SubSearch: Robust Estimation and Outlier Detection for Stochastic Block Models via Subgraph Search

Leonardo Martins Bianco

leonardo.martins-bianco@universite-paris-saclay.fr

Laboratoire de Mathématiques d'Orsay (IMO)

November 07, 2024

Graphs

Graph: nodes linked by edges.

$$G = (V, E)$$

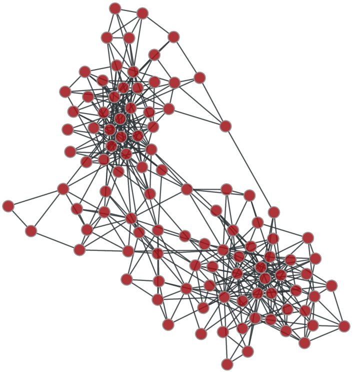
$$V = \{1, \ldots, n\}, E \subset V \times V$$

Undirected graph:

$$(i, j) \in E \Rightarrow (j, i) \in E$$

Adjacency matrix:

$$A_{ij} = \begin{cases} 1 & \text{if } i \sim j \\ 0 & \text{otherwise} \end{cases}$$



The Stochastic Block Model

Models graphs *w/ communities*.

n nodes

K communities

Size parameters π

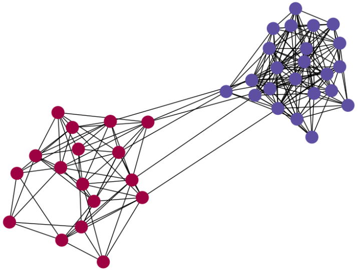
Z_i = label of node i

$\mathbb{P}(Z_i = k) = \pi_k$

Communities $\{\Omega_k\}_{k=1,\ldots,K}$

Connectivity parameters Γ

$\mathbb{P}(A_{ij} = 1 | Z_i = k, Z_j = l) = \Gamma_{kl}$



Motivation for Robustness

Classical Estimation

Studies $\hat{Z}(A), \hat{\Gamma}(A)$ with $(Z, A) \sim \text{SBM}_{n,K}(\pi, \Gamma)$.

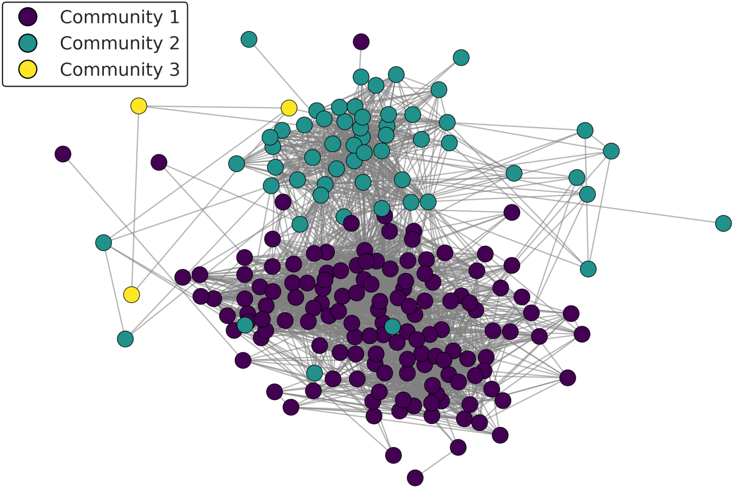
What if the data does not come *exactly* from the model?

Robust Estimation

Studies $\hat{Z}(\tilde{A}), \hat{\Gamma}(\tilde{A})$ with

$$(Z, A) \sim \text{SBM}_{n,k}(\pi, \Gamma) \quad \xrightarrow[\gamma n \text{ nodes}]{\text{Corruption}} \quad \tilde{A}$$

Motivation for Robustness



Previous Work for Robust SBM

On estimating Z:

- **Spectral methods:** spectrum of $\tilde{A}$ or related matrices. Faster, "less robust" [Stephan and Massoulié, 2019].
- **SDP methods:** "more robust", computationally expensive [Cai and Li, 2015, Ding et al., 2023].

Previous Work for Robust SBM

On estimating Z:

- **Spectral methods:** spectrum of $\tilde{A}$ or related matrices. Faster, "less robust" [Stephan and Massoulié, 2019].
- **SDP methods:** "more robust", computationally expensive [Cai and Li, 2015, Ding et al., 2023].

On estimating Γ:

- **Filtering:** for Erdős-Rényi ($K = 1$) graphs [Acharya et al., 2022].

Question:

How to robustly estimate Γ for $K > 1$?

Filtering for Erdős-Rényi Model ($K = 1$)

Idea: *filter out* "bad" outliers $\Rightarrow$ find a "good" subgraph S.
[Diakonikolas et al., 2019]

- 1st) *Error bound*: for any $S \subset V$,

$$|p - \hat{p}_S| \lesssim \frac{\|A_S - \hat{p}_S\|_{\text{op}}}{n}$$

  where A_S restriction of A to S and $\hat{p}_S = \left(\sum_{i,j \in S} A_{ij}\right) / |S|^2$.

- 2nd) *Optimizing the bound*: let $\lambda_{\max}$ be the top eigenvalue of $A_S - \hat{p}_S$ and v eigenvector of $\lambda_{\max}$. Removing $i \sim (v_j^2) \Rightarrow$ w.h.p. $\|A_S - \hat{p}_S\| \approx \sqrt{n}$.

Filtering for the SBM ($K > 1$)

Additional complexity: communities $\Rightarrow S = S_1 \cup \ldots \cup S_K$.

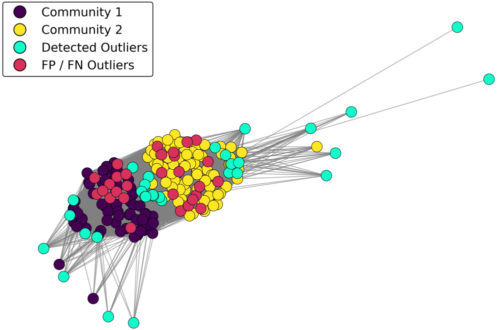
- 1st) *Error bound:* denote $\mathcal{I}$ the set of inliers. Then,

$$\|\Gamma - \hat{\Gamma}\|_1 \lesssim \frac{\|A_S - \hat{Q}(S)\|_{\text{op}}}{\min_{1\leq k\leq K}|\Omega_k \cap S_k \cap \mathcal{I}|} \tag{1}$$

  where $\hat{\Gamma} = \left(\sum_{i\in S_k j\in S_l} A_{ij}\right)/|S_k||S_l|$ and $\hat{Q}(S)_{ij} = \hat{\Gamma}_{S(i)S(j)}$.

- 2nd) *Optimizing the bound:*
  - Now v places weight on outliers *or misclassified* nodes.
  - Removing outliers is still good.
  - Removing a misclassified node $\longrightarrow$ unclear.

Failure of Filtering for $K > 1$

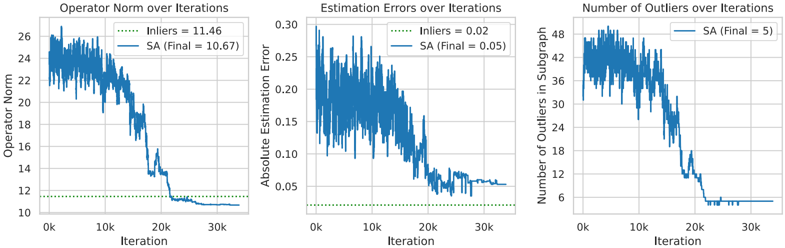


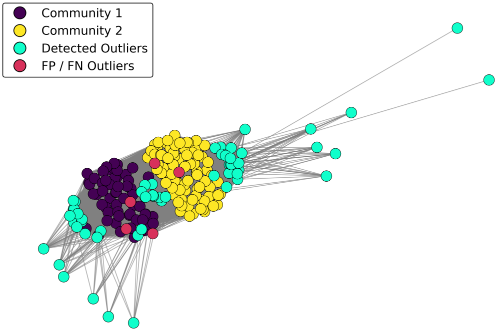
SubSearch: Subgraph Search via S.A.

We propose exploring the space $\mathcal{S}$ of all subgraphs $S \subset G$ of size $(1-\gamma)n$, in search of a minimizer of $c(S) = \|A_S - \hat{Q}(S)\|$.

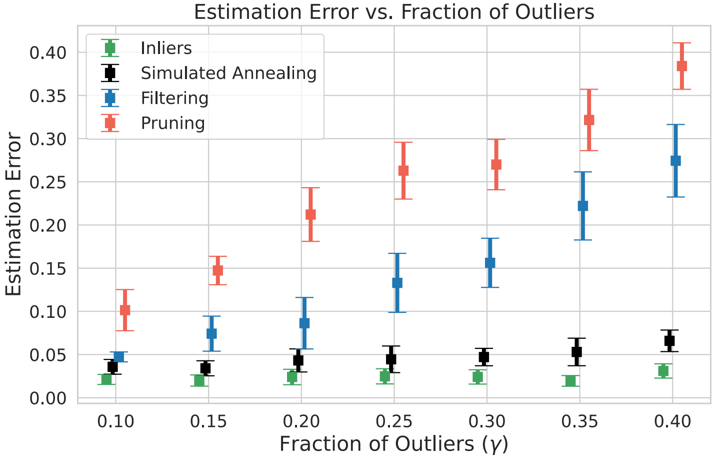
- Start with a high enough "temperature" T_0.
- $S_{\text{candidate}}$ is proposed by swapping random nodes $i \in S_{\text{current}}$ and $j \notin S_{\text{current}}$.
- $S_{\text{candidate}}$ is accepted with probability $\min(1, \exp(\Delta/T_t))$, where $\Delta := c(S_{\text{current}}) - c(S_{\text{candidate}})$.
- Temperature is decreased as $T_{t+1} = cT_t$, where the cooling rate $c \approx 1$.

Numerical Results

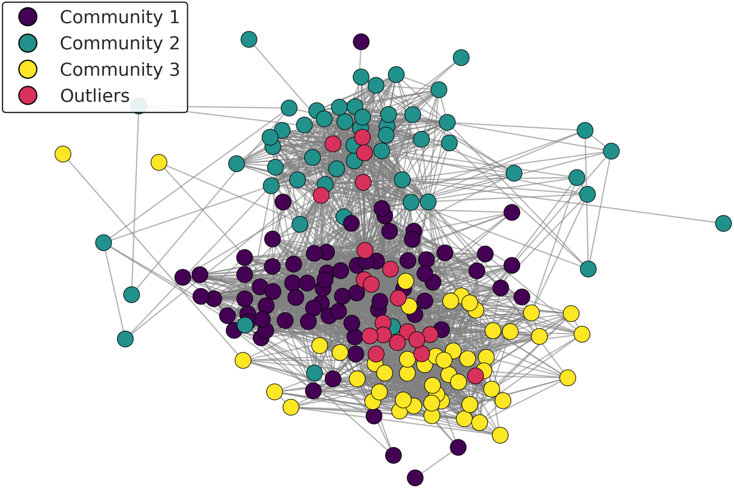




Numerical Results



Numerical Results



Take Away & Perspectives

Take Away:

- Error bound of Equation (1) $\Rightarrow$ objective function for robust estimation.
- Filtering is "greedy" and tries to remove worst node at each step $\Rightarrow$ bad solutions!
- We propose an algorithm with exploration to search for a "good" solution.

Perspectives:

- Actual robustness guarantees?
- Faster (non-asymptotic) convergence guarantees?

References I

Acharya, J., Jain, A., Kamath, G., Suresh, A. T., and Zhang, H. (2022).
Robust estimation for random graphs.
In Loh, P.-L. and Raginsky, M., editors, *Proceedings of Thirty Fifth Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, pages 130–166. PMLR.

Cai, T. T. and Li, X. (2015).
Robust and computationally feasible community detection in the presence of arbitrary outlier nodes.
The Annals of Statistics, 43(3).

Diakonikolas, I., Kamath, G., Kane, D., Li, J., Moitra, A., and Stewart, A. (2019).
Robust estimators in high dimensions without the computational intractability.

Ding, J., d'Orsi, T., Hua, Y., and Steurer, D. (2023).
Reaching Kesten-Stigum Threshold in the Stochastic Block Model under Node Corruptions.
arXiv:2305.10227 [cs, stat].

Gleiser, P. M. and Danon, L. (2003).
Community structure in jazz.
Advances in complex systems, 6(04):565–573.

References II

Liu, A. and Moitra, A. (2022).
Minimax rates for robust community detection.

Stephan, L. and Massoulié, L. (2019).
Robustness of spectral methods for community detection.
In Beygelzimer, A. and Hsu, D., editors, *Proceedings of the Thirty-Second Conference on Learning Theory*, volume 99 of *Proceedings of Machine Learning Research*, pages 2831–2860. PMLR.