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

Leonardo Martins Bianco[1], Christine Keribin[1], Zacharie Naulet[2]

[1]Université Paris-Saclay, CNRS, Inria, Laboratoire de Mathématiques d'Orsay, [2]Université Paris-Saclay, INRAE

Motivation

Community detection is a fundamental task in graph analysis, with methods often relying on fitting models like the Stochastic Block Model (SBM) [1] to observed networks. While many algorithms can accurately estimate SBM parameters and communities when the input graph is a perfect sample from the model, real-world graphs rarely conform to such idealized assumptions. Therefore, *robust* algorithms, capable of recovering model parameters even when the data deviates from the assumed distribution, are crucial. We propose **SubSearch**, an algorithm for robustly estimating SBM parameters by exploring the space of subgraphs in search of one that closely aligns with the model's assumptions. Our approach also works as an outlier detection method, properly identifying nodes responsible for the graph's deviation from the model and going beyond simple techniques like pruning high-degree nodes.

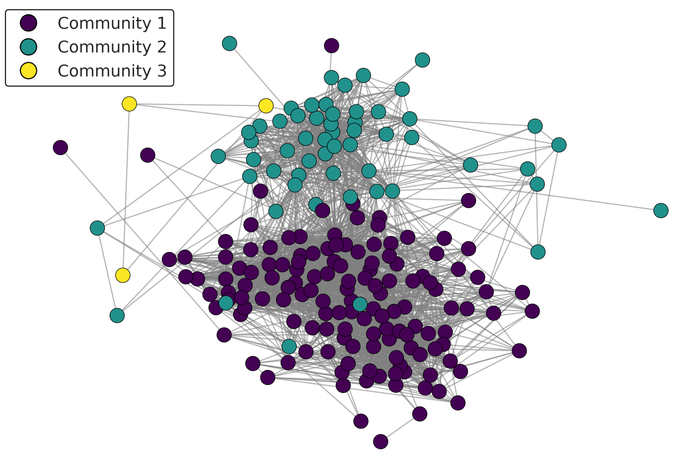

Fig. 1: Spectral clustering applied to the jazz collaboration dataset [2]. Nodes represent jazz musicians, with edges being collaborations during 1912 - 1940. This algorithm, which has guarantees under the SBM, fails to separate the graph into its three main collaboration groups.

Model

The *adjacency matrix* of a graph $G = (V, E)$ is defined as $A_{ij} = 1$ if $(i, j) \in E$, zero otherwise. A *graph with communities* has a partition of V into K communities $(\Omega_k)_{k=1,\dots,K}$ that can be represented by $z \in \{1, \dots, K\}^n$ or $Z \in \{0, 1\}^{n \times K}$ such that $\sum_j Z_{ij} = 1$ for every $i \in \{1, \dots, n\}$. The *Stochastic Block Model* (SBM) is a popular model for graphs with communities. Given size parameters $\Pi = (\pi_1, \dots, \pi_K)$ such that $\forall k, 0 < \pi_k < 1$ and $\sum_k \pi_k = 1$, and connectivity parameters $\Gamma \in [0, 1]^{K \times K}$, the SBM poses

$$\mathbb{P}(Z) = \prod_{k=1}^{K} \pi_k^{|\Omega_k|},$$
$$\mathbb{P}(A|Z) = \prod_{i \neq j} \Gamma_{z_i z_j}^{A_{ij}} (1 - \Gamma_{z_i z_j})^{1 - A_{ij}}.$$

If $S_1, \dots, S_K$ are disjoint subsets of $\{1, \dots, n\}$, we can estimate the connectivity parameters associated to them, for $k, l \in \{1, \dots, K\}$, by

$$\hat{\Gamma}_{kl} = \frac{1}{|S_k||S_l|} \sum_{i=1}^{|S_k|} \sum_{j=1}^{|S_l|} (A_{S_k \times S_l})_{ij},$$

defining a $K \times K$ matrix $\hat{\Gamma}$. This can be extended to an $n \times n$ matrix $\hat{Q}(S) \coloneqq \mathbf{S}\hat{\Gamma}\mathbf{S}^t$, where $\mathbf{S}$ is the $|S| \times K$ matrix such that $\mathbf{S}_{ij} = 1$ if $S(i) \in S_j$ and 0 otherwise.

Problem Statement

We consider the node adversary perturbation model, where an adversary receives a sample (Z, A_0) of an SBM and is allowed to *arbitrarily* modify the adjacency of up to γn nodes, where $\gamma \in [0, 1/2)$ is a known parameter representing the amount of corruption. This leads to the observation of a *corrupted* adjacency matrix A. The nodes whose adjacencies were directly modified by the adversary are called *outlier* nodes, while the rest are called *inlier* nodes. We denote the set of inlier nodes as F. The goal is to accurately estimate the connectivity parameters Γ of the original SBM from A, in the sense of minimizing the empirical estimation error $\sum_{kl} |\Gamma_{kl} - \hat{\Gamma}_{kl}|$.

Error Bound

We prove the following bound on the estimation error, extending the proof in [3] for $K = 1$.

Theorem. *Let A be an adjacency matrix sampled from a γ-corrupt* SBM *with K communities $(\Omega_k)_{k=1,\dots,K}$, connectivity parameters Γ, and inlier nodes F. Furthermore, let $S_1, \dots, S_K$ be non-empty disjoint subsets of $\{1, \dots, n\}$, S be their union, and $\hat{Q}(S)$ be the estimation of the expected adjacency matrix restricted to S. Then,*

$$\sum_{k=1}^{K} \sum_{l=k}^{K} |\Gamma_{kl} - \hat{\Gamma}_{kl}| \leq \frac{K^2}{\min_k |\Omega_k \cap S_k \cap F|} \times \left(\max_k \Gamma_{kk} + \|A_F - \mathbb{E}[A]_F\| + \|A_S - \hat{Q}(S)\| \right)$$

Minimizing the right-hand side of this bound gives us an idea for an algorithm.

Algorithm

We propose using Simulated Annealing (S.A.) to minimize the cost function $c(S) = \|A_S - \hat{Q}(S)\|$.

Algorithm 1 SubSearch

Require: $A, K, \gamma, c, (l_t)_{t=0,\dots,t_{\max}}, t_{\max}, t_{\text{tol}}, \varepsilon$.

$S_{\text{current}} \leftarrow$ connected subgraph with $|S| = (1 - \gamma)n$
$S_{\text{best}} \leftarrow S_{\text{current}}$
$T_0 \leftarrow \texttt{set\_initial\_temp}(S_{\text{current}})$
for $t = 1, \dots, t_{\max}$ **do**
  for $l = 1, \dots, l_t$ **do**
    $S_{\text{candidate}} \leftarrow \texttt{neighbor}(S_{\text{current}})$
    $\Delta \leftarrow c(S_{\text{current}}) - c(S_{\text{candidate}})$
    $u \sim \mathcal{U}([0, 1])$
    $\texttt{accept\_prob} \leftarrow \min(1, \exp(\Delta / T_t))$
    if $u < \texttt{accept\_prob}$ **then**
      $S_{\text{current}} \leftarrow S_{\text{candidate}}$
      if $c(S_{\text{current}}) < c(S_{\text{best}})$ **then**
        $S_{\text{best}} \leftarrow S_{\text{current}}$
  $T_{t+1} \leftarrow cT_t$
  if $\texttt{stopping\_conditions}(\varepsilon, t_{\text{tol}})$ **then**
    break
return S_{best}

Intuitively, this explores the space of subgraphs in search of one well explained by the model, and this implies removing "bad" outliers.

Experiments

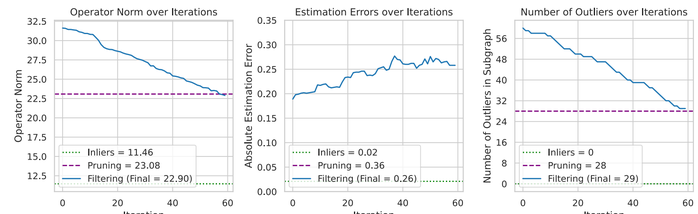

Fig. 2: Filtering decreases the cost (operator norm) and the number of outliers by considering smaller subgraphs at each step, but fails to decrease the error due to the lack of exploration.

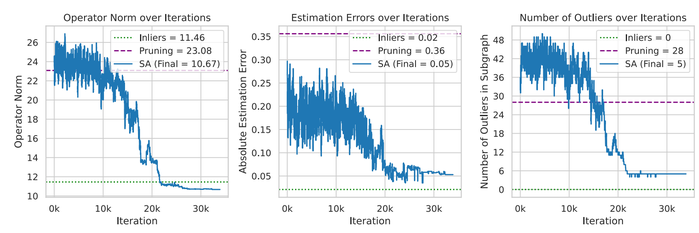

Fig. 3: Our method (S.A.) decreases the cost (operator norm) and the number of outliers by exploring subgraph space and finding good solutions, while keeping subgraph size constant.

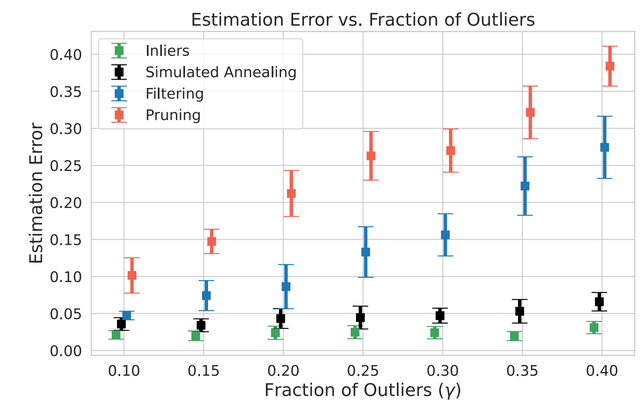

Fig. 4: Estimation error of different methods as the amount of perturbation increases. Our method stays close to the inlier baseline.

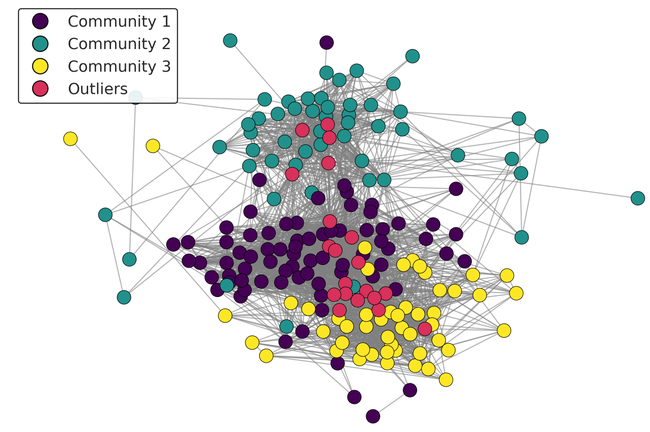

Fig. 5: Community partition of the jazz collaboration graph using our method (S.A.). It identifies outlier nodes to be ignored, allowing Spectral Clustering to find the three main collaboration groups (corresponding to bands in New York, Chicago, and other cities), something it had not been capable in Figure 1.

Conclusion

We argue that *exploring* the subgraph-space in search of one that can be well-explained by the model allows for robust estimation of the model's parameters. Future research directions include finding stronger theoretical robustness guarantees and considering the use of other optimization algorithms over S.A..

References

[1] Paul W. Holland et al. "Stochastic blockmodels: First steps". In: *Social Networks* 5.2 (1983), pp. 109–137. ISSN: 0378-8733. DOI: https://doi.org/10.1016/0378-8733(83)90021-7.
[2] Pablo M Gleiser et al. "Community structure in jazz". In: *Advances in complex systems* 6.04 (2003), pp. 565–573.
[3] Jayadev Acharya et al. "Robust Estimation for Random Graphs". In: *Proceedings of Thirty Fifth Conference on Learning Theory*. Vol. 178. Proceedings of Machine Learning Research. PMLR, Feb. 2022, pp. 130–166.

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