Advanced Graph Mining for Community Evaluation in Social Networks and the Web

Christos Giatsidis
École Polytechnique, France
giatsidis@lix.polytechnique.fr

Fragkiskos D. Malliaros
École Polytechnique, France
fmalliaros@lix.polytechnique.fr

Michalis Vazirgiannis
AUEB, Greece and
École Polytechnique, France
mvazirg@lix.polytechnique.fr

ABSTRACT

Graphs constitute a dominant data structure and appear essentially in all forms of information. Examples are the Web graph, numerous social networks, protein interaction networks, terms dependency graphs and network topologies. The main features of these graphs are their huge volume and rate of change. Presumably, there is important hidden knowledge in the macroscopic topology and features of these graphs. A cornerstone issue here is the detection and evaluation of communities – bearing multiple and diverse semantics.

The tutorial reports the basic models of graph structures for undirected, directed and signed graphs and their properties. Next we offer a thorough review of fundamental methods for graph clustering and community detection, on both undirected and directed graphs. Then we survey community evaluation measures, including both the individual node based ones as well as those that take into account aggregate properties of communities. A special mention is made on approaches that capitalize on the concept of degeneracy (k-cores and extensions), as a novel means of community detection and evaluation. We justify the above foundational framework with applications on citation graphs, trust networks and protein graphs.

Categories and Subject Descriptors

H.4 [**Database Management**]: Database applications—*Data mining*

General Terms

Algorithms; Measurement

Keywords

Community structure; Community detection; Social network analysis; Graph mining

1. INTRODUCTION

Graphs (or networks) appear in several diverse domains, including sociology, biology, neuroscience and information management. An interesting feature of real networks is the clustering or community structure property, i.e., the structure is based into a modular organization; nodes within the same module (cluster or community) tend to be highly similar sharing common features, while on the other hand, nodes of different modules show low similarity. Typically, the communities correspond to densely connected groups of nodes, where the number of edges within a community is much higher than the number of edges across different communities.

Detecting and evaluating the community structure of real-world graphs constitutes an essential task in the area of graph mining and social network analysis. For example, in the link structure of the Web, communities correspond to groups of web pages that share common topics, and therefore, revealing the underlying community structure is a crucial application from a web search engine perspective. Similarly, communities in a social network (e.g., Facebook, Twitter) correspond to individuals with increased social ties (e.g., friendship relationships, common interests). Broadly speaking, community discovery and evaluation can contribute in our understanding of a social system, summarizing the interactions within the system in a concise manner.

The aim of this tutorial is to review basic methods and tools for the task of community detection and evaluation in real graphs. Since a plethora of diverse approaches have been presented in the area, we focus on the fundamental ones, demonstrating their basic methodological principles.

2. BRIEF OUTLINE

The goal of the tutorial is to present community detection and evaluation techniques as mining tools for social networks and the Web. More precisely, the following key topics are covered by the tutorial: (i) Introduction and preliminaries on graphs and graph mining - social network analysis; (ii) topics in graph clustering and community detection; (iii) clustering and community detection in directed graphs; (iv) degeneracy-based community evaluation.

The tutorial commences by presenting fundamental graph concepts and models for undirected, directed and signed graphs along with their properties. It continues with a thorough review of graph clustering and community detection techniques. Then, we present how the graph clustering task can be treated in directed graphs, where the presence of directed edges conveys essential semantics. Next, some topics on community evaluation measures are presented for both individual nodes, as well as aggregate metrics. Special mention is made to the degeneracy (k-cores and extensions) approach for community evaluation in directed and signed graphs, presenting also several case studies on real-world

WSDM'13, February 4–8, 2013, Rome, Italy.
ACM 978-1-4503-1869-3/13/02.

networks, such as co-authorship networks, citation networks, trust networks and the Web graph.

Next, we outline the basic points that will be covered in the tutorial.

- Graph fundamentals
  - i. The Erdős-Rényi model
  - ii. Basic graph concepts including clustering, density, locality, connectivity, coherency
- Fundamental techniques of graph clustering and community detection
  - i. A taxonomy of graph clustering algorithms
- Graph partitioning
- Hierarchical graph clustering algorithms
- Partitional clustering
- Spectral graph clustering
- Modularity-based methods
- Modularity optimization
  - i. Greedy techniques
  - ii. Simulated annealing
  - iii. Extremal optimization
  - iv. Spectral optimization
- Dynamic algorithms
- Clustering and community detection in directed graphs
  - i. A taxonomy of clustering algorithms in directed graphs
  - ii. Graph transformations maintaining edge directionality
  - iii. Extending objective functions to directed graphs
  - iv. Alternative approaches: information theoretic based methods, probabilistic models and statistical inference
- Community evaluation measures
  - i. Individual node metrics
  - ii. Aggregate metrics
- Degeneracy based community evaluation
  - i. Degeneracy concept
  - ii. Degeneracy for undirected/directed/signed graphs
  - iii. Case studies: DBLP and arXiv citation graphs; Wikipedia graph; Epinions signed graph
- New directions for research in the area of graph mining for community evaluation

More details about the tutorial can be found in the following url: `http://www.lix.polytechnique.fr/~mvazirg/WSDM2013_tutorial/`.

Acknowledgements

The authors are partially supported by the DIGITEO Chair grant LEVETONE in France. Fragkiskos D. Malliaros is a recipient of the Google Europe Fellowship in Graph Mining, and this research is supported in part by this Google Fellowship.

3. REFERENCES

[1] D. Chakrabarti and C. Faloutsos. Graph mining: Laws, generators, and algorithms. *ACM Comput. Surv.*, 38(1), 2006.

[2] A. Clauset, M. E. J. Newman, and C. Moore. Finding community structure in very large networks. *Physical Review E*, 70(6):066111+, 2004.

[3] M. Coscia, F. Giannotti, and D. Pedreschi. A classification for community discovery methods in complex networks. *Statistical Analysis and Data Mining*, 4(5):512–546, 2011.

[4] L. Danon, A. D. Guilera, J. Duch, and A. Arenas. Comparing community structure identification. *Journal of Statistical Mechanics: Theory and Experiment*, 2005(9):P09008–09008, 2005.

[5] P. Erdős and A. Rényi. On the evolution of random graphs. *Publ. Math. Inst. Hung. Acad. Sci*, 5:17–61, 1960.

[6] S. Fortunato. Community detection in graphs. *Physics Reports*, 486(3-5):75 – 174, 2010.

[7] C. Giatsidis, D. M. Thilikos, and M. Vazirgiannis. D-cores: Measuring collaboration of directed graphs based on degeneracy. In *ICDM*, pages 201–210, 2011.

[8] C. Giatsidis, D. M. Thilikos, and M. Vazirgiannis. Evaluating cooperation in communities with the k-core structure. In *ASONAM*, pages 87–93, 2011.

[9] M. Girvan and M. E. J. Newman. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences*, 99(12):7821–7826, 2002.

[10] 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. *Internet Mathematics*, 6(1):29–123, 2009.

[11] M. E. J. Newman. Modularity and community structure in networks. *Proceedings of the National Academy of Sciences*, 103(23):8577–8582, 2006.

[12] M. E. J. Newman and M. Girvan. Finding and evaluating community structure in networks. *Physical Review E*, 69(2):026113+, 2003.

[13] M. Rosvall and C. T. Bergstrom. Maps of random walks on complex networks reveal community structure. *Proceedings of the National Academy of Sciences*, 105(4):1118–1123, 2008.

[14] V. Satuluri and S. Parthasarathy. Symmetrizations for clustering directed graphs. In *EDBT*, pages 343–354, 2011.

[15] S. Schaeffer. Graph clustering. *Computer Science Review*, 1(1):27–64, 2007.

[16] U. von Luxburg. A tutorial on spectral clustering. *Statistics and Computing*, 17(4):395–416, 2007.