Consistency of Spectral Algorithms for Hypergraphs under Planted Partition Model

Debarghya Ghoshdastidar

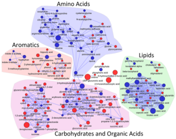
Ph.D. (4^{th} year)
Dept. of Computer Science & Automation
Advisor: Prof. Ambedkar Dukkipati

April, 2016

Networks and communities

Networks are ubiquitous:

Social network



Biological network

Image segmentation

Community structure is universal:

- Behavioral similarity of friends
- Functional similarity of proteins / molecules
- Pixels associated with same object

Community detection is crucial:

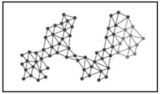
- Server load balancing & efficient data storage in networking sites
- Finding functional relationships in biological networks

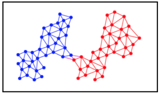
From networks to graphs

$$\text{Network} \Longrightarrow \text{Graph}$$
$$\text{Community detection} \Longrightarrow \text{Graph partitioning}$$

Graph partitioning:

- Group vertices into disjoint sets
- Each group has high edge density
- Few edges cross boundaries
- Groups of comparable sizes (balanced partition)





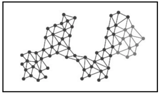
From networks to graphs

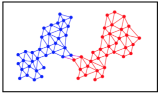
Network $\Longrightarrow$ Graph

Community detection $\Longrightarrow$ Graph partitioning

Graph partitioning:

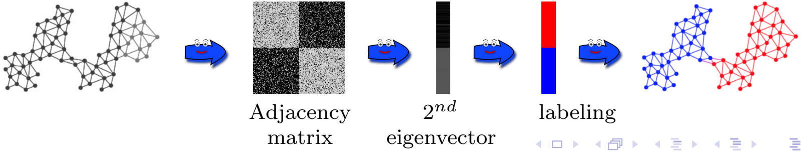
- Group vertices into disjoint sets
- Each group has high edge density
- Few edges cross boundaries
- Groups of comparable sizes (balanced partition)





Spectral algorithms:

[Fiedler '73; Shi & Malik '00; McSherry '01; Rohe et al '11; Vu '14]



Question: How 'good' are these algorithms?

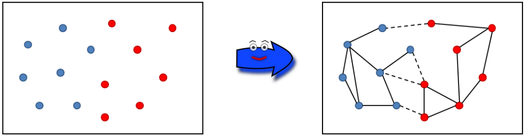
Planted partition model

[Holland et al '83; McSherry '01]

Question: How 'good' are these algorithms?

Framework:

- Given n blue nodes, and n red nodes
- Connect vertices of same color with probability p
- Connect vertices of different color with probability $q < p$



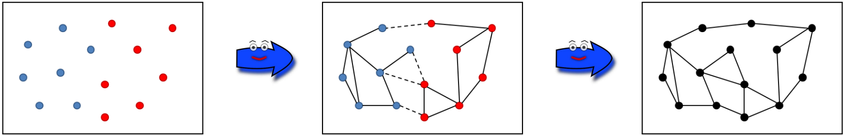
Planted partition model

[Holland et al '83; McSherry '01]

Question: How 'good' are these algorithms?

Framework:

- Given n blue nodes, and n red nodes
- Connect vertices of same color with probability p
- Connect vertices of different color with probability $q < p$



Remove labels

Question (formal): How many vertices are incorrectly labelled?

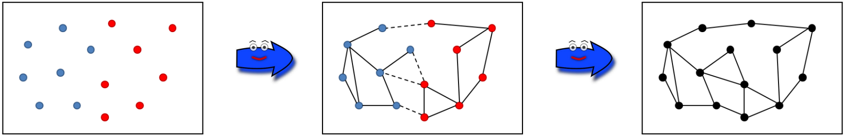
Planted partition model

[Holland et al '83; McSherry '01]

Question: How ‘good’ are these algorithms?

Framework:

- Given n blue nodes, and n red nodes
- Connect vertices of same color with probability p
- Connect vertices of different color with probability $q < p$



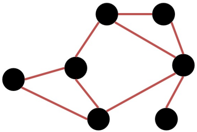
Remove labels

Question (formal): How many vertices are incorrectly labelled?

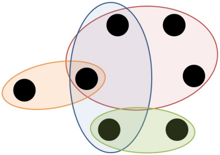
Answer: %error $\to 0$ as $n \to \infty$ [Rohe et al '11; Lei & Rinaldo '15]

Hypergraphs

- Each edge can connect more than two nodes

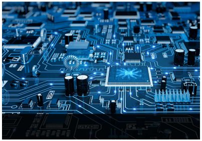


Graph

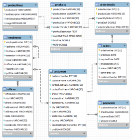


Hypergraph

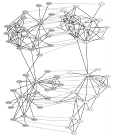
Applications:



Electronic circuits



Database systems



Molecular interations

Computer vision

Hypergraph partitioning: Algorithms & applications

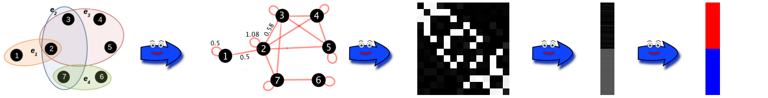
- Circuit design [Schweikert & Kernighan '79; Karypis & Kumar '00]
- Graph approximation for hypergraphs [Hadley '95]
- Spectral hypergraph partitioning [Zien et al '99]
- Categorical data clustering with hypergraphs [Gibson et al '00]
- Hypergraphs in computer vision [Agarwal et al '05]
- Tensor based algorithms [Govindu '05; Duchenne et al '11]
- Higher order learning [Zhou et al '07; Rota Bulo & Pellilo '13; etc.]

Hypergraph partitioning: Algorithms & applications

- Circuit design [Schweikert & Kernighan '79; Karypis & Kumar '00]
- Graph approximation for hypergraphs [Hadley '95]
- Spectral hypergraph partitioning [Zien et al '99]
- Categorical data clustering with hypergraphs [Gibson et al '00]
- Hypergraphs in computer vision [Agarwal et al '05]
- Tensor based algorithms [Govindu '05; Duchenne et al '11]
- Higher order learning [Zhou et al '07; Rota Bulo & Pellilo '13; etc.]

Spectral approach:

[Agarwal et al '05; Govindu '05; Zhou et al '07; Arias-Castro et al '11]



Reduced graph | Adjacency matrix | 2^{nd} eigenvector | labeling

The problem at hand

Question: How 'good' are the hypergraph partitioning algorithms?

The problem at hand

Question: How ‘good’ are the hypergraph partitioning algorithms?

Challenge 1:

- Existing planted partition model can only generate graphs
- Need a generalized model for hypergraphs

The problem at hand

Question: How 'good' are the hypergraph partitioning algorithms?

Challenge 1:

- Existing planted partition model can only generate graphs
- Need a generalized model for hypergraphs

Challenge 2:

- Graph adjacency: Random matrix with independent entries
- Adjacency of reduced graph:
  - Entries not independent
  - Need alternative tools for analysis

The problem at hand

Question: How ‘good’ are the hypergraph partitioning algorithms?

Challenge 1:

- Existing planted partition model can only generate graphs
- Need a generalized model for hypergraphs

Challenge 2:

- Graph adjacency: Random matrix with independent entries
- Adjacency of reduced graph:
  - Entries not independent
  - Need alternative tools for analysis

Challenge 3:

- Hypergraphs can have too many edges (computationally expensive)
- Practical approach: Edge sampling
- **Question:** How good are sampled algorithms?

Glimpse of the answer

Planted partition model:

- Hypergraph on n nodes, and each edge of size m
- k unknown classes of equal size, $k = O\left(\frac{n^{1/4}}{\log n}\right)$
- Unknown edge probabilities within class $= p$, and across classes $= q < p$

Theorem [Ghoshdastidar & Dukkipati '15; '16a; '16b]

With probability $1 - o(1)$

- number of vertices incorrectly labelled $= O\left(\frac{n^{(3-m)/2}}{(\log n)^{2m-3}}\right)$
- For $m = 2$, %error$\rightarrow 0$ as $n \rightarrow \infty$; For $m > 2$, error$\rightarrow 0$ as $n \rightarrow \infty$
- %error$\rightarrow 0$ even if only $\Omega\left(\frac{1}{n^{(m-1.5)/2}(\log n)^{2m-3}}\right)$ fraction of edges sampled

---

- Ghoshdastidar, D. & Dukkipati, A. (2015). In *Int. Conf. on Machine Learning (ICML)*.
- Ghoshdastidar, D. & Dukkipati, A. (2016a). *The Annals of Statistics* (in press). arXiv:1505.01582.
- Ghoshdastidar, D. & Dukkipati, A. (2016b). *Manuscript submitted*. arXiv:1602.06516.

Thank you

Publications based on this work

1. Ghoshdastidar, D., & Dukkipati, A. (2016). Consistency of spectral hypergraph partitioning under planted partition model. *Annals of Statistics* (in press) arXiv:1505.01582.
2. Ghoshdastidar, D., Adsul, A. P., & Dukkipati, A. (2016). Learning with Jensen-Tsallis kernels. *IEEE Transactions on Neural Networks and Learning Systems* (in press).
3. Ghoshdastidar, D. & Dukkipati, A. (2016). Uniform hypergraph partitioning: Provable tensor methods and sampling techniques. *Manuscript submitted* arXiv:1602.06516.
4. Ghoshdastidar, D. & Dukkipati, A. (2015). Coloring random non-uniform bipartite hypergraphs. *Manuscript submitted* arXiv:1507.00763.
5. Ghoshdastidar, D. & Dukkipati, A. (2015). A provable generalized tensor spectral method for uniform hypergraph partitioning. In *Proceedings of International Conference on Machine Learning (ICML)* 400–409.
6. Ghoshdastidar, D. & Dukkipati, A. (2015). Spectral clustering using multilinear SVD: Analysis, approximations and applications. In *Proceedings of the AAAI Conference on Artificial Intelligence* 2610–2616.
7. Ghoshdastidar, D. & Dukkipati, A. (2014). Consistency of spectral partitioning of uniform hypergraphs under planted partition model. In *Advances in Neural Information Processing Systems (NIPS)* 397–405.
8. Ghoshdastidar, D., Dukkipati, A., Adsul, A. P., & Vijayan, A. S. (2014). Spectral clustering with Jensen-type kernels and their multi-point extensions. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* 1472–1477.

References

- Agarwal, S., Lim, J., Zelnik-Manor, L., Perona, P., Kriegman, D. & Belongie, S. (2005). In *IEEE Computer Vision and Pattern Recognition* 838-845.
- Arias-Castro, E., Chen, G., & Lerman, G. (2011). *Electronic Journal of Statistics* **5** 15371587. In *IEEE Computer Vision and Pattern Recognition* 838-845.
- Duchenne, O., Bach, F., Kweon, I.-S. & Ponce, J. (2011). *IEEE Transactions on Pattern Analysis and Machine Intelligence* **33(12)** 2383-2395.
- Fiedler, M. (1973). *Czechoslovak Mathematical Journal* **23(2)** 298305.
- Gibson, D., Kleinberg, J. & Raghavan, P. (2000). *VLDB Journal* **8** 222-236.
- Govindu, V. M. (2005). In *IEEE Computer Vision and Pattern Recognition* 1150-1157.
- Hadley, S. W. (1995). *Discrete Applied Mathematics* **59** 115-127.
- Holland, P. W., Laskey, K. B. & Leinhardt, S. (1983). *Social Networks* **5** 109-137.
- Karypis, G. & Kumar, V. (2000). *VLSI Design* **11** 285-300.
- Lei, J. & Rinaldo, A. (2015). *Annals of Statistics* **43** 215-237.
- McSherry, F. (2001). In *Foundations of Computer Science* 529537.
- Rohe, K., Chatterjee, S., & Yu, B. (2011). *Annals of Statistics* **39** 1878-1915.
- Rota Bulo, S. & Pellilo, M. (2013). *IEEE Transactions on Pattern Analysis and Machine Intelligence* **35(6)** 1312-1327.
- Schweikert, G. & Kernighan, B. W. (1979). In *Design Automation Workshop* 57-62.
- Shi, J. & Malik, J. (2000). *IEEE Transactions on Pattern Analysis and Machine Intelligence* **22(8)** 888-905.
- Vu, V. (2014). *arXiv preprint* arXiv:1404.3918.
- Zien, J. Y., Schlag, M. D. F. and Chan, P. K. (1999). *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems* **13(9)** 1088-1096.
- Zhou, D., Huang, J. and Schölkopf, B. (2007). In *Advances in Neural Informal Processing Systems* 1601-1608.