Community Detection Problem

Eric Sheng

Supervised by Charilaos Efthymiou

March 05, 2024

Overview

Community Detection Problem

- Community detection in graphs is the problem of finding groups of vertices which are more densely connected than they are to the rest of the graph.
- NP-Hard in general (Schaeffer 2007).

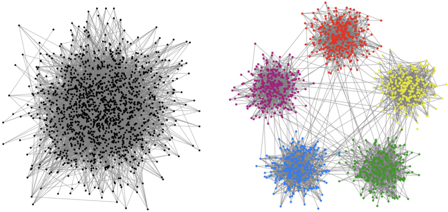


Figure: The right graph is the community structrue hidden in the left one

More Formally

- **Stochastic Block Model (SBM):** For a graph of n vertices and k groups. Each vertex i belongs to a group $\sigma_i \in \{1, \ldots, k\}$; where σ is the planted assignment. $\Pr(i,j) \in E$ depends only on whether i and j are in the same group or different group:

$$\Pr(i,j) \in E = \begin{cases} p_{\text{in}} & \text{if } \sigma_i = \sigma_j \\ p_{\text{out}} & \text{if } \sigma_i \neq \sigma_j \end{cases}$$

More Formally

- $c :=$ the expected degree, $p_{in} = \frac{c_{in}}{n}$, $p_{out} = \frac{c_{out}}{n}$ and $c = \frac{c_{in}+(k-1)c_{out}}{k}$
- $SBM(n, k, c_{in}/n, c_{out}/n) \rightarrow (G, \sigma)$, where G is a SBM with k communities, probability c_{in}/n inside the communities and c_{out}/n aross, σ is the planted assignment.

More Formally

- **Detection:** can we distinguish G generated by SBM from the Erdős-Rényi random graph $G(n, c/n)$ with the same average degree?

- **Recovery:** label the vertices with an assignment τ that is correlated with the planted assignment σ. Better than random guess.

Conjecture

Conjecture (Kesten-Stigum (KS) threshold)

Let (G, σ) be drawn from $\mathrm{SBM}(n, k, c_{in}/n, c_{out}/n)$. Define $\mathrm{SNR} = \frac{(c_{in}-c_{out})^2}{k(c_{in}+(k-1)c_{out})}$. Then,

- For any $k \geq 2$, if $\mathrm{SNR} > 1$ (the KS threshold), detection and weak recovery are possibly solvable in polynomial time.
- If $k \geq 4$, detection and weak recovery are possibly information-theoretically solvable (not necessarily in polynomial time) for $\mathrm{SNR} < 1$.

Some Progress So Far

- It was shown that weak recovery can be achieved efficiently for SNR > 1 and $k = 2$ (Bordenave et al. 2015).

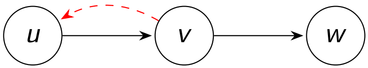
- It was proved that weak recovery is not solvable if SNR < 1 (Mossel et al. 2012).

Existing Solution (Spectral Algorithms)

The **non-backtracking Matrix** B can defined as:

$$B_{(u,v)(w,x)} = \begin{cases} 1 & \text{if } v = w \text{ and } u \neq x \\ 0 & \text{otherwise} \end{cases}$$

This matrix corresponds to a non-backtracking walk, which is a walk that does not repeat a vertex within 2 steps.



Non-backtracking Matrix

Claim: the eigenvector λ_μ associated with the second largest eigenvalue μ of B is correlated with the true communities whenever it is outside the bulk of spectrum of the B (Krzakala et al. 2013).

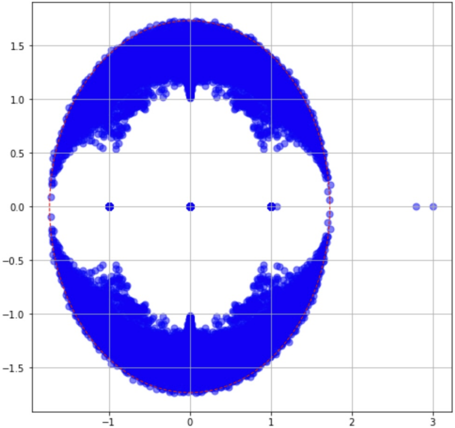


Figure: Spectrum Distribution

Properties:

1. leading eigenvalue = average degree $c = c_{in} + c_{out}/2$
2. second largest eigenvalue μ is approaching to $c_{in} - c_{out}/2$
3. the bulk of B's spectrum is confined to disk of radius $\sqrt{c}$

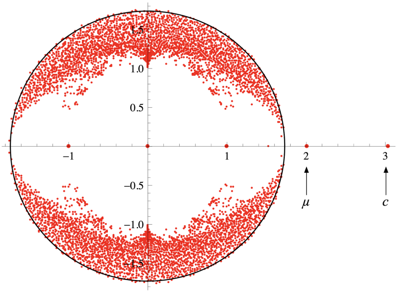


Figure: here, $c_{in} = 5, c_{out} = 1$, so $c = 3$ and $\mu = (c_{in} - c_{out})/2 = 2$

Spectral Algorithm based on Non-backtracking Operator: at each vertex we sum the eigenvector λ_μ of μ over all its incoming edges and label vertices according to the sign of this sum.

Claim: Non-backtracking-based spectral algorithm can succeed all the way down to the KS threshold no matter in sparse or dense graph (Abbe 2018).

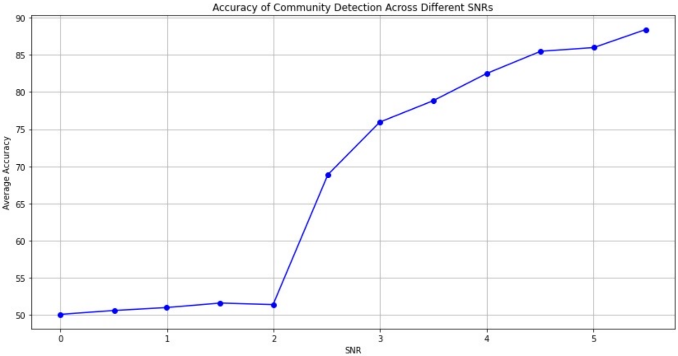


Figure: Accuracy of spectral algorithm

Community Detection with Partial Pre-defined Assignment

Problem Statement

Let (G, σ) be drawn from $\mathrm{SBM}(n, k, c_{in}/n, c_{out}/n)$. Given a subset $R \subset V$ where the community assignment σ_i for $i \in R$ is known, our task is to recover the complete assignment σ. Furthermore, $|R| = \alpha|V|$ for some $\alpha \in (0, 1)$

Greedy Recovery Algorithm

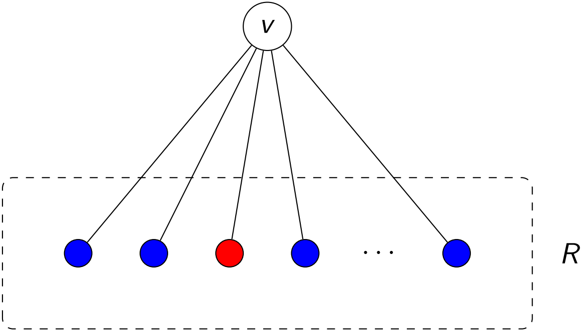
Algorithm 1 Greedy Recovery Algorithm

Input: Graph $G = (V, E)$; the set R of vertices with pre-defined assignments.

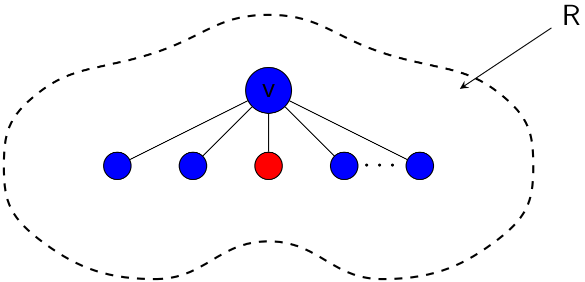
Output: R, a set of vertices for which the assignment has been determined.

procedure $\textsc{GreedyRecovery}(G, R)$
  if $|R| = |V|$ **then**
    return R
  end if
  for all $v \in V \setminus R$ **do**
    if $\exists u \in N(v) : u \in R$ **then**
      $R[v] \leftarrow$ the majority assignment of v's neighbors
    end if
  end for
  $\textsc{GreedyRecovery}(G, R)$ until no new assignment can be made
  return R
end procedure

Example



Example



Experimental Results

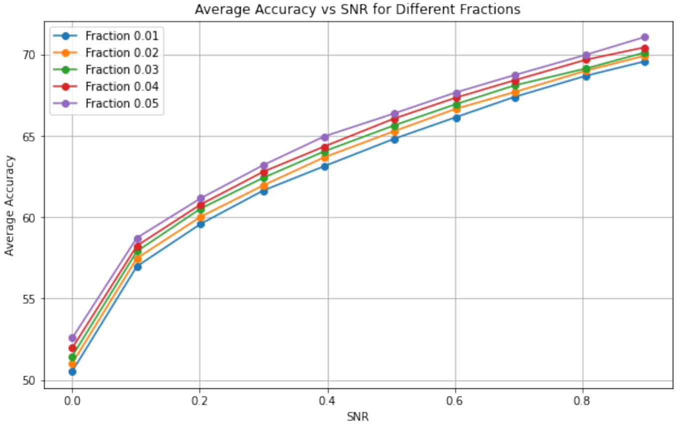


Figure: Average accuracy over 10 instance with SNR ranges from **0 to 0.9** and Fraction α ranges from 0.01 to 0.05

Experimental Results

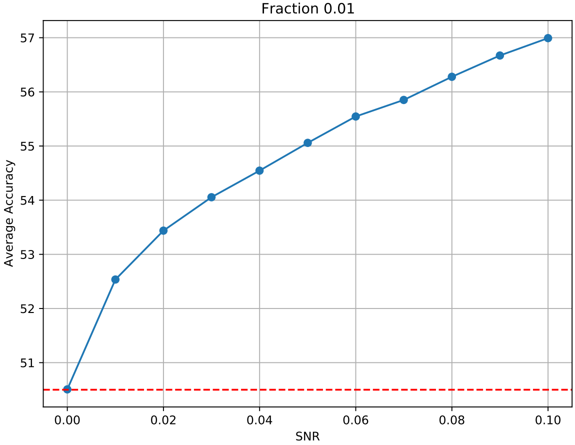


Figure: Average accuracy over 10 instance with SNR ranges from **0 to 0.1** and Fraction $\alpha = 0.01$

Experimental Results

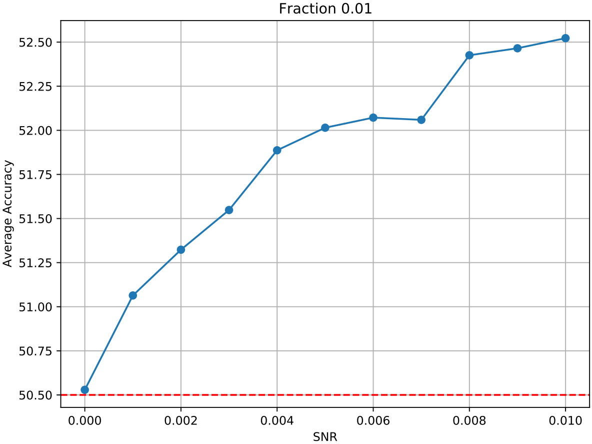


Figure: Average accuracy over 10 instance with SNR ranges from **0 to 0.01** and Fraction $\alpha = 0.01$

Evaluation

Pros

- simple and fast
- good accuracy, achieve the aim of this project

Cons

- don't have time to generalise it to case $k > 2$
- don't have time to delve deeply into and analyse the relationship between SNR, fraction value and accuracy.

Future Work

1. the inital R also contains some noise (Robustness).
2. combine spectral algorithm with our greedy recovery algorithm for $SNR > 1$
3. generalise it to more communities, particularly $K \geq 4$.
4. in-depth analysis of the relationship between SNR, fraction value and accuracy.

Robustness

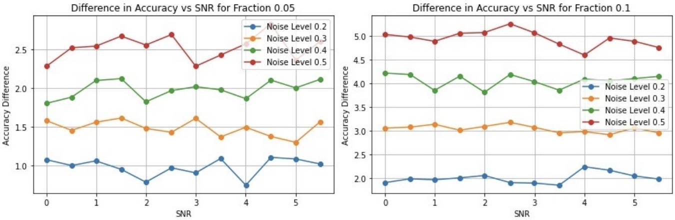


Figure: Accuracy difference when $|R|$ is small

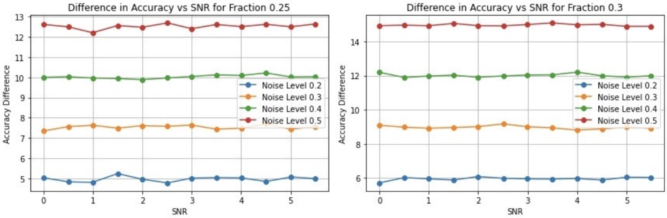


Figure: Accuracy difference when $|R|$ is large

Some Preliminary Results

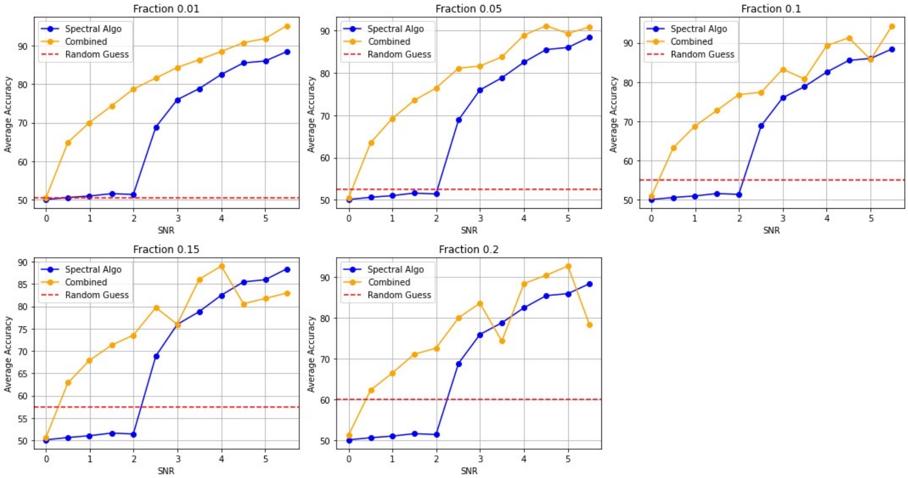
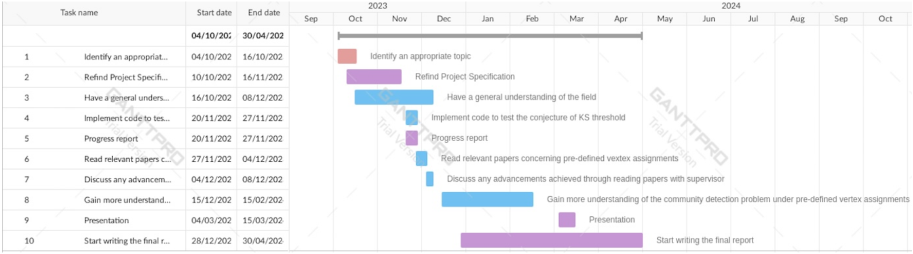


Figure: Accuracy Comparison: Spectral Algorithm vs. Combined Algorithm

Project Management



	Task name	Start date	End date
		04/10/202	30/04/202
1	Identify an appropriat...	04/10/202	16/10/202
2	Refind Project Specifi...	10/10/202	16/11/202
3	Have a general unders...	16/10/202	08/12/202
4	Implement code to tes...	20/11/202	27/11/202
5	Progress report	20/11/202	27/11/202
6	Read relevant papers c...	27/11/202	04/12/202
7	Discuss any advancem...	04/12/202	08/12/202
8	Gain more understand...	15/12/202	15/02/202
9	Presentation	04/03/202	15/03/202
10	Start writing the final r...	28/12/202	30/04/202

- Behind schedule in Term 1.
- All primary objectives are complete on time.
- Allows me to have a try for the further works.

Thank You!

References I

Abbe, Emmanuel (2018). "Community Detection and Stochastic Block Models: Recent Developments". In: *Journal of Machine Learning Research* 18.177, pp. 1–86. URL: http://jmlr.org/papers/v18/16-480.html.

Bordenave, Charles et al. (2015). *Non-backtracking spectrum of random graphs: community detection and non-regular Ramanujan graphs*. arXiv: 1501.06087 [math.PR].

Krzakala, Florent et al. (Nov. 2013). "Spectral redemption in clustering sparse networks". In: *Proceedings of the National Academy of Sciences* 110.52, pp. 20935–20940. ISSN: 1091-6490. DOI: 10.1073/pnas.1312486110. URL: http://dx.doi.org/10.1073/pnas.1312486110.

References II

Moore, Cristopher (2017). "The Computer Science and Physics of Community Detection: Landscapes, Phase Transitions, and Hardness". In: *Bull. EATCS* 121. URL: https://api.semanticscholar.org/CorpusID:1213533.

Mossel, Elchanan et al. (2012). "Reconstruction and estimation in the planted partition model". In: *Probability Theory and Related Fields* 162, pp. 431–461. URL: https://api.semanticscholar.org/CorpusID:120425378.

Nadakuditi et al. (May 2012). "Graph Spectra and the Detectability of Community Structure in Networks". In: *Physical Review Letters* 108.18. ISSN: 1079-7114. DOI: 10.1103/physrevlett.108.188701. URL: http://dx.doi.org/10.1103/PhysRevLett.108.188701.

Schaeffer, Satu Elisa (2007). "Graph clustering". In: *Computer Science Review* 1.1, pp. 27–64. ISSN: 1574-0137.