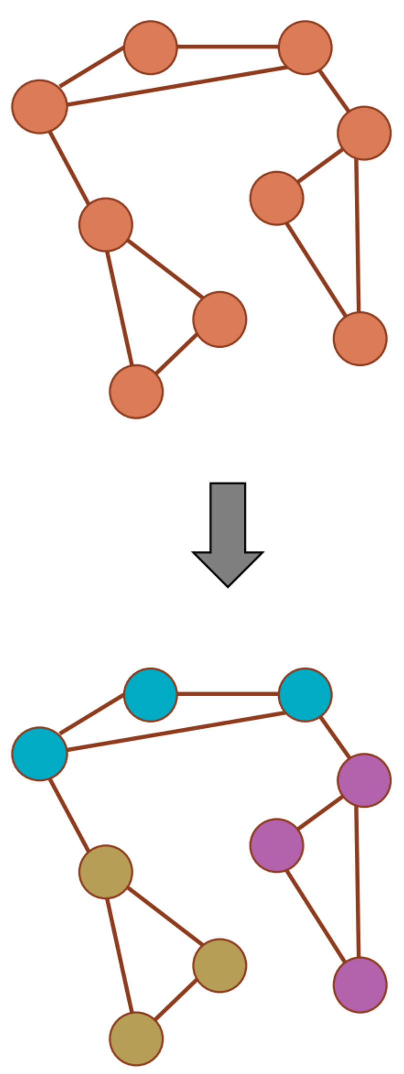
Robust and Efficient Multi-Way Spectral Clustering

Anil Damle[1,3], Victor Minden[1], Lexing Ying[1,2]

1. Institute for Computational and Mathematical Engineering, Stanford University
2. Department of Mathematics, Stanford University
3. Department of Mathematics, University of California, Berkeley

Introduction



- Consider clustering a graph distributed according to the *stochastic block model (SBM)*, where each edge is independent Bernoulli with probability p (within a cluster) or q (between two clusters).

$$\mathbb{E}[A] = M = \Pi \begin{pmatrix} p & p & q & q & q & q \\ p & p & q & q & q & q \\ q & q & p & p & q & q \\ q & q & p & p & q & q \\ q & q & q & q & p & p \\ q & q & q & q & p & p \end{pmatrix} \Pi^T$$

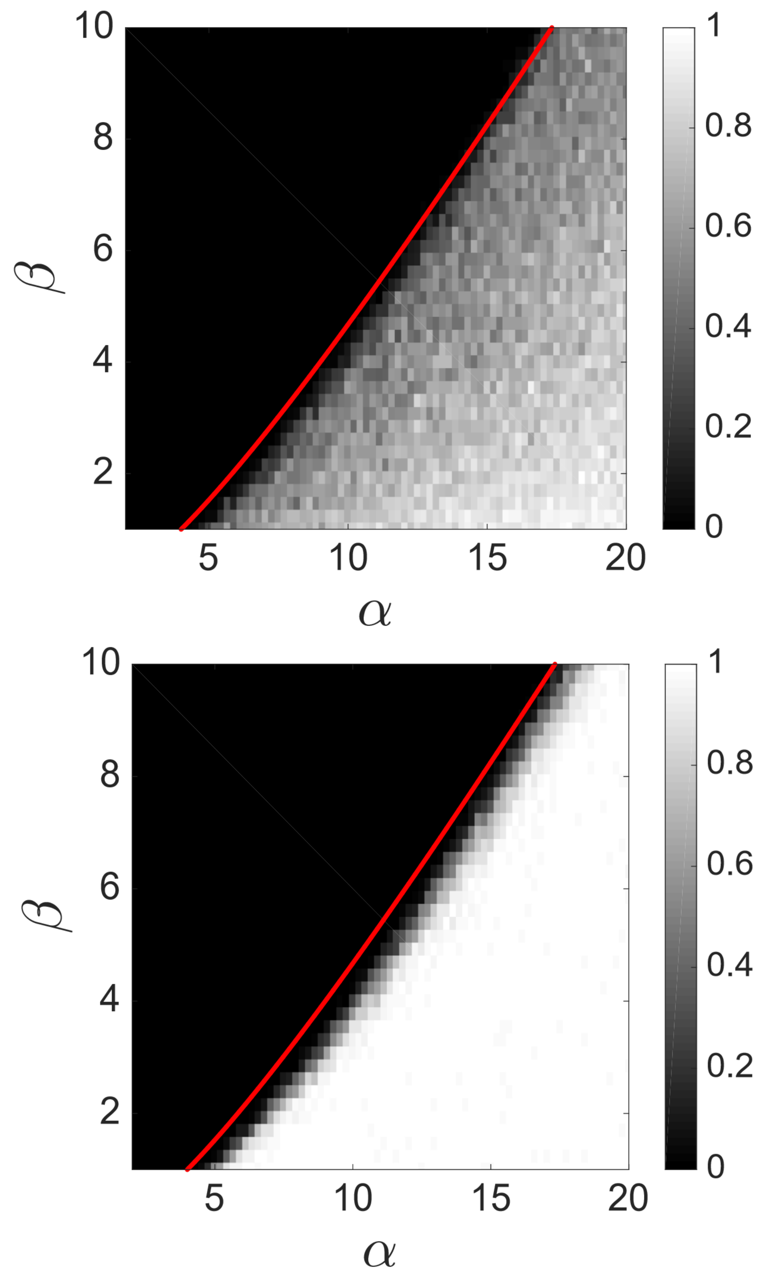
- To recover the clustering, *spectral clustering* typically takes eigendecomposition of adjacency matrix (or Laplacian) and then runs k-means on the eigenvectors.
- But, k-means is a non-convex optimization problem and suffers from local minima.
- There is more structure to be used! Eigenvectors of M are rotated indicator vectors.

Numerical Results

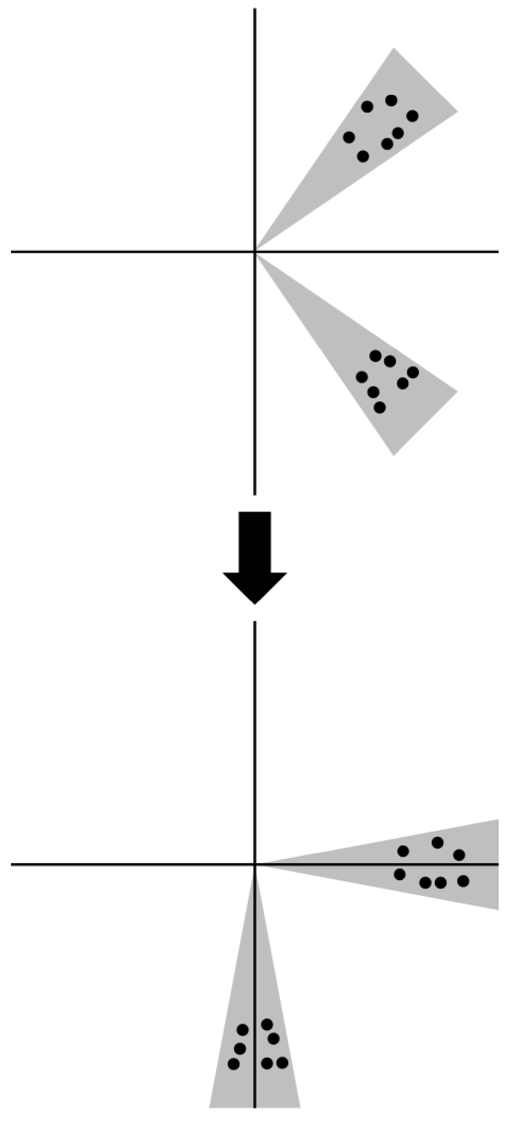
- Right: k-means++ (Arthur & Vassilvitskii) versus our algorithm, compared on the task of exact recovery of the SBM in semi-sparse regime, which is known to exhibit a phase transition (see Abbe et al.). Our method gives clean recovery near the theoretical limit, though proving this remains future work.
- Below: k-means++ versus our algorithm, compared on the ArXiV astrophysics collaboration graph. When seeking six clusters, we find that seeding k-means with our approach gives the best clustering according to two different metrics, when compared to 50 different random initializations using k-means++.

Table 1. Comparison of deterministic CPQR-based clustering and `k-means++`.

Algorithm	`k-means` objective	multi-way cut
`k-means++` mean	1.36	8.48
`k-means++` median	1.46	10.21
`k-means++` minimum	0.76	1.86
`k-means++` maximum	2.52	42.03
CPQR-based algorithm	2.52	1.92
`k-means` seeded with our algorithm	0.76	1.86



Algorithm: Basic Idea



- Eigenvectors of A are "almost" a rotation of indicator vectors on a cluster for SBM, similar structure for other applications (see Fiedler and Schiebinger et al.)
- Idea: rotate back to near-indicator vectors, then read off cluster assignment
- Key points
  I. Choose one node per cluster
     (pivoted QR of eigenvector matrix)
  II. Find basis describing clusters
     (polar factorization restricted to selected nodes)
  III. Rotate to align all nodes with the selected nodes
     (Apply polar factor to eigenvector matrix)
- Based on ideas from the quantum chemistry literature (see Damle et al.)
- Pivoted QR for k-means previously explored by Zha et al.

Key References

ABBE, E., BANDEIRA, A. S. & HALL, G. (2016) Exact Recovery in the Stochastic Block Model. *IEEE Transactions on Information Theory*, 62(1), 471–487.

ARTHUR, D. & VASSILVITSKII, S. (2007) k-means++: The advantages of careful seeding. in *Proc of the 18th annual ACM-SIAM symposium on discrete algorithms*, pp. 1027–1035. SIAM.

DAMLE, A., LIN, L. & YING, L. (2015) Compressed Representation of Kohn-Sham Orbitals via Selected Columns of the Density Matrix. *Journal of Chemical Theory and Computation*, 11(4), 1463–1469, PMID: 26574357.

FIEDLER, M. (1973) Algebraic connectivity of graphs. *Czechoslovak mathematical journal*, 23(2), 298–305.

HIGHAM, N. J. (1986) Computing the Polar Decomposition with Applications. *SIAM Journal on Scientific and Statistical Computing*, 7(4), 1160–1174.

SCHIEBINGER, G., WAINWRIGHT, M. J., YU, B. ET AL. (2015) The geometry of kernelized spectral clustering. *The Annals of Statistics*, 43(2), 819–846.

ZHA, H., HE, X., DING, C., GU, M. & SIMON, H. D. (2001) Spectral relaxation for k-means clustering. in *Advances in neural information processing systems*, pp. 1057–1064.

Acknowledgments

This work was supported by the National Science Foundation [DMS-1606277 to A.D.]; Stanford [Stanford Graduate Fellowship to V.M.]; and the United States Department of Energy [DE-FG02-97ER25308 to V.M., DE-FC02-13ER26134 to L.Y., and DE-SC0009409 to L.Y.].

The authors thank Austin Benson, Sven Schmit, Nolan Skochdopole, and Yuekai Sun for useful discussion, as well as Stanford University and the Stanford Research Computing Center for providing computational resources and support that have contributed to these research results

ICME INSTITUTE for COMPUTATIONAL & MATHEMATICAL ENGINEERING at STANFORD UNIVERSITY

DOE CSGF