Graph Mining and Graph Kernels

Xifeng Yan and Karsten Borgwardt

March 15, 2008

Abstract

Social and biological networks have led to a huge interest in data analysis on graphs. Various groups within the KDD community have begun to study the task of data mining on graphs, including researchers from database-oriented graph mining, and researchers from kernel machine learning. Their approaches are often complementary, and we feel that exciting research problems and techniques can be discovered by exploring the link between these different approaches to graph mining.

The goal of this tutorial is (i) to introduce newcomers to the field of graph mining, (ii) to introduce people with database background to graph mining using kernel machines, (iii) to introduce people with machine learning background to database-oriented graph mining, and (iv) to present exciting research problems at the interface of both fields.

1 Target audience and prerequisites

Target Audience This tutorial is intended for students, researchers and practitioners from graph mining, system biology, social network analysis, software engineering and related fields. The intended audience of this tutorial includes novice researchers, advanced experts, as well as practitioners from any application domain dealing with graph-structured or network data.

The goal of this tutorial is

- to introduce newcomers to the field of graph mining,
- to introduce people with database background to graph mining using kernel machines,
- to introduce people with machine learning background to database-oriented graph mining,
- to present exciting research problems at the interface of both fields.

In the past five years, a lot of research work on graph mining and graph kernels has been done in parallel in data mining and machine learning societies. So far, there is no single tutorial dedicated to bring these two research areas together. This tutorial presents a comprehensive overview of the techniques developed in graph mining and graph kernels and examines the connection between them. We believe it is timely and in high demand to have such a converged tutorial on this growing theme.

Prerequisites The tutorial assumes familiarity with basic knowledge about efficient algorithms and data structures, as taught in most undergraduate courses in Computer Science.

2 Outline of the tutorial

This tutorial provides a comprehensive and unified view of graph mining, reaching from the field of graph data mining to kernel machine learning on graphs, and illustrating exciting links and intersections between these two branches of graph mining. Part of the materials from our previous tutorial at SIGKDD'06 (`http://www.xifengyan.net/tutorial/kdd06_graph.htm`) will be included, with a significant update of new research results. Sections marked (*) are materials new in this tutorial.

1. Graph Mining from a Pattern Discovery Perspective
    - Frequent graph pattern mining
    - Contrast graph pattern mining*
    - Constrained graph pattern mining
    - Optimal graph pattern mining*
    - Graph mining in single graphs*
    - Graph pattern summarization
    - Graph search*
    - Applications in bioinformatics
    - Applications in software engineering*
2. Graph Mining from a Kernel Machines Perspective
    - Introduction to kernel machine learning*
    - Introduction to graph kernels*
    - The classic random walk graph kernel
    - Speeding up the random walk graph kernel*
    - Shortest-path graph kernel
    - Subtree pattern kernels
    - Cyclic pattern kernels*
    - Graph kernels for chemoinformatics*
    - Scalability of graph kernels*
    - Applications in bioinformatics
3. Bridging the two branches of Graph Mining
    - Frequent subgraph mining via sampling
    - Feature selection on frequent subgraphs
    - Graph classification via kernels and patterns

3 List of related tutorials by the same authors

Xifeng Yan has presented three tutorials on graph mining and management as

1. Mining and Searching Graphs and Structures at SIGKDD'06, with Jiawei Han and Philip S. Yu, 2006 (Tutorial, the 12th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining),
2. Mining, Indexing, and Similarity Search in Graphs and Complex Structures at ICDE'06, with Jiawei Han and Philip S. Yu, 2006 (Tutorial, the 22nd Int. Conf. on Data Engineering),
3. Mining and Searching of Graph-Structured Databases at ICDM'05, with Jiawei Han and Philip S. Yu, 2005 (Invited Tutorial, the Fifth IEEE Int. Conf. on Data Mining).

Difference to our tutorial: Our previous tutorials presented graph mining and managing from a data mining/database point of view. This new tutorial will focus on the convergence of graph pattern mining (data mining) and graph kernels (machine learning). By introducing the state-of-the-art results in both fields, we hope to bring researchers from data mining and machine learning together for studying their connections and shared research issues.

Karsten Borgwardt has presented the material on graph kernels as

1. part of a one-semester course at the Ludwig-Maximilians-University in Munich
2. a guest lecture at the University of Tübingen
3. invited lecture at the European Bioinformatics Institute, Hinxton
4. invited lecture at the University of British Columbia
5. invited lecture at Simon Fraser University
6. invited lecture at the University of Saarbrücken
7. invited lecture at the University of Cambridge, Computer Lab

4 List of related tutorials by other researchers

- **KDD 2004**: Graph Structures in Data Mining by Soumen Chakrabarti and Christos Faloutsos
  Difference to our tutorial: In Soumen and Christos' tutorial, topics related to graphs' global structures are presented, which include graph topology, graph generator, eigenvalue, influence and inference of graphs' topology. Our tutorial examines the graph mining problem through graphs' local structures such as graph patterns, graph features, and graph kernels. Therefore, it will be orthogonal to Soumen and Christos'tutorial.
- **KDD 2007**: Statistical Modeling of Relational Data by Pedro Domingos
  Difference to our tutorial: While relational data can be modelled as graphs, the focus of Prof. Domingo's tutorial are Markov networks and inductive logic programming. In contrast, our tutorial will focus on graph pattern mining and graph kernel computation.

5 Bio and Expertise of authors

Xifeng Yan is a Research Staff Member at IBM T. J. Watson Research Center. He obtained his BE degree in computer engineering from Zhejiang University in 1997 and his PhD degree in computer science from the University of Illinois at Urbana-Champaign in 2006. His current research interests are data mining, bioinformatics, and database systems. He has published more than 40 papers in refereed journals and conferences, including TODS, Bioinformatics, TSE, SIGMOD, SIGKDD, VLDB, and ISMB. His dissertation, "Mining, Indexing and Similarity Search in Large Graph Data Sets" received the 2006 ACM-SIGMOD Best Dissertation Runner-up Award. Xifeng also received the best student paper award at the 2007 International Conference on Data Engineering, the 2007 Pacific-Asia Conference on Knowledge Discovery and Data Mining , and a runner-up award at the 2005 International Conference on Knowledge Discovery and Data Mining.

Karsten Borgwardt has been working on 'mining and learning on graphs' with leading experts both in Data Mining (Prof. Hans-Peter Kriegel, LMU Munich) and Machine Learning (Dr. Alexander Smola, NICTA Australia and Prof. Zoubin Ghahramani, University of Cambridge).

In 2004, he wrote his master thesis on 'Protein function prediction via graph kernels' at the NICTA Statistical Machine Learning Group in Canberra, headed by Dr Alexander Smola. From January 2005 to August 2007, he has been a research and teaching assistant in Prof. Hans-Peter Kriegel's database and data mining group at the Ludwig-Maximilians-University in Munich.

His PhD thesis on 'Graph Kernels' won the Heinz-Schwärtzel-Dissertation Award (best PhD thesis award) by the the three computer science departments in Munich (LMU, TU, UniBw) and has recently been nominated for the German Best PhD thesis Award in Computer Science. Currently, Karsten Borgwardt is a postdoctoral research associate in the Machine Learning Group at the University of Cambridge, UK, headed by Prof. Zoubin Ghahramani, working on machine learning on graphs. He has published 20 research articles on kernel machine learning, in particular on learning on graphs (including 4 NIPS, 3 ISMB and 2 ICML papers).

6 List of top 20 references covered in tutorial

A list of up to 20 most important references that will be covered in the tutorial.

Graph Mining

[1] L. Holder, D. Cook, and S. Djoko. Substructure discovery in the subdue system. In *Proc. AAAI'94 Workshop on Knowledge Discovery in Databases (KDD'94)*, pages 169 – 180, 1994.

[2] J. M. Kleinberg, R. Kumar, P. Raghavan, S. Rajagopalan, and A. Tomkins. The web as a graph: Measurements, models, and methods. In *Proc. Int. Conf. Combinatorics and Computing*, pages 1–17, 1999.

[3] M. Faloutsos, P. Faloutsos, and C. Faloutsos. On power-law relationships of the internet topology. In *Proc. ACM SIGCOMM'99 Conf. Applications, Technologies, Architectures, and Protocols for Computer Communication*, pages 251–262, 1999.

[4] A. Inokuchi, T. Washio, and H. Motoda. An apriori-based algorithm for mining frequent substructures from graph data. In *Proc. 2000 European Symp. Principle of Data Mining and Knowledge Discovery (PKDD'00)*, pages 13–23. Lyon, France, 2000.

[5] M. Kuramochi and G. Karypis. Frequent subgraph discovery. In *Proc. 2001 Int. Conf. on Data Mining (ICDM'01)*, pages 313–320, 2001.

[6] X. Yan and J. Han. gSpan: Graph-based substructure pattern mining. In *Proc. 2002 Int. Conf. on Data Mining (ICDM'02)*, pages 721–724. Maebashi, Japan, 2002.

[7] X. Yan, P. Yu, and J. Han. Graph indexing: A frequent structure-based approach. In *Proc. of 2004 Int. Conf. on Management of Data (SIGMOD'04)*, pages 335 – 346. Paris, France, 2004.

[8] X. Yan, H. Cheng, J. Han, and P. S. Yu. Mining significant graph patterns by scalable leap search. In *Proc. 2008 ACM SIGMOD Int. Conf. on Management of Data (SIGMOD'08)*, 2008.

[9] J. Huan, W. Wang, D. Bandyopadhyay, J. Snoeyink, J. Prins, and A. Tropsha. Mining spatial motifs from protein structure graphs. In *Proc. of the 8th Annual Int. Conf. on Research in Computational Molecular Biology (RECOMB'04)*, pages 308–315, 2004.

[10] M. Koyuturk, A. Grama, and W. Szpankowski. An efficient algorithm for detecting frequent subgraphs in biological networks. In *Proc. of the 12th International Conference on Intelligent Systems for Molecular Biology (ISMB'04)*, pages 200–207, 2004.

[11] M. Acharya, T. Xie, J. Pei, and J. Xu. Mining API patterns as partial orders from source code: From usage scenarios to specifications. In *Proc. 2007 ACM SIGSOFT Symposium on the Foundations of Software Engineering (FSE'07)*, pages 25–34, 2007.

Graph Kernels

[1] T. Gärtner, P. Flach, and S. Wrobel. On graph kernels: Hardness results and efficient alternatives. In B. Schölkopf and M. K. Warmuth, editors, *Proc. Annual Conf. Computational Learning Theory*, pages 129–143. Springer, 2003.

[2] H. Kashima, K. Tsuda, and A. Inokuchi. Marginalized kernels between labeled graphs. In *Proc. Intl. Conf. Machine Learning*, pages 321–328. Morgan Kaufmann, San Francisco, CA, 2003.

[3] S. V. N. Vishwanathan, K. Borgwardt, and N. N. Schraudolph. Fast computation of graph kernels. In B. Schölkopf, J. Platt, and T. Hofmann, editors, *Advances in Neural Information Processing Systems 19*. MIT Press, Cambridge MA, 2007.

[4] K. M. Borgwardt and H.-P. Kriegel. Shortest-path kernels on graphs. In *Proc. Intl. Conf. Data Mining*, pages 74–81, 2005.

[5] J. Ramon and T. Gärtner. Expressivity versus efficiency of graph kernels. Technical Report, First International Workshop on Mining Graphs, Trees and Sequences (held with ECML/PKDD'03), 2003.

[6] T. Horvath, T. Gärtner, and S. Wrobel. Cyclic pattern kernels for predictive graph mining. In *Proceedings of the International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 158–167, 2004.

[7] P. Mahé, N. Ueda, T. Akutsu, J.-L. Perret, and J.-P. Vert. Extensions of marginalized graph kernels. In *Proceedings of the Twenty-First International Conference on Machine Learning*, pages 552–559, 2004.

[8] L. Ralaivola, S. J. Swamidass, H. Saigo, and P. Baldi. Graph kernels for chemical informatics. *Neural Networks*, 18(8):1093–1110, 2005.

[9] K. M. Borgwardt, C. S. Ong, S. Schönauer, S. V. N. Vishwanathan, A. J. Smola, and H. P. Kriegel. Protein function prediction via graph kernels. *Bioinformatics (and Intelligent Systems in Molecular Biology, ISMB 2005)*, 21(Suppl 1):i47–i56, 2005.

7 Links to presentation material

`http://www.xifengyan.net/tutorial/kdd08_graph.htm`