

Data Mining: Concepts and Techniques (2nd edition)

Jiawei Han and Micheline Kamber
Morgan Kaufmann Publishers, 2006

Bibliographic Notes for Chapter 9 Graph Mining, Social Network Analysis, and Multirelational Data Mining

Research into graph mining has developed many frequent subgraph mining methods. Washio and Motoda [WM03] performed a survey on graph-based data mining. Many well-known pair-wise isomorphism testing algorithms were developed, such as Ullmann's Backtracking [Ull76] and McKay's Nauty [McK81]. Dehaspe, Toivonen and King [DTK98] applied inductive logic programming to predict chemical carcinogenicity by mining frequent substructures. Several Apriori-based frequent substructure mining algorithm have been proposed, including AGM by Inokuchi, Washio, and Motoda [IWM98], FSG by Kuramochi and Karypis [KK01], and an edge-disjoint path-join algorithm by Vanetik, Gudes, and Shimony [VGS02]. Pattern-growth-based graph pattern mining algorithms include gSpan by Yan and Han [YH02], MoFa by Borgelt and Berthold [BB02], FFMSM and SPIN by Huan, Wang, and Prins [HWP03] and Prins, Yang, Huan, and Wang [PYHW04], respectively, and Gaston by Nijssen and Kok [NK04]. These algorithms are inspired by PrefixSpan [PHMA⁺01] for mining sequences, and TreeMinerV [Zak02] and FREQT [AAK⁺02] for mining trees. A disk-based frequent graph mining method was proposed by Wang, Wang, Pei, et al. [WWP⁺04].

Mining closed graph patterns was studied by Yan and Han [YH03] with the proposal of the algorithm, CloseGraph, as an extension of gSpan and CloSpan [YHA03]. Holder, Cook and Djoko [HCD94] proposed SUBDUE for approximate substructure pattern discovery based on minimum description length and background knowledge. Mining coherent subgraphs was studied by Huan, Wang, Bandyopadhyay, et al. [HWB⁺04]. For mining relational graphs, Yan, Zhou and Han [YZH05] proposed two algorithms, CloseCut and Splat, to discover exact dense frequent substructures in a set of relational graphs.

There have been many studies that explore the applications of mined graph patterns. Path-based graph indexing approaches are used in GraphGrep, developed by Shasha, Wang, and Giugno [SWG02], and in Daylight, developed by James, Weininger, and Delany [JWD03]. Frequent graph patterns were used as graph indexing features in the gIndex and Grafil methods proposed by Yan, Yu and Han [YYH04, YYH05] to perform fast graph search and structure similarity search. Borgelt and Berthold [BB02] illustrated the discovery of active chemical structures in an HIV-screening dataset by contrasting the support of frequent graphs between different classes. Deshpande, Kuramochi and Karypis [DKK02] used frequent structures as features to classify chemical compounds. Huan, Wang, Bandyopadhyay, et al. [HWB⁺04] successfully applied the frequent graph mining technique to study protein structural families. Koyuturk, Grama, and Szpankowski [KGS04] proposed a method to detect frequent subgraphs in biological networks. Hu, Yan, Yu, et al. [HYY⁺05] developed an algorithm called CoDense to find dense subgraphs across multiple biological networks.

There has been a great deal of research on social networks. For texts on social network analysis, see Wasserman and Faust [WF94], Degenne and Forse [DF99], Scott [Sco05], Watts [Wat03a], Barabasi [Bar03], and Carrington, Scott, and Wasserman [CSW05]. For a survey of work on social network analysis, see Newman [New03]. Barabasi, Oltvai, Jeong, Albert, et al. have several comprehensive tutorials on the topic, available at www.nd.edu/~networks/publications.htm#talks0001. Books on small world networks include Watts [Wat03b] and Buchanan [Buc03]. Milgram's "six degrees of separation" experiment is presented in [Mil67].

The *Forest Fire model* for network generation was proposed in Leskovec, Kleinberg, and Faloutsos [LKF05]. The *preferential attachment model* was studied in Albert and Barabasi [AB99] and Cooper and Frieze [CF03]. The *copying model* was explored in Kleinberg, Kumar, Raghavan, et al. [KKR⁺99] and Kumar, Raghavan, Rajagopalan, et al. [KRR⁺00].

Link mining tasks and challenges were overviewed by Getoor [Get03]. A link-based classification method was proposed in Lu and Getoor [LG03]. Iterative classification and inference algorithms have been proposed for hypertext classification by Chakrabarti, Dom, and Indyk [CDI98] and Oh, Myaeng, and Lee [OML00]. Bhattacharya and Getoor [BG04] propose a method for clustering linked data, which can be used to solve the data mining tasks of entity deduplication and group discovery. A method for group discovery was proposed by Kubica, Moore, and Schneider [KMS03]. Approaches to link prediction, based on measures for analyzing the “proximity” of nodes in a network, are described in Liben-Nowell and Kleinberg [LNK03]. The Katz measure was presented in Katz [Kat53]. A probabilistic model for learning link structure is given in Getoor, Friedman, Koller, and Taskar [GFKT01]. Link prediction for counterterrorism was proposed by Krebs [Kre02]. Viral marketing was described by Domingos [Dom05] and his work with Richardson [DR01, RD02]. BLOG (Bayesian LOGic), a language for reasoning with unknown objects, was proposed by Milch, Marthi, Russell, et al. [MMR⁺05] to address the closed world assumption problem. Mining newsgroups to partition discussion participants into opposite camps using quotation networks was proposed by Agrawal, Rajagopalan, Srikant, and Xu [ARSX04]. The relation selection and extraction approach to community mining from multirelational networks was described in Cai, Shao, He, et al. [CSH⁺05].

Multirelational data mining has been investigated extensively in the Inductive Logic Programming (ILP) community. Lavrac and Dzeroski [LD94] and Muggleton [Mug95] provide comprehensive introductions to Inductive Logic Programming (ILP). An overview of multirelational data mining was given by Dzeroski [Dze03]. Well-known ILP systems include FOIL by Quinlan and Cameron-Jones [QCJ93], Golem by Muggleton and Feng [MF90], and Progol by Muggleton [Mug95]. More recent systems include TILDE by Blockeel, De Raedt, and Ramon [BRR98], Mr-SMOTI by Appice, Ceci, and Malerba [ACM03], and RPTs by Neville, Jensen, Friedland and Hay [NJFH03], which inductively constructs decision trees from relational data. Probabilistic approaches to multirelational classification include probabilistic relational models by Getoor, Friedman, Koller, and Taskar [GFKT01] and by Taskar, Segal, and Koller [TSK01]. Popescul, Ungar, Lawrence, and Pennock [PULP03] propose an approach to integrate ILP and statistical modelling for document classification and retrieval. The CrossMine approach is described in Yin, Han, Yang, and Yu [YHYY04]. The look-one-ahead method used in CrossMine was developed by Blockeel, De Raedt, and Ramon [BRR98]. Multirelational clustering was explored by Gartner, Lloyd, and Flach [GLF04], and Kirsten and Wrobel [KW98, KW00]. CrossClus performs multirelational clustering with user guidance by Yin, Han, and Yu [YHY05].

Bibliography

- [AAK⁺02] T. Asai, K. Abe, S. Kawasoe, H. Arimura, H. Satamoto, and S. Arikawa. Efficient substructure discovery from large semi-structured data. In *Proc. 2002 SIAM Int. Conf. Data Mining (SDM'02)*, pages 158–174, Arlington, VA, April 2002.
- [AB99] R. Albert and A.-L. Barabasi. Emergence of scaling in random networks. *Science*, 286:509–512, 1999.
- [ACM03] A. Appice, M. Ceci, and D. Malerba. Mining model trees: A multi-relational approach. In *Proc. 2003 Int. Conf. Inductive Logic Programming (ILP'03)*, pages 4–21, Szeged, Hungary, Sept. 2003.
- [ARSX04] R. Agrawal, S. Rajagopalan, R. Srikant, and Y. Xu. Mining newsgroups using networks arising from social behavior. In *Proc. 2003 Int. World Wide Web Conf. (WWW'03)*, pages 529–535, New York, NY, May 2004.
- [Bar03] A.-L. Barabasi. *Linked: How Everything Is Connected to Everything Else and What It Means*. Plume, 2003.
- [BB02] C. Borgelt and M. R. Berthold. Mining molecular fragments: Finding relevant substructures of molecules. In *Proc. 2002 Int. Conf. Data Mining (ICDM'02)*, pages 211–218, Maebashi, Japan, Dec. 2002.
- [BG04] I. Bhattacharya and L. Getoor. Iterative record linkage for cleaning and integration. In *Proc. SIGMOD 2004 Workshop on Research Issues on Data Mining and Knowledge Discovery (DMKD'04)*, pages 11–18, Paris, France, June 2004.
- [BRR98] H. Blockeel, L. De Raedt, and J. Ramon. Top-down induction of logical decision trees. In *Proc. 1998 Int. Conf. Machine Learning (ICML'98)*, pages 55–63, Madison, WI, Aug. 1998.
- [Buc03] M. Buchanan. *Nexus: Small Worlds and the Groundbreaking Theory of Networks*. W. W. Norton & Company, 2003.
- [CDI98] S. Chakrabarti, B. E. Dom, and P. Indyk. Enhanced hypertext classification using hyper-links. In *Proc. 1998 ACM-SIGMOD Int. Conf. Management of Data (SIGMOD'98)*, pages 307–318, Seattle, WA, June 1998.
- [CF03] C. Cooper and A. Frieze. A general model of web graphs. *Algorithms*, 22:311–335, 2003.
- [CSH⁺05] D. Cai, Z. Shao, X. He, X. Yan, and J. Han. Community mining from multi-relational networks. In *Proc. 2005 European Conf. Principles and Practice of Knowledge Discovery in Databases (PKDD'05)*, pages 445–452, Porto, Portugal, Oct. 2005.
- [CSW05] P. J. Carrington, J. Scott, and S. Wasserman. *Models and methods in social network analysis*. Cambridge University Press, 2005.
- [DF99] A. Degenne and M. Forse. *Introducing Social Networks*. Sage Publications, 1999.

- [DKK02] M. Deshpande, M. Kuramochi, and G. Karypis. Automated approaches for classifying structures. In *Proc. 2002 Workshop on Data Mining in Bioinformatics (BIOKDD'02)*, pages 11–18, Edmonton, Canada, July 2002.
- [Dom05] P. Domingos. Mining social networks for viral marketing. *IEEE Intelligent Systems*, 20:80–82, 2005.
- [DR01] P. Domingos and M. Richardson. Mining the network value of customers. In *Proc. 2001 ACM SIGKDD Int. Conf. Knowledge Discovery in Databases (KDD'01)*, pages 57–66, San Francisco, CA, Aug. 2001.
- [DTK98] L. Dehaspe, H. Toivonen, and R. King. Finding frequent substructures in chemical compounds. In *Proc. 1998 Int. Conf. Knowledge Discovery and Data Mining (KDD'98)*, pages 30–36, New York, NY, Aug. 1998.
- [Dze03] S. Dzeroski. Multirelational data mining: An introduction. *ACM SIGKDD Explorations*, 5:1–16, July 2003.
- [Get03] L. Getoor. Link mining: a new data mining challenge. *SIGKDD Explorations*, 5:84–89, 2003.
- [GFKT01] L. Getoor, N. Friedman, D. Koller, and B. Taskar. Learning probabilistic models of relational structure. In *Proc. 2001 Int. Conf. Machine Learning (ICML'01)*, pages 170–177, Williamstown, MA, 2001.
- [GLF04] T. Garner, J. W. Lloyd, and P. A. Flach. Kernels and distances for structured data. *Machine Learning*, 57:205–232, 2004.
- [HCD94] L. B. Holder, D. J. Cook, and S. Djoko. Substructure discovery in the subdue system. In *Proc. AAAI'94 Workshop Knowledge Discovery in Databases (KDD'94)*, pages 169–180, Seattle, WA, July 1994.
- [HWB⁺04] J. Huan, W. Wang, D. Bandyopadhyay, J. Snoeyink, J. Prins, and A. Tropsha. Mining spatial motifs from protein structure graphs. In *Proc. 8th Int. Conf. Research in Computational Molecular Biology (RECOMB)*, pages 308–315, San Diego, CA, March 2004.
- [HWP03] J. Huan, W. Wang, and J. Prins. Efficient mining of frequent subgraph in the presence of isomorphism. In *Proc. 2003 Int. Conf. Data Mining (ICDM'03)*, pages 549–552, Melbourne, FL, Nov. 2003.
- [HYY⁺05] H. Hu, X. Yan, H. Yu, J. Han, and X. J. Zhou. Mining coherent dense subgraphs across massive biological networks for functional discovery. In *Proc. 2005 Int. Conf. Intelligent Systems for Molecular Biology (ISMB'05)*, pages 213–221, Ann Arbor, MI, June 2005.
- [IWM98] 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, Sept. 1998.
- [JWD03] C. A. James, D. Weininger, and J. Delany. *Daylight Theory Manual Daylight Version 4.82*. Daylight Chemical Information Systems, Inc., 2003.
- [Kat53] L. Katz. A new status index derived from sociometric analysis. *Psychometrika*, 18:39–43, March 1953.
- [KGS04] M. Koyuturk, A. Grama, and W. Szpankowski. An efficient algorithm for detecting frequent subgraphs in biological networks. *Bioinformatics*, 20:I200–I207, 2004.
- [KK01] M. Kuramochi and G. Karypis. Frequent subgraph discovery. In *Proc. 2001 Int. Conf. Data Mining (ICDM'01)*, pages 313–320, San Jose, CA, Nov. 2001.
- [KKR⁺99] 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. Computing and Combinatorics (COCOON'99)*, pages 1–17, Tokyo, Japan, July 1999.

- [KMS03] J. Kubica, A. Moore, and J. Schneider. Tractable group detection on large link data sets. In *Proc. 2003 Int. Conf. Data Mining (ICDM'03)*, pages 573–576, Melbourne, FL, Nov. 2003.
- [Kre02] V. Krebs. Mapping networks of terrorist cells. *Connections*, 24:43–52, Winter 2002.
- [KRR⁺00] R. Kumar, P. Raghavan, S. Rajagopalan, D. Sivakumar, A. Tomkins, and E. Upfal. Stochastic models for the web graph. In *Proc. 2000 IEEE Symp. Foundations of Computer Science (FOCS'00)*, pages 57–65, Redondo Beach, CA, Nov. 2000.
- [KW98] M. Kirsten and S. Wrobel. Relational distance-based clustering. In *Proc. 1998 Int. Conf. Inductive Logic Programming (ILP'98)*, pages 261–270, Madison, WI, 1998.
- [KW00] M. Kirsten and S. Wrobel. Extending k-means clustering to first-order representations. In *Proc. 2000 Int. Conf. Inductive Logic Programming (ILP'00)*, pages 112–129, London, UK, July 2000.
- [LD94] N. Lavrac and S. Dzeroski. *Inductive Logic Programming: Techniques and Applications*. Ellis Horwood, 1994.
- [LG03] Q. Lu and L. Getoor. Link-based classification. In *Proc. 2003 Int. Conf. Machine Learning (ICML'03)*, pages 496–503, Washington, DC, Aug. 2003.
- [LKF05] J. Leskovec, J. Kleinberg, and C. Faloutsos. Graphs over time: Densification laws, shrinking diameters and possible explanations. In *Proc. 2005 ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (KDD'05)*, pages 177–187, Chicago, IL, Aug. 2005.
- [LNK03] D. Liben-Nowell and J. Kleinberg. The link prediction problem for social networks. In *Proc. 2003 Int. Conf. Information and Knowledge Management (CIKM'03)*, pages 556–559, New Orleans, LA, Nov. 2003.
- [McK81] B. D. McKay. Practical graph isomorphism. *Congressus Numerantium*, 30:45–87, 1981.
- [MF90] S. Muggleton and C. Feng. Efficient induction of logic programs. In *Proc. 1990 Conf. Algorithmic Learning Theory (ALT'90)*, pages 368–381, Tokyo, Japan, Oct. 1990.
- [Mil67] S. Milgram. The small world problem. *Psychology Today*, 2:60–67, 1967.
- [MMR⁺05] B. Milch, B. Marthi, S. Russell, D. Sontag, D. L. Ong, and A. Kolobov. BLOG: Probabilistic models with unknown objects. In *Proc. 19th Int. Joint Conf. on Artificial Intelligence (IJCAI'05)*, pages 1352–1359, Aug. 2005.
- [Mug95] S. Muggleton. Inverse entailment and prolog. *New Generation Computing, Special issue on Inductive Logic Programming*, 3:245–286, 1995.
- [New03] M. E. J. Newman. The structure and function of complex networks. *SIAM Review*, 45:167–256, 2003.
- [NJFH03] J. Neville, D. Jensen, L. Friedland, and M. Hay. Learning relational probability trees. In *Proc. 2003 ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining (KDD'03)*, pages 625–630, Washington, DC, Aug 2003.
- [NK04] S. Nijssen and J. Kok. A quickstart in frequent structure mining can make a difference. In *Proc. 2004 ACM SIGKDD Int. Conf. Knowledge Discovery in Databases (KDD'04)*, pages 647–652, Seattle, WA, Aug. 2004.
- [OML00] H.-J. Oh, S. H. Myaeng, and M.-H. Lee. A practical hypertext categorization method using links and incrementally available class information. In *Proc. Int. 2000 ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR'00)*, pages 264–271, Athens, Greece, July 2000.
- [PHMA⁺01] J. Pei, J. Han, B. Mortazavi-Asl, H. Pinto, Q. Chen, U. Dayal, and M.-C. Hsu. PrefixSpan: Mining sequential patterns efficiently by prefix-projected pattern growth. In *Proc. 2001 Int. Conf. Data Engineering (ICDE'01)*, pages 215–224, Heidelberg, Germany, April 2001.

- [PULP03] A. Popescul, L. Ungar, S. Lawrence, and M. Pennock. Statistical relational learning for document mining. In *Proc. 2003 Int. Conf. Data Mining (ICDM'03)*, pages 275–282, Melbourne, FL, Nov. 2003.
- [PYHW04] J. Prins, J. Yang, J. Huan, and W. Wang. Spin: Mining maximal frequent subgraphs from graph databases. In *Proc. 2004 ACM SIGKDD Int. Conf. Knowledge Discovery in Databases (KDD'04)*, pages 581–586, Seattle, WA, Aug. 2004.
- [QCJ93] J. R. Quinlan and R. M. Cameron-Jones. FOIL: A midterm report. In *Proc. 1993 European Conf. Machine Learning*, pages 3–20, Vienna, Austria, 1993.
- [RD02] M. Richardson and P. Domingos. Mining knowledge-sharing sites for viral marketing. In *Proc. 2002 ACM SIGKDD Int. Conf. Knowledge Discovery in Databases (KDD'02)*, pages 61–70, Edmonton, Canada, July 2002.
- [Sco05] J. P. Scott. *Social Network Analysis: A Handbook*. Sage Publications, 2005.
- [SWG02] D. Shasha, J. T.-L. Wang, and R. Giugno. Algorithmics and applications of tree and graph searching. In *Proc. 2002 ACM Symp. Principles of Database Systems (PODS'02)*, pages 39–52, Madison, WI, June 2002.
- [TSK01] B. Taskar, E. Segal, and D. Koller. Probabilistic classification and clustering in relational data. In *Proc. 2001 Int. Joint Conf. Artificial Intelligence (IJCAI'01)*, pages 870–878, Seattle, WA, 2001.
- [Ull76] J. R. Ullmann. An algorithm for subgraph isomorphism. *J. ACM*, 23:31–42, 1976.
- [VGS02] N. Vanetik, E. Gudes, and S. E. Shimony. Computing frequent graph patterns from semistructured data. In *Proc. 2002 Int. Conf. on Data Mining (ICDM'02)*, pages 458–465, Maebashi, Japan, Dec. 2002.
- [Wat03a] D. J. Watts. *Six Degrees: The Science of a Connected Age*. W. W. Norton & Company, 2003.
- [Wat03b] D. J. Watts. *Small Worlds: The Dynamics of Networks between Order and Randomness*. Princeton University Press, 2003.
- [WF94] S. Wasserman and K. Faust. *Social Network Analysis: Methods and Applications*. Cambridge University Press, 1994.
- [WM03] T. Washio and H. Motoda. State of the art of graph-based data mining. *SIGKDD Explorations*, 5:59–68, 2003.
- [WWP⁺04] C. Wang, W. Wang, J. Pei, Y. Zhu, and B. Shi. Scalable mining of large disk-base graph databases. In *Proc. 2004 ACM SIGKDD Int. Conf. Knowledge Discovery in Databases (KDD'04)*, pages 316–325, Seattle, WA, Aug. 2004.
- [YH02] X. Yan and J. Han. gSpan: Graph-based substructure pattern mining. In *Proc. 2002 Int. Conf. Data Mining (ICDM'02)*, pages 721–724, Maebashi, Japan, Dec. 2002.
- [YH03] X. Yan and J. Han. CloseGraph: Mining closed frequent graph patterns. In *Proc. 2003 ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining (KDD'03)*, pages 286–295, Washington, DC, Aug. 2003.
- [YHA03] X. Yan, J. Han, and R. Afshar. CloSpan: Mining closed sequential patterns in large datasets. In *Proc. 2003 SIAM Int. Conf. Data Mining (SDM'03)*, pages 166–177, San Fransisco, CA, May 2003.
- [YHY05] X. Yin, J. Han, and P.S. Yu. Cross-relational clustering with user's guidance. In *Proc. 2005 ACM SIGKDD Int. Conf. Knowledge Discovery in Databases (KDD'05)*, pages 344–353, Chicago, IL, Aug. 2005.
- [YHYY04] X. Yin, J. Han, J. Yang, and P. S. Yu. CrossMine: Efficient classification across multiple database relations. In *Proc. 2004 Int. Conf. Data Engineering (ICDE'04)*, pages 399–410, Boston, MA, Mar. 2004.

- [YYH04] X. Yan, P. S. Yu, and J. Han. Graph indexing: A frequent structure-based approach. In *Proc. 2004 ACM-SIGMOD Int. Conf. Management of Data (SIGMOD'04)*, pages 335–346, Paris, France, June 2004.
- [YYH05] X. Yan, P. S. Yu, and J. Han. Substructure similarity search in graph databases. In *Proc. 2005 ACM-SIGMOD Int. Conf. Management of Data (SIGMOD'05)*, pages 766–777, Baltimore, MD, June 2005.
- [YZH05] X. Yan, X. J. Zhou, and J. Han. Mining closed relational graphs with connectivity constraints. In *Proc. 2005 ACM SIGKDD Int. Conf. Knowledge Discovery in Databases (KDD'05)*, pages 324–333, Chicago, IL, Aug. 2005.
- [Zak02] M. J. Zaki. Efficiently mining frequent trees in a forest. In *Proc. 2002 ACM SIGKDD Int. Conf. Knowledge Discovery in Databases (KDD'02)*, pages 71–80, Edmonton, Canada, July 2002.