The Propagation Puzzle: Unraveling Community Clusters in the Stochastic Block Model Network

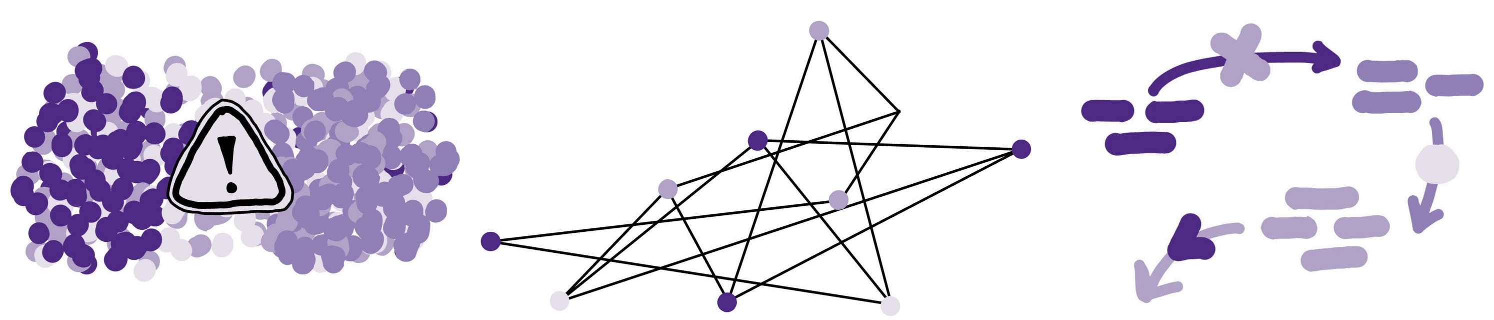
Sophia Pi, Northwestern University
Advised by Prof. Miklos Racz, Mentored by Shuwen Chai

Northwestern University

Northwestern OFFICE OF UNDERGRADUATE RESEARCH

Searching for Clusters

Clustering (also known as **community detection**) is a classic problem in machine learning and network science that seeks to **identify clusters of related nodes in a network**



Community detection has critical applications in epidemiology (disease hotspots), finance (fraud detection), sociology (influence groups), and more.

LPA Algorithm

The **Label Propagation Algorithm (LPA)** is a popular clustering algorithm for many real-world applications due to its speed and scalability.

Despite its popularity, there is surprisingly little literature published on the behavior of LPA on the **stochastic block model (SBM)**

Goal: Provide empirical and theoretical characterizations of the LPA on the 2-community SBM

Graph Model

- Networks modeled as **unweighted, undirected graphs**
- Erdős-Rényi: one community (N, p)
- SBM: 2+ communities (N, p, q)

Algorithm

- The LPA is a way of identifying communities by propagating "opinions" through a network:
  1. Randomly assign labels 0 through N-1 to each node
  2. For each node, relabel it with the majority label of its neighbors, breaking ties towards the smaller label
  3. Repeat step 2 until **convergence** (no more changing labels) or termination

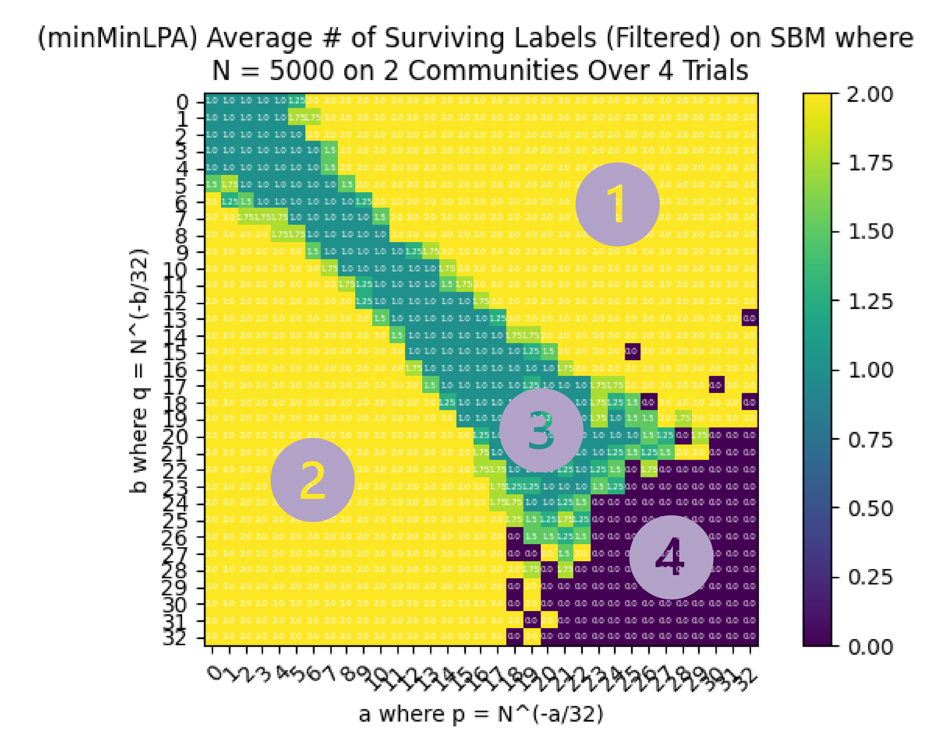
Empirical Simulations

Parameters

- **N** - # of nodes in network
- **numComm** - # of communities
- **p** – probability that any two nodes in the same community are connected
- **q** – probability that any two nodes in different communities are connected
- **cap** - # of iterations allowed before termination

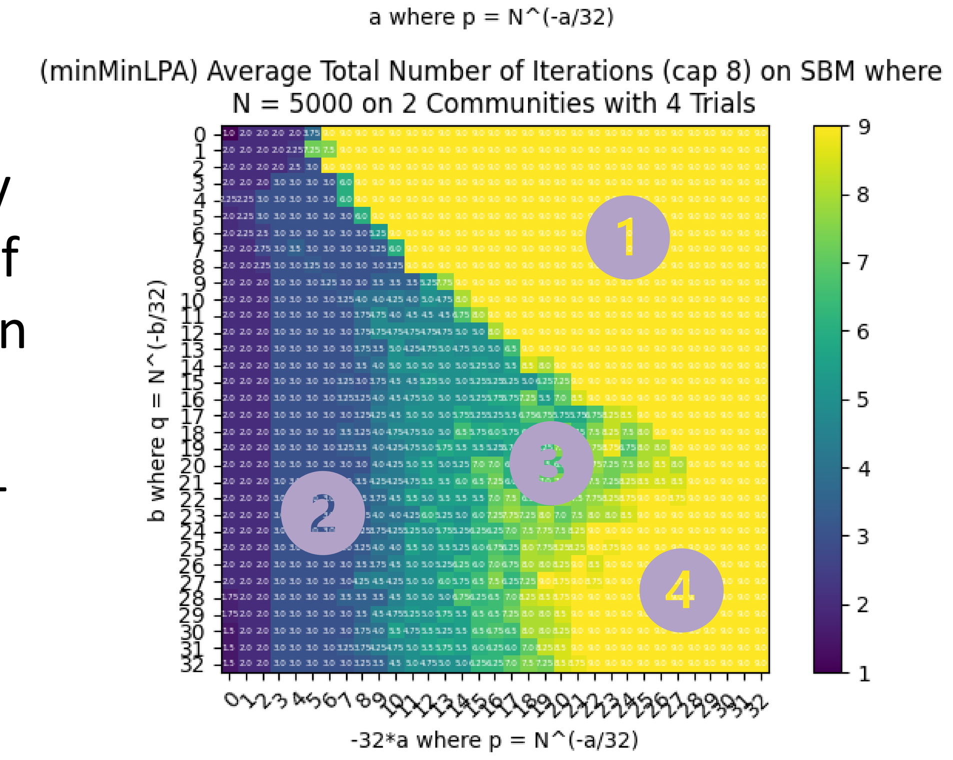
Behavioral Partition

Identified four distinct regions of the p-q parameter space that partition the space by convergence behavior

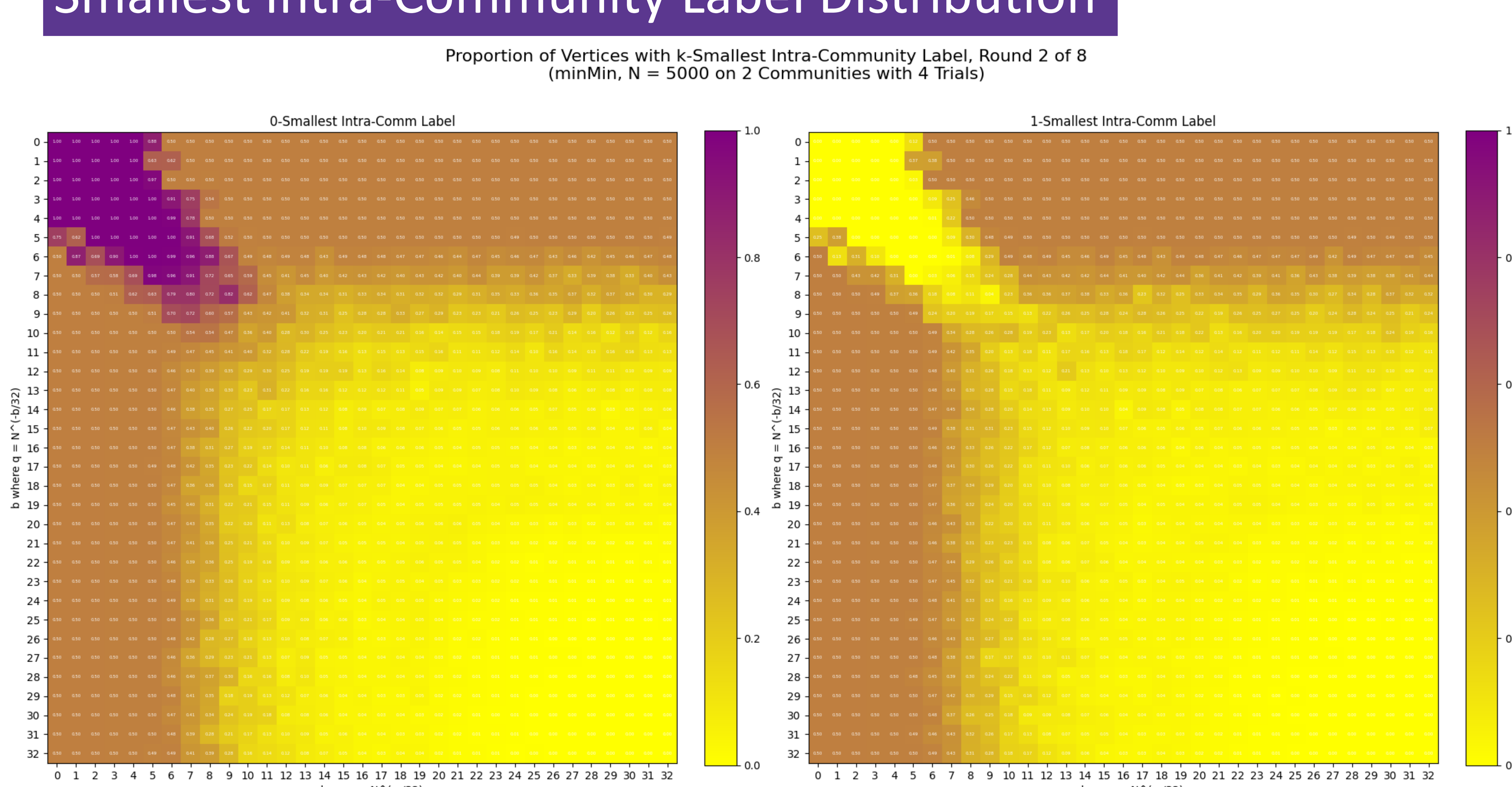


Switching

Despite the seemingly symmetric behavior of regions 1 and 2, region 2 exhibits stable convergence; region 1 exhibits periodic switching of labels



Smallest Intra-Community Label Distribution



Mathematical Proofs

Theorem 1 (Stable Convergence). *Let $G \sim SSBM(n,p,q)$ be a symmetric stochastic block model. All vertices are labeled as the smallest label of its community within 2 rounds of MinLPA w.h.p. if (i) $p \gg q$ and $p = \Omega(n^{-\frac{1}{4}+\epsilon})$ (an improvement from the current result [2], or (ii) $q \gg p$ and $q = \Omega(n^{-\frac{1}{4}+\epsilon})$.*

On the other side, not all vertices are labeled as the smallest label of its community within 2 rounds of MinLPA w.h.p. if $\max(p,q) = O(n^{-\frac{1}{4}})$.

Lemma 1 (Cross-Community Superiority). *Let $G \sim SSBM(n,p,q)$ with uniform random label initialization. Assume $p = \alpha n^{-a}$ and $q = \beta n^{-b}$. Assume that $v_{(1)} \in V_1$ and $v_{(2)} \in V_2$. Then $\forall v \in V_1$,*

$$Y_1(v) > Y_2(v)$$

where $Y_i(v)$ denotes the number of vertices in the neighborhood of v with label i after the first round. Also, $\forall v \in V_2$,

$$Y_2(v) > Y_1(v)$$

Lemma 2 (Local Superiority). *Let $G \sim SBM(n,p,q)$. Assume that $v_{(1)} \in V_1$ and $v_{(2)} \in V_2$. Then $\forall v \in V_1$,*

$$Y_1(v) > Y_i(v), \forall i \neq 1 \text{ with } v_{(i)} \in V_1$$

where $Y_i(v)$ denotes the number of vertices in the neighborhood of v with label i after the first round. Also, $\forall v \in V_2$,

$$Y_2(v) > Y_i(v), \forall i \neq 2 \text{ with } v_{(i)} \in V_2$$

Theoretical analysis proved tighter convergence conditions for accurate identification in two rounds; guarantee asymptotically accurate community detection for $SSBM(n,p,q)$ at the $\max\{p,q\} > n^{-1/4}$ threshold

Conclusions and Future Work

We have identified several critical behaviors (guaranteed convergence, guaranteed non-convergence, switching) of the LPA on the symmetric SBM in 2 communities, and have improved upon existing theoretical bounds on convergence criteria in two rounds. Future work may examine theoretical convergence in three or more rounds, precise behavior of the "Erdős-Rényi strip", LPA behavior over three or more communities, non-symmetric SBM, or different tiebreaking variation of the LPA.

References

[1] Kiwi, Marcos, Lyuben Lichev, Dieter Mitsche, and Paweł Prałat. "Label propagation on binomial random graphs." arXiv preprint arXiv:2302.03569 (2023).

[2] Kishore Kothapalli, Sriram V Pemmaraju, and Vivek Sardeshmukh. On the analysis of a label propagation algorithm for community detection. In International Conference on Distributed Computing and Networking, pages 255–269. Springer, 2013.