ICML 2010 Tutorial: Geometric Tools for Identifying Structure in Large Social and Information Networks

Michael W. Mahoney *

1 Topic Overview

The tutorial will cover recent algorithmic and statistical work on identifying and exploiting "geometric" structure in large informatics graphs such as large social and information networks. Such tools (*e.g.*, Principal Component Analysis and related non-linear dimensionality reduction methods) are popular in many areas of machine learning and data analysis due to their relatively-nice algorithmic properties and their connections with regularization and statistical inference. These tools are *not*, however, immediately-applicable in many large informatics graphs applications since graphs are more combinatorial objects; due to the noise and sparsity patterns of many real-world networks, etc. Recent theoretical and empirical work has begun to remedy this, and in doing so it has already elucidated several surprising and counterintuitive properties of very large networks. Topics include: underlying theoretical ideas; tips to bridge the theory-practice gap; empirical observations; and the usefulness of these tools for such diverse applications as community detection, routing, inference, and visualization.

2 Target Audience

This tutorial will provide an opportunity for the ICML community, including both mathematically-oriented researchers as well as practitioners, to learn about recent algorithmic advances for dealing with very large social and information networks. Many of these algorithmic tools have implicit geometric properties associated with them; and these geometric properties often have implicit statistical properties and consequences that indicate where these tools are more or less useful in real-world applications. As such, this tutorial should be of interest to and accessible by a large fraction of the ICML community—including both:

- **established researchers** who have done work in this or related areas, as well as researchers whose interests are not directly in the topic of the tutorial; and
- **graduate students and postdocs**, as well as junior and more senior researchers.

Many of the algorithmic and statistical techniques to be discussed have a strong overlap with seemingly-different problems and questions in statistics, optimization, numerical analysis, and machine learning—these connections will be highlighted throughout. Relatedly, many of these questions have been studied by researchers in theoretical computer science, scientific computing, statistics, machine learning, and data analysis—the complementary aspects of these different

*Stanford University, Stanford, CA 94305, mmahoney@cs.stanford.edu.

approaches, including their applicability to solving real-world problems from different application domains, will be emphasized. Depending on one's background, one can expect to benefit in different ways from the tutorial. In particular:

- **Practitioners** of machine learning and data analysis should gain just enough insight into the theoretical underpinnings of relevant algorithms to see how and why algorithms work well or fail to work well in real-world settings.
- **Application-oriented theorists** should gain insight into how the inner-workings of algorithms have practical implications for machine learning and data analysis on large networks, as well as learn about interesting theoretical problems raise by recent empirical findings.
- **Knowledgeable members of the ICML community** should gain a broad overview of the area of large-scale graph mining and network analysis, including where ICML-type methods with which they are familiar are well-suited or ill-suited.

As the body of work to be covered has connections with work done in the machine learning and data mining communities, as well as related work in related areas, an improved understanding of this work in a broader context could significantly impact various Machine Learning and Data Mining applications. In addition, members of the audience with a range of backgrounds could benefit from an exposition of the theory in this broader context, as well as a description of applications that have already been considered or are natural future targets for these methods.

3 Content Details

Graphs and matrices arise naturally in many areas of data mining, machine learning, pattern recognition, and large-scale network analysis. For example, a common way to model a large social or information network is with an *interaction graph model*, $G = (V, E)$, in which nodes in the vertex set V represent "entities" and the edges (whether directed, undirected, weighted or unweighted) in the edge set E represent "interactions" between pairs of entities. Alternatively, these and other data sets can be modeled as matrices, since an $m \times n$ real-valued matrix A provides a natural structure for encoding information about m objects, each of which is described by n features. Due to their large size, their extreme sparsity, and their complex and often adversarial noise properties, data graphs and data matrices arising in modern informatics applications present considerable challenges and opportunities for interdisciplinary research.

In particular, note that tools such as Principal Component Analysis and related non-linear dimensionality reduction methods are popular in many areas of machine learning and data analysis due to their relatively-nice algorithmic properties and their connections with regularization and statistical inference. On the other hand, these tools are not immediately-applicable in many large informatics graphs applications, such as when dealing with large social and information networks. There are numerous reasons for this, including that graphs are often more-naturally viewed as combinatorial objects; that the noise and sparsity patterns of many real-world networks render such geometric and statistical interpretations difficult; and that many informatics graphs are "expander-like" and thus not even approximately "low-dimensional."

Recent theoretical and applied work has begun to remedy this disconnect, and in doing so it has already elucidated very precisely several surprising and counterintuitive properties of very large networks. Thus, the proposed tutorial will cover recent algorithmic and statistical work on identifying and exploiting "geometric" structure in large informatics graphs such as large social and information networks. Topics to be covered will include underlying theoretical ideas; tips to bridge the theory-practice gap; empirical observations; and the usefulness of these tools for

such diverse applications as community detection, routing, inference, and visualization. In more detail, here is an outline of the intended topics to be covered in this tutorial:

- **Popular algorithmic tools with a geometric flavor** (ca. 0.5 hours)
  - Computational/algorithmic as well as statistical/geometric issues underlying the use of PCA and SVD (including the relationship between geometric structure and issues of smoothing, regularization, and inference).
  - Interpretation of spectral methods (*e.g.*, eigenvector localization and connections to underlying social/economics concepts like homophily and centrality) and extensions of these tools to kernels and manifold-based non-linear dimensionality-reduction settings.
  - Difficulties applying these tools in very large informatics graph applications (including scaling and noise issues, as well as diagnostic tools to indicate failings or "inappropriateness" of these methods).
  - Representative references: [1], [2], [3].
- **Graph algorithms and their geometric underpinnings** (ca. 1.0 hours)
  - Spectral, flow-based, local-improvement, and multi-resolution methods for graph partitioning, including theoretical underpinnings and implementation issues.
  - Geometric and statistical perspectives, including the "worst case" examples for each method, the behavior on "typical" low-dimensional and expander-like graphs, and geometric assumptions underlying the use of these methods.
  - Recent "local" spectral methods, spectral-based and flow-based "cut improvement" methods, and methods that "interpolate" between spectral and flow, including empirical/implementational issues and connections with boosting and online learning.
  - Tools for identifying "negatively-curved" or "hyperbolic" structure, including uses in routing and visualization, connections with expander-like structure, and intuitive relationships with low-dimensional Euclidean and manifold-like structure.
  - Representative references: [4], [5], [6], [7], [8], [9], [10], [11],
- **Novel insights on geometric structure in large informatics graphs** (ca. 1.5 hours)
  - Heavy-tailed and small-world models that attempt to capture local clustering and/or large-scale heterogeneity of large graphs (including question of "pre-existing" versus "generated" geometry—*e.g.*, models that start with an underlying geometric scaffolding versus models that implicitly generate structures that are operationally geometric).
  - Empirical successes/failings of these popular models (including graph densification, diameters versus time, and small-scale versus large-scale clustering and community structure) in large-scale applications of interest in machine learning and data analysis, as well as technical and domain-specific challenges in validating success or failure.
  - Empirical results on the structure of large real-world graphs—including "experimental" methodologies that takes into account challenges related to the size, sparsity, and noise properties of real-world networks, as well as technical and non-technical issues in validation.
  - Empirical results on "local" geometric structure, "global" metric structure, and the coupling between the two size scales—including implications for clustering and community structure, diffusion of information, routing, visualization, and the design of machine learning and data analysis tools for large networks.

- Empirical results on implicit regularization by worst-case approximation algorithms that have geometric underpinnings (including behavior of different classes of algorithms on different classes of graphs, relationships to other operationally-defined regularization procedures, and applications to scalable community identification).
- Implications of these results for dynamics of processes *on* graphs and dynamic evolution *of* graphs; connections to hyperbolic metrics and implications for robust and reliable statistical inference on networks; and relationships to and implications for recently-developed statistical models of networks, including stochastic blockmodels.
- Representative references: [12], [13], [14], [15], [16], [17], [18], [19], [20].

Most of the tutorial will focus on tools that are appropriate for networks with between, say, 10^4 and 10^7 nodes. On the one hand, this size regime is well above the size scale that is commonly-studied in the "complex networks" literature, and numerous subtle and counterintuitive properties have already been observed in this regime. On the other hand, this size regime is well below the size scale where the data need to be stored in a distributed environments and memory-management issues become paramount. Thus, this size regime represents a nice domain for the development of novel algorithmic tools, drawing on existing machinery from theoretical computer science, machine learning, scientific computing, and data analysis.

Examples graphs that will be considered will include a large corpus of social and information networks drawn from Internet domains, and if possible graphs from large-scale biology and finance applications of the sort that have not been commonly-studied in this area thus far. Of course, although they will not be the primary focus of the tutorial, larger graphs (say, with billions of nodes) will be of interest insofar as one hopes that the "geometric network analysis" tools that will be described will be able to be integrated (in principle and hopefully in practice) into much larger-scale systems.

References

[1] A. N. Langville and C. D. Meyer. A survey of eigenvector methods of web information retrieval. *SIAM Review*, 47(1):135–161, 2005.

[2] L. K. Saul, K. Q. Weinberger, J. H. Ham, F. Sha, and D. D. Lee. Spectral methods for dimensionality reduction. In O. Chapelle, B. Schoelkopf, and A. Zien, editors, *Semisupervised Learning*, pages 293–308. MIT Press, 2006.

[3] M. Belkin and P. Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation*, 15(6):1373–1396, 2003.

[4] A. Pothen. Graph partitioning algorithms with applications to scientific computing. In D. E. Keyes, A. H. Sameh, and V. Venkatakrishnan, editors, *Parallel Numerical Algorithms*. Kluwer Academic Press, 1996.

[5] S. Guattery and G.L. Miller. On the quality of spectral separators. *SIAM Journal on Matrix Analysis and Applications*, 19:701–719, 1998.

[6] J. Shi and J. Malik. Normalized cuts and image segmentation. *IEEE Transcations of Pattern Analysis and Machine Intelligence*, 22(8):888–905, 2000.

[7] T. Leighton and S. Rao. Multicommodity max-flow min-cut theorems and their use in designing approximation algorithms. *Journal of the ACM*, 46(6):787–832, 1999.

[8] N. Linial, E. London, and Y. Rabinovich. The geometry of graphs and some of its algorithmic applications. *Combinatorica*, 15(2):215–245, 1995.

[9] R. Andersen, F.R.K. Chung, and K. Lang. Local graph partitioning using PageRank vectors. In *FOCS '06: Proceedings of the 47th Annual IEEE Symposium on Foundations of Computer Science*, pages 475–486, 2006.

[10] R. Andersen and K. Lang. An algorithm for improving graph partitions. In *SODA '08: Proceedings of the 19th ACM-SIAM Symposium on Discrete algorithms*, pages 651–660, 2008.

[11] S. Arora, S. Rao, and U. Vazirani. Geometry, flows, and graph-partitioning algorithms. *Communications of the ACM*, 51(10):96–105, 2008.

[12] M.E.J. Newman. The structure and function of complex networks. *SIAM Review*, 45(2):167–256, 2003.

[13] J. Leskovec, J. Kleinberg, and C. Faloutsos. Graph evolution: Densification and shrinking diameters. *ACM Transactions on Knowledge Discovery from Data*, 1(1), 2007.

[14] J. Leskovec, K.J. Lang, A. Dasgupta, and M.W. Mahoney. Statistical properties of community structure in large social and information networks. In *WWW '08: Proceedings of the 17th International Conference on World Wide Web*, pages 695–704, 2008.

[15] J. Leskovec, K.J. Lang, A. Dasgupta, and M.W. Mahoney. Community structure in large networks: Natural cluster sizes and the absence of large well-defined clusters. arXiv:0810.1355, October 2008.

[16] J. Leskovec, K.J. Lang, and M.W. Mahoney. Empirical comparison of algorithms for network community detection. In *WWW '10: Proceedings of the 19th International Conference on World Wide Web*, pages 000–000, 2010.

[17] R. Andersen and K. Lang. Communities from seed sets. In *WWW '06: Proceedings of the 15th International Conference on World Wide Web*, pages 223–232, 2006.

[18] K. Lang, M. W. Mahoney, and L. Orecchia. Empirical evaluation of graph partitioning using spectral embeddings and flow. In *Proc. 8-th International SEA*, pages 197–208, 2009.

[19] T. Munzner and P. Burchard. Visualizing the structure of the World Wide Web in 3D hyperbolic space. In *Proceedings of the first symposium on Virtual reality modeling language*, pages 33–38, 1995.

[20] M. Boguñá, D. Krioukov, and K.C. Claffy. Navigability of complex networks. *Nature Physics*, 5:74–80, 2009.