Discovering Block Structure in Graphs with Approximate Eigenvectors

James Fairbanks[1] **and Geoff Sanders**[2]
Georgia Institute of Technology[1] and Lawrence Livermore National Laboratory[2]

Contact Information:
Email: `james.fairbanks@gatech.edu`
Email: `sanders29@llnl.gov`

Abstract

Graphs and Networks are important in modeling structure across disciplines. We demonstrate that graph partitioning in a minimum cut sense can be improved with an ensemble of low-fidelity eigenvectors which can outperform a single high-fidelity eigenvector. These ensembles can be computed faster and can be more helpful than one high-fidelity eigenvector. Since the individual ensemble members are independent they can be computed in parallel. This effect arises because of the discrepancy between solutions to a continuous relaxation of a discrete optimization problem and the discrete optimization solutions.

Introduction

Graphs can represent structure in diverse application areas. Stochastic Block Models (SBM) are controlled models of community structure, where the probability of each edge depends only on the two communities of the vertices. Eigenvectors can recover minimum cut partitions of graphs. SBM parameters can be learned using spectral partitioning [2].

- Adjacency Matrix A: $A_{i,j}$ counts the edges between vertex i and j
- Laplacian: $L = D - A$ where $D_{ii} = degree(i) = \sum_j A_{ij}$
- Normalized Laplacian: $\hat{L} = I - D^{-1/2}AD^{-1/2}$
- Eigenvalue λ Eigenvector x: $\hat{L}x = \lambda x$

The eigenvectors of associated matrices can be useful for solving graph problems such as the min cut problem $min_{b\in\{-1,1\}^n} b^T Lb$. The Fiedler Vector solves the continuous relaxation $min_{\{x\in\mathbb{R}^n, x\perp 1, \|x\|=1\}} x^T Lx$ [1].

Graphs can be represented by their matrices and visualized to see regular or block structure. If that structure is not known it can be hard to detect.

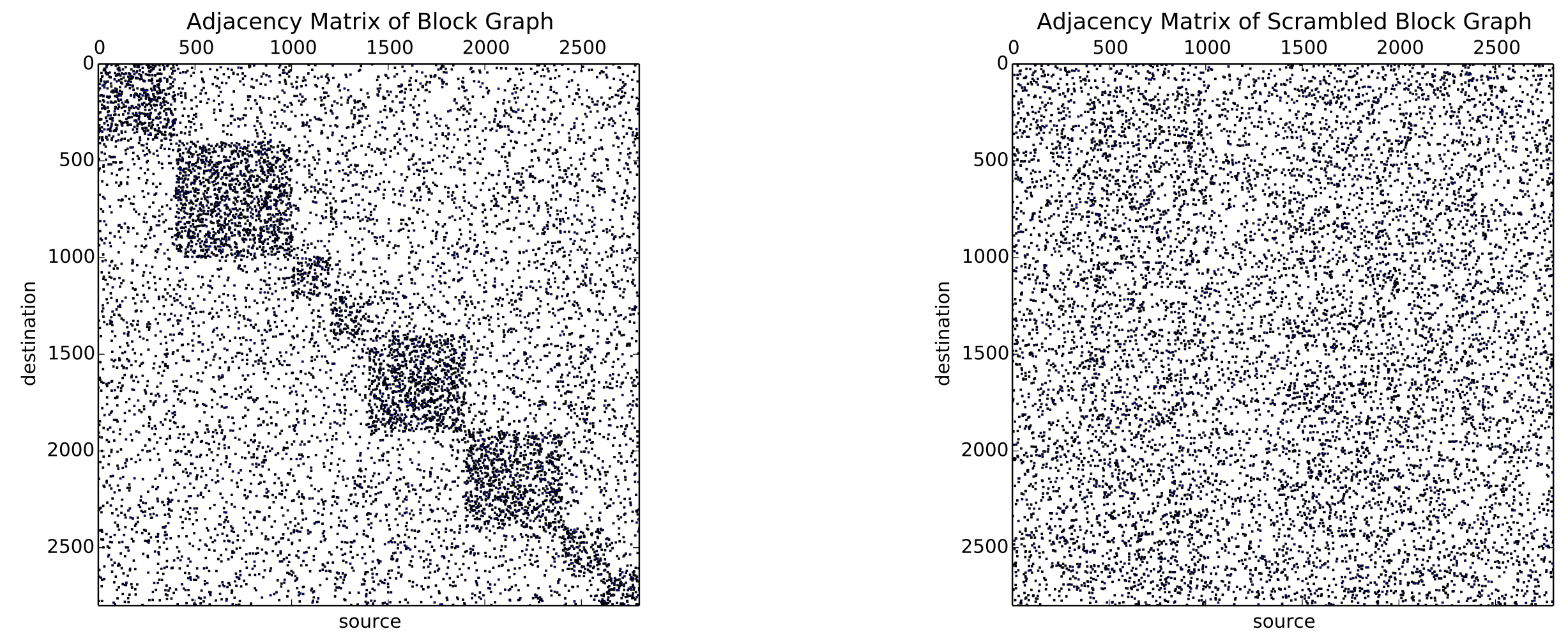

Figure 1: Matrix from Stochastic Block Model (left), and the same matrix permuted to hide structure (right).

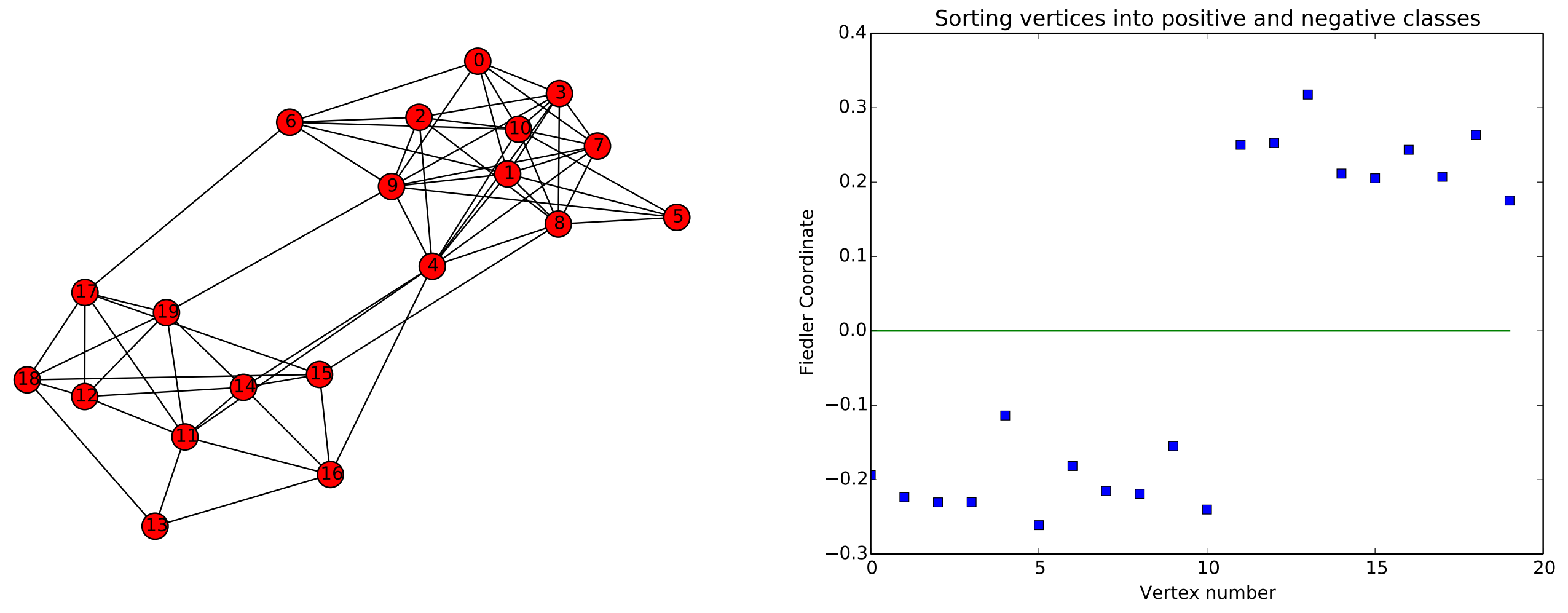

Figure 2: Small graph showing block structure (left) Graph partitioned according to Fiedler Vector

If we can solve min-cut with low-fidelity eigenvectors then eigenvectors can be used in a streaming environment, where new connections are being formed rapidly.

Results

Eigenvector solvers such as `ARPACK` [3] use random seed vectors to compute a candidate eigenvector. We can see the effects of treating the eigenvector returned from `ARPACK` as a random variable. For an Erdős-Rényi random graph Laplacian, we see that as the tolerance on the eigenresidual decreases the spread of the approximate solution vectors also decreases.

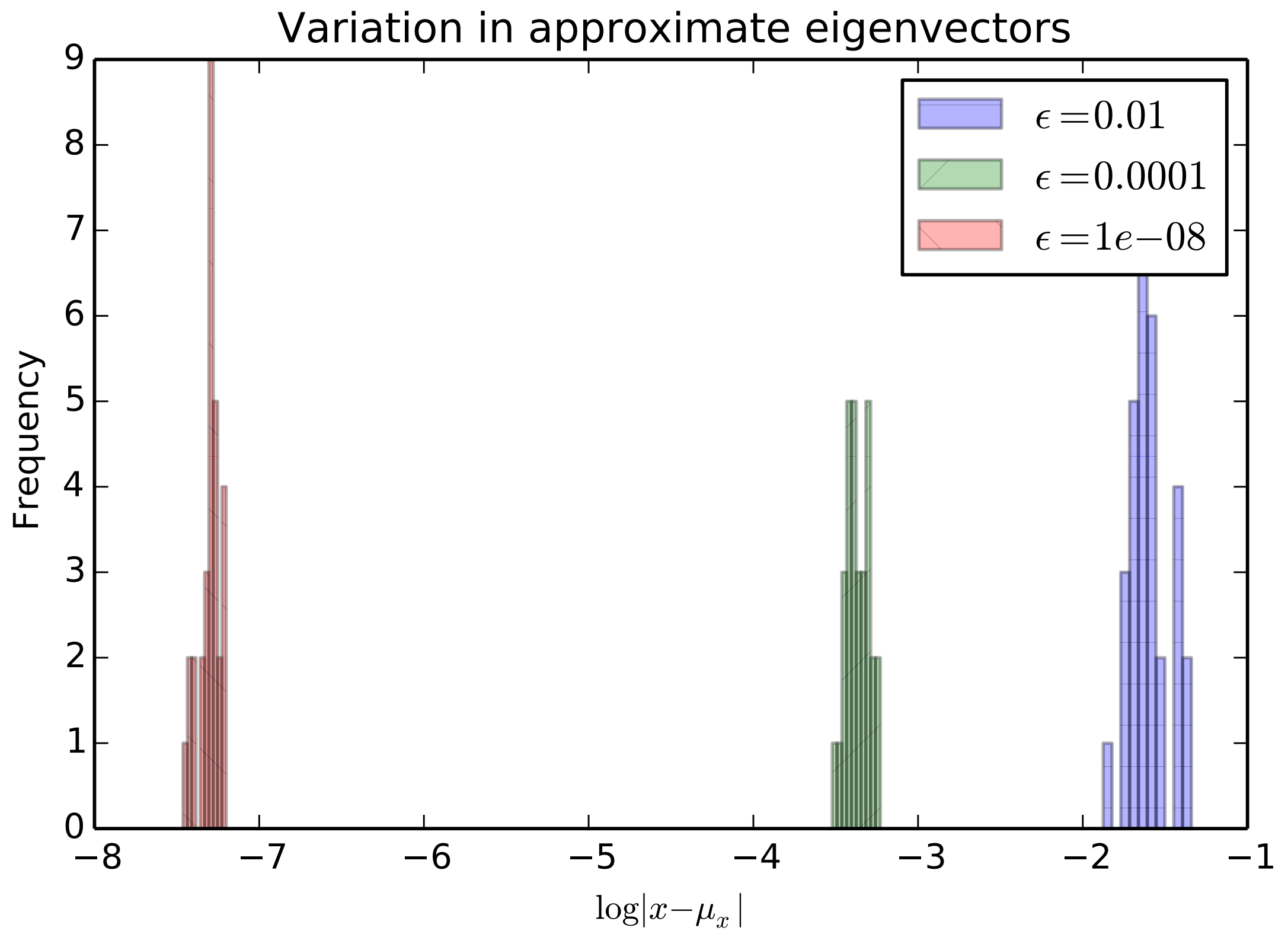

Figure 3: Tighter eigenresidual bounds imply tighter distributions of approximate solutions. Log scale implies that the distributions on the left have smaller variances.

When examining a near-bipartite community graph, we see that the cut size is more concentrated for tighter tolerances than for looser tolerances. As we get more digits of accuracy on the eigenvectors, we are more likely to get the equivalent partitionings from the vectors. Since the goal is to find the minimum cut of the graph, we can take several approximate eigenvectors and take the best induced cut. Figure 4 shows the distribution of cut size when an ensemble of eigenvectors are computed. The minimum of the distribution is the most important part.

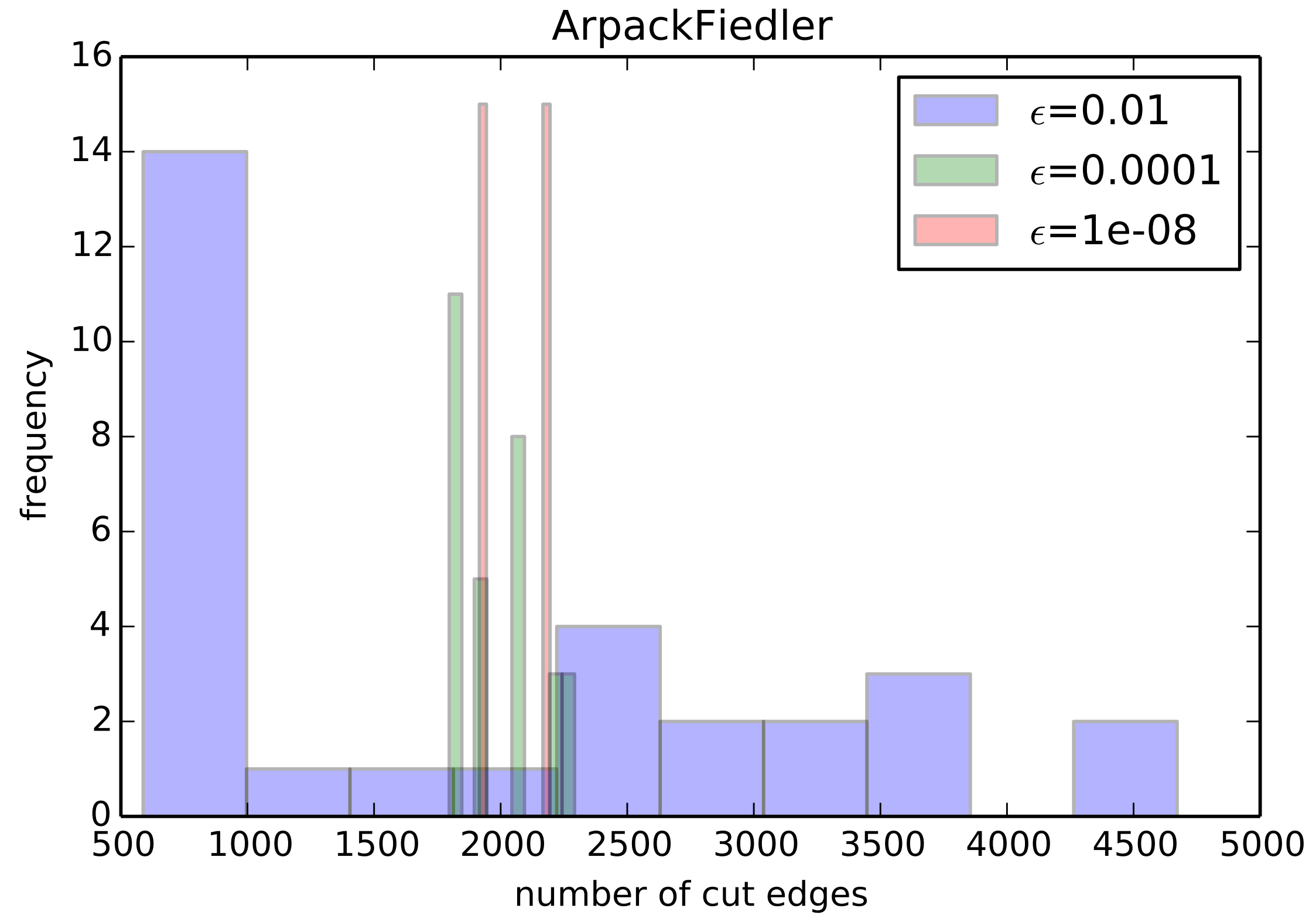

Figure 4: Distribution of supervised cut size using Fiedler vectors of various accuracy

Given a vector, we can split the graph at any threshold value. This gives a set of possible cuts. Evaluating the cut size at every possible split produces an indication of how many true blocks are in the graph. We can also see that a random vector used for splitting vastly underperforms eigenvector based splitting.

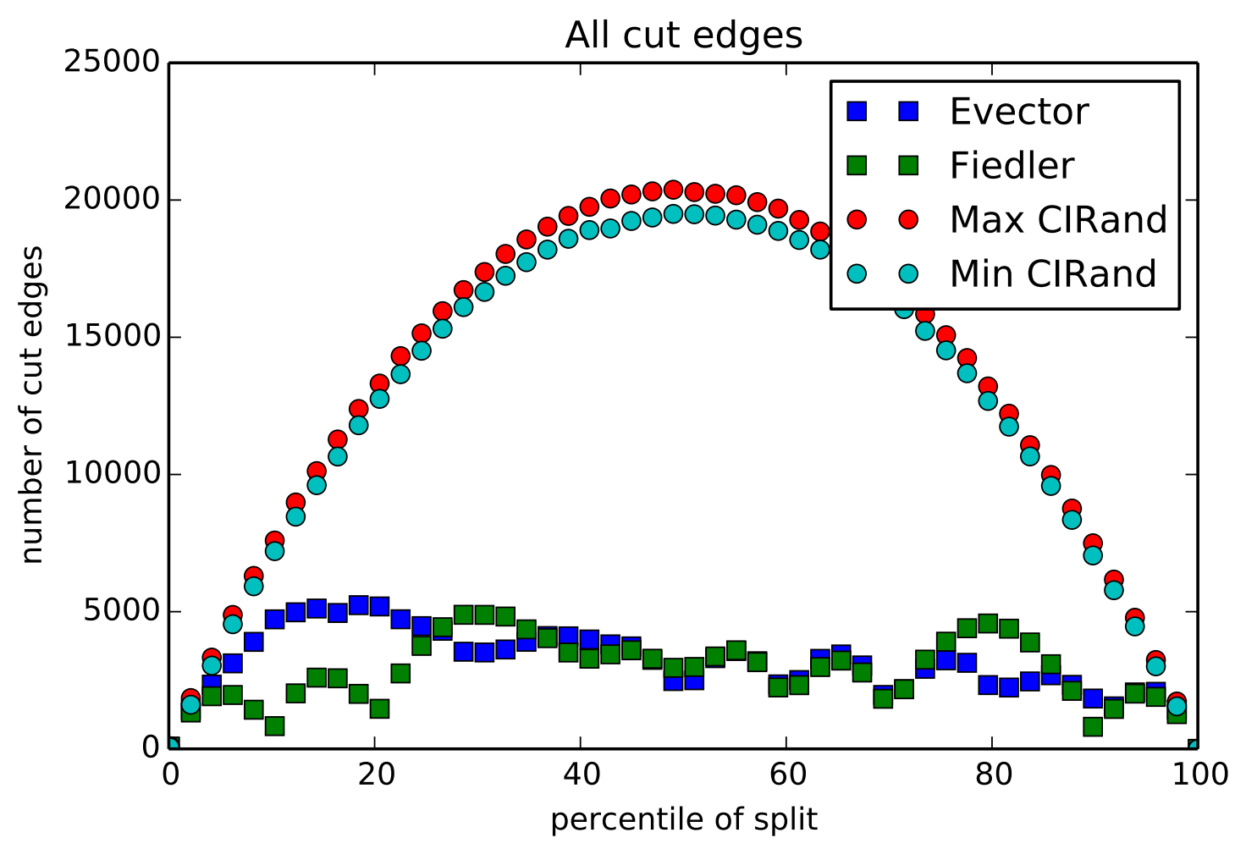

Figure 5: Comparing high-fidelity eigenvectors to random vectors. Here Evector refers to the absolute value of the maximal eigenvector of the Laplacian, which reveals structure in near bipartite graphs.

Conclusions

- Large residual Fiedler vector approximations can outperform the small residual approximations at finding a minimum cut.
- At large tolerances, randomized eigensolvers have useful random variation.
- Ensembles of weak numerical solutions can outperform strong numerical solutions in data mining.
- Stochastic Block Models reveal the quality performance of graph partitioning algorithms.

Forthcoming Research

- Further research should examine these techniques in the streaming environment.
- Thorough analysis of cost to perform multiple low-fidelity solves against the cost to perform one high-fidelity solve.
- Using other structured graphs to examine the interaction between approximate numerical solutions and data analysis solutions.
- Extension to other Machine Learning and Data Mining tasks.

References

[1] Fan RK Chung. *Spectral graph theory*, volume 92. American Mathematical Soc., 1997.

[2] D. Fishkind, D. Sussman, M. Tang, J. Vogelstein, and C. Priebe. Consistent adjacency-spectral partitioning for the stochastic block model when the model parameters are unknown. *SIAM Journal on Matrix Analysis and Applications*, 34(1):23–39, 2013.

[3] R. Lehoucq and D. Sorensen. Deflation techniques for an implicitly restarted arnoldi iteration. *SIAM Journal on Matrix Analysis and Applications*, 17(4):789–821, 1996.

Acknowledgments

This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344 and the ASEE-NDSEG fellowship supporting my PhD at Georgia Tech. LLNL-POST-668628