SAND2015-20785C

EFFECTS OF GRAPH STRUCTURE ON 2D PARTITIONING OF SCALE-FREE GRAPHS WITH SAMPLING

Michael M. Wolf, Benjamin A. Miller

SIAM Workshop on Network Science 2015
May 16-17 · Snowbird

Summary

Eigenspace analysis of large-scale graphs is useful in a number of important application domains, such as social network analysis, cyber security, and biology. When performing this sort of analysis across many parallel processes, the data partitioning scheme is extremely important and may have a significant impact on the overall running time. Previous work demonstrated that partitioning based on a subset of edges still yields a substantial improvement in running time, and in this work we explore the effect of community structure and degree distribution within this context.

Eigenspace analysis of large networks

Computing extreme eigenvectors of matrices based on graphs is a fundamental problem in a wide variety of application areas. Common methods for community detection include eigenvector analysis of the graph Laplacian [4] or the modularity matrix [9]. A common measure of vertex importance is PageRank [10], which is based on the principal eigenvector of a modified adjacency matrix. Eigenvector analysis is also frequently used in anomaly detection, either in detection of anomalous clusters [7, 8] or of global anomalous connectivity [6]. In all of these applications, a sparse eigensolver is required to perform the desired computation.

Eigenspace analysis of extremely large graphs requires parallel processing, and in this context significant complications arise. Data partitioning across several processes when the data are in the form of a social or computer network is a relatively new problem. Traditional matrix partitioning algorithms have, for the most part, focused on and been most effective for sparse matrices that have a regular structure (e.g., matrices derived from meshes). However, there has been recent on research focused on partitioning large networks—with community structure and skewed degree distributions, as in many graphs of interest—to best improve the time required to multiply a sparse matrix by a dense vector (denoted SpMV), which is the key kernel for a sparse eigensolver.

Partitioning for large, scale-free graphs

Matrix partitioning for SpMV operations has, in practice, largely focused on 1-dimensional distributions, where a set of rows is assigned to a process. In cases where there is no known structure to the rows, this may require all processes to communicate with all other processes, since all rows of the dense vector may have to be multiplied by each block of rows. It was shown by Yoo et al. that simply randomly partitioning in two dimensions provides a substantial speedup over 1D partitioning [14]. Later, Boman et al. demonstrated that using hypergraph partitioning to determine the 2D structure provides an even better speedup in real-world graphs [2].

The hypergraph partitioning algorithm, however, is expensive and may not be amortized over the SpMV operations required to compute the desired eigenvectors. Therefore, in order for the method to be computationally beneficial, the partition must either (1) be useful for multiple time steps in a dynamic graph, or (2) be computable from a sampled graph which will take much less time. These were explored in earlier work [12, 13], on R-MAT synthetic graphs [3] and a film actor's network. In this work, we consider the effect of community structure on performance when the graph is sampled.

Experiments

We generated R-MAT matrices with 2^{20} vertices and an average of 100 nonzeros per row. These matrices were partitioned with the random 2D method (2DR) or the hypergraph method (2DH). For the hypergraph method, we sampled a $1/2^i$ edges at random for $0 \leq i \leq 10$. The

M. M. Wolf is with Sandia National Laboratories (email: mmwolf@sandia.gov).

B. A. Miller is with MIT Lincoln Laboratory (email: bamiller@ll.mit.edu).

Sandia National Laboratories is a multi-program laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.

base probability matrix for the R-MAT graphs was given by

$$\begin{bmatrix} a & (0.75-a)/2 \\ (0.75-a)/2 & 0.25 \end{bmatrix}, \tag{1}$$

with a set to 0.25, 0.375, 0.5, and 0.625. For $a = 0.25$, the resulting graph will be an Erdős–Rényi graph, where all edges occur with equal probability. As a increases, the community structure in the graph increases as the two halves begin to interact less.

The SpMV operation was performed with the Anasazi library [1], part of the Trilinos project [5]. Results are shown in Figure 1. As expected, partitioning time is significantly reduced in all cases as the proportion of edges sampled decreases. The 2DH method provides a benefit in most cases for the values of a corresponding to more community structure (0.5 and 0.625), while not providing a benefit for cases with less.

Discussion

Consistent with previous results [13], we see that sampling (even at fairly low sampling rates), is very effective at reducing the 2DH partitioning time and even improves the partitioning quality, reducing the SpMV time. In the SpMV running times shown in the figure, when the matrix is more similar to an Erdős–Rényi graph, 2DH partitioning never helps a great deal. This makes intuitive sense: Since there is no structure to the graph, partitioning randomly will provide about the same performance as looking for ideal cuts. One surprising aspect of the results is that, for the cases with more community structure, performance actually *improves* when the data were sampled. This may be an artifact of the R-MAT generator. It could be that, when there are fewer edges, the weak community structure of the R-MAT graph is more apparent, and as more edges between the two halves (and their recursive counterparts) are added, the partitioning method does not see the advantage of separating in certain parts of the graph from others. This would imply that there still is an advantage for SpMV, which is revealed in the sparsified graph.

This is an interesting phenomenon that will be explored in more depth in future work. In addition, while the R-MAT generator was used here for the sake of continuity with earlier work, future experiments will also use the Block Two-level Erős–Rényi (BTER) generator [11], which has much more substantial community structure.

References

[1] C. G. Baker, U. L. Hetmaniuk, R. B. Lehoucq, and H. K. Thornquist. Anasazi software for the numerical solution of large-scale eigenvalue problems. *ACM Trans. Math. Softw.*, 36(3):13:1–13:23, July 2009.

[2] E. G. Boman, K. D. Devine, and S. Rajamanickam. Scalable matrix computations on large scale-free graphs using 2D graph partitioning. In *Proc. Supercomputing*, pages 50:1–50:12, 2013.

[3] D. Chakrabarti, Y. Zhan, and C. Faloutsos. R-MAT: A recursive model for graph mining. In *Proc. of the 4th SIAM Conference on Data Mining*, pages 442–446, 2004.

[4] S. Fortunato. Community detection in graphs. *Physics Reports*, 486:75–174, February 2010.

[5] M. A. Heroux, R. A. Bartlett, V. E. Howle, R. J. Hoekstra, J. J. Hu, T. G. Kolda, R. B. Lehoucq, K. R. Long, R. P. Pawlowski, E. T. Phipps, A. G. Salinger, H. K. Thornquist, R. S. Tuminaro, J. M. Willenbring, A. Williams, and K. S. Stanley. An overview of the Trilinos project. *ACM Trans. Math. Softw.*, 31(3):397–423, 2005.

[6] T. Idé and H. Kashima. Eigenspace-based anomaly detection in computer systems. In *Proc. ACM Int. Conf. Knowledge Discovery and Data Mining*, pages 440–449, 2004.

[7] B. A. Miller, N. T. Bliss, P. J. Wolfe, and M. S. Beard. Detection theory for graphs. *Lincoln Laboratory J.*, 20(1):10–30, 2013.

[8] R. R. Nadakuditi. On hard limits of eigen-analysis based planted clique detection. In *Proc. IEEE Statistical Signal Process. Workshop*, pages 129–132, 2012.

[9] M. E. J. Newman. Finding community structure in networks using the eigenvectors of matrices. *Phys. Rev. E*, 74(3), 2006.

[10] L. Page, S. Brin, R. Motwani, and T. Winograd. The PageRank citation ranking: Bringing order to the web. Technical report, Stanford InfoLab, 1999.

[11] C. Seshadhri, T. G. Kolda, and A. Pinar. Community structure and scale-free collections of erdős-rényi graphs. *Physical Review E*, 85(5):056109, 2012.

[12] M. M. Wolf and B. A. Miller. Detecting anomalies in very large graphs. *SIAM Workshop Combinatorial Scientific Computing*, pages 43–44, 2014.

[13] M. M. Wolf and B. A. Miller. Sparse matrix partitioning for parallel eigenanalysis of large static and dynamic graphs. In *Proc. IEEE High Performance Extreme Computing Conf.*, 2014.

[14] A. Yoo, A. H. Baker, R. Pearce, and V. E. Henson. A scalable eigensolver for large scale-free graphs using 2D graph partitioning. In *Proc. Supercomputing*, pages 63:1–63:11, 2011.

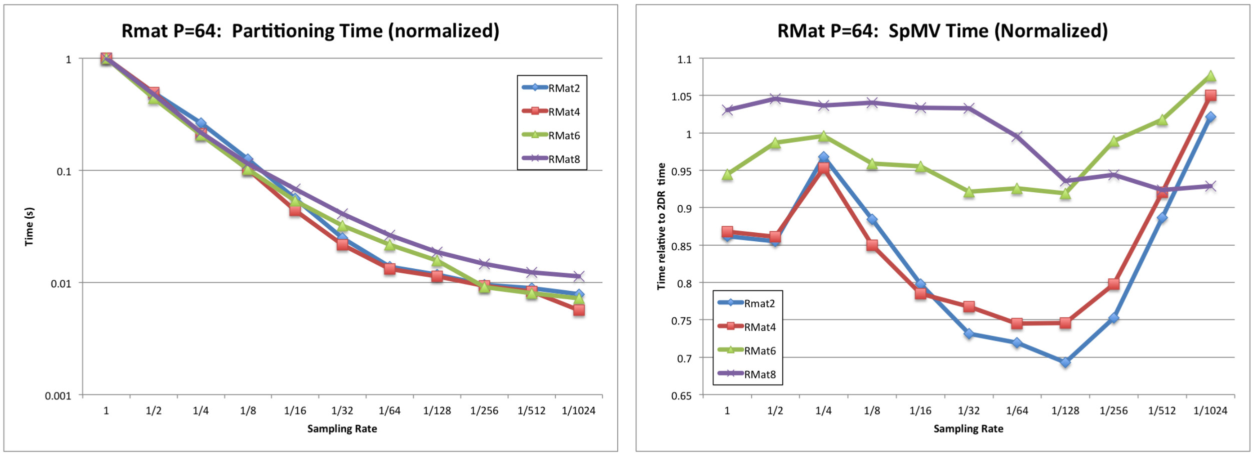

Figure 1: Relative time to partition the sparse matrix (left) and perform a SpMV operation (right).