}_2$ synchronization problem

- Here, the error could arise from two sources: first, the edges themselves are random; second, we could only observe the adjacency matrix in a noisy environment.
- I applied the stochastic block model algorithm to the noisy adjacency matrices and evaluated the average accuracy of the reconstruction with the added noise.
- For my case,  $\sigma = 0.4$  was the highest noise level where the misclassification rate did not exceed 5%.

# Own simulation-combining the SBM and the $\mathbb{Z}_2$ synchronization problem

	Sigma	M.R.
0	0.0	0.008
1	0.2	0.017
2	0.4	0.045
3	0.6	0.102
4	0.8	0.155
5	1.0	0.275

**Figure:** The average misclassification rates with different noises,  $p_{in} = 0.5$ ,  $p_{out} = 0.4$



# Own simulation-combining the SBM and the $\mathbb{Z}_2$ synchronization problem

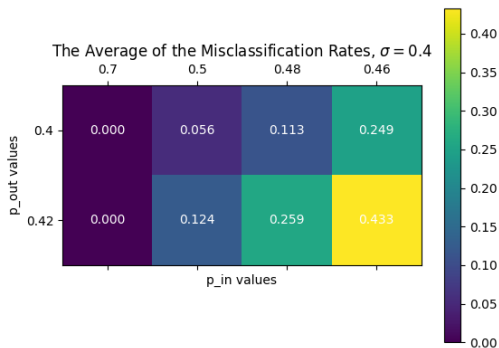


Figure: The average misclassification rates with noise  $\sigma = 0.4$

# Own simulation-combining the SBM and the $\mathbb{Z}_2$ synchronization problem

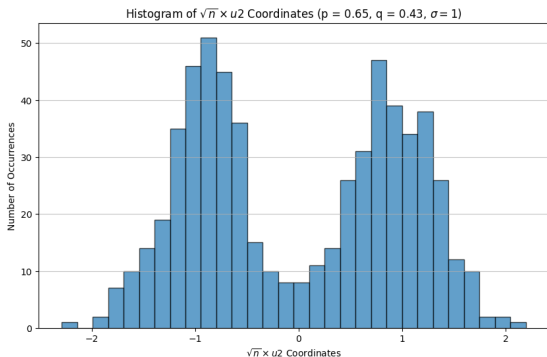


Figure: The histogram of the coordinates of  $\sqrt{n} \cdot u_2$  with  $p_{in} = 0.65$ ,  $p_{out} = 0.43$ ,  $\sigma = 1$

# Own simulation-combining the SBM and the $\mathbb{Z}_2$ synchronization problem

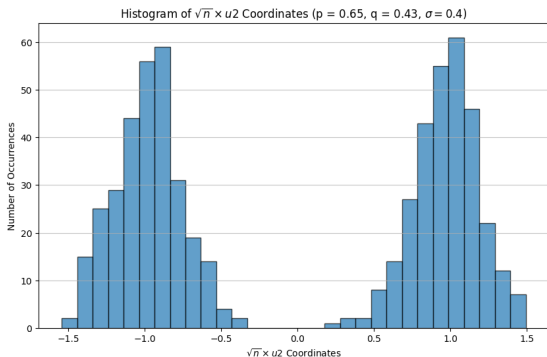


Figure: The histogram of the coordinates of  $\sqrt{n} \cdot u_2$  with  $p_{in} = 0.65$ ,  $p_{out} = 0.43$ ,  $\sigma = 0.4$

# Application on real data

- I tested the stochastic block model algorithm on a deterministic graph as well, where the edge between two vertices does not depend on randomness.
- The description of our graph is as follows: "This network was constructed from the USA's FAA (Federal Aviation Administration) National Flight Data Center (NFDC), Preferred Routes Database. Nodes in this network represent airports or service centers and links are created from strings of preferred routes recommended by the NFDC." ([7])
- The graph contained 1,226 vertices and 2,615 edges.

- An interesting question is whether the individual cohesions within the graph are stronger within the groups generated by the Stochastic Block Model. Such cohesion-related indicators include edge density and the clustering coefficient:
- The density of a graph with  $n$  vertices and  $m$  edges is  $\frac{m}{\binom{n}{2}}$ .  
This indicates how dense the edges are in the graph relative to the complete graph.
- The overall clustering coefficient of the graph is the average of the clustering coefficients of all vertices:

$$C = \frac{1}{n} \cdot \sum_{v \in V} \frac{|\{\{u, w\} \in E : u, w \in N(v)\}|}{\binom{\deg(v)}{2}}.$$

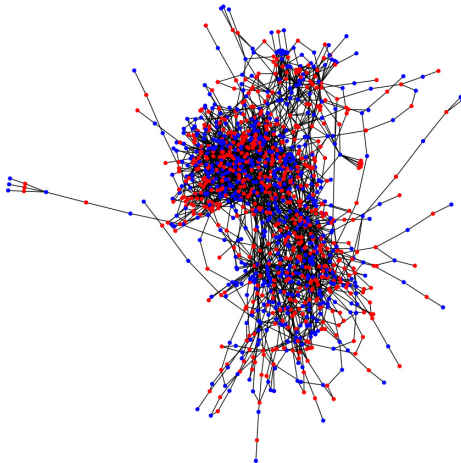







Figure: The two groups of our graph


# Edge density and clustering coefficient in the entire graph and in its groups

- Number of nodes in Group 1: 616
- Number of nodes in Group 2: 610
- Edge density in the entire graph: 0.0032
- Clustering coefficient in the entire graph: 0.0675
- Edge density in Group 1: 0.0054
- Edge density in Group 2: 0.0063
- Clustering coefficient in Group 1: 0.0900
- Clustering coefficient in Group 2: 0.1031

-  Emmanuel Abbe, Jianqing Fan, Kaizheng Wang, Yiqiao Zhong: *Entrywise eigenvector analysis of random matrices with low expected rank*, Ann Stat. 2020 June ; 48(3): 1452–1474. doi:10.1214/19-aos1854.(<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8046180/>)
-  Benaych-Georges, F.; Nadakuditi, R. R.: *The singular values and vectors of low rank perturbations of large rectangular random matrices*, J. Multivariate Anal. **111** (2012), 120–135.
-  Blum, A.; Hopcroft, J.; Kannan, R.: *Foundations of Data Science*, Cambridge University Press, 2020.
-  Chen, Y.; Cheng, Ch.; Fan, Y.: *Asymmetry helps: Eigenvalue and eigenvector analyses of asymmetrically perturbed low-rank matrices*, Annals of statistics, 2021, 49.1: 435.
-  Fan, Jianqing; Wang, Weichen; Zhong, Yiqiao: *An  $l^\infty$  eigenvector perturbation bound and its application to robust*



*covariance*, J. Mach. Learn. Res. 18 (2017), Paper No. 207, 42 pp.

 O'Rourke, S.; Vu, V.; Wang, K.: *Random perturbation of low rank matrices: Improving classical bounds*, Linear Algebra and its Applications **540** (2016), 26–59.

 Air traffic control  
(<http://konect.cc/networks/maayan-faa/>)

# Thank you for your attention!