Extending Classical Enumeration Techniques for Graph Mining

1 General Informations

Phd advisors.

- Amedeo Napoli (supervisor)
- Chedy Raïssi (co-supervisor)

Address	LORIA, Campus Scientifique - BP 239, 54506 Vandœuvre-lès-Nancy
Phone	+33-383-5920-91/79
Email	amedeo.napoli@loria.fr,chedy.raissi@inria.fr
Office	B 140/162

2 Context and Motivations

Many datasets of interest today are best described as graphs, as a result, the field of graph mining has seen a rapid growth in recent years because of new applications in chemistry, computational biology, and social and communication networking [AW10, PN09]. Graph mining is actively and extensively studied by the theoretical graph community in the context of numerous problems such as graph partitioning, vertices and edges coloring, node clustering, matching, and connectivity analysis [dCRTB05]. However the traditional work in the theoretical graph community cannot be directly used in practice. In real applications, the problem often differ from theoretical definition and may have different variations such as a multi-graph representation, or incomplete informations on vertices or edges. These graphs may represent homogeneous networks, in which there is a single vertex type and link type, or richer, heterogeneous networks, in which there may be multiple object and link types. Another limitation is that real datasets are incredibly huge in size and complexity (think for instance about the Facebook graph). In such a case, graphs may not be stored in main memory and the output of mining algorithms may be available only on disk (or distributed storages such as clouds). As a result, the main challenges, for the graph mining community, is to design specialized algorithms which are sensitive to disk access constraints but also to improve the runtime performances and provide concise and complete summarization mechanisms for the mining results.

As in the case of other data types such as sequential or textual data, one can design mining approaches for graph data. This includes techniques such as frequent pattern mining, clustering and classification [AW10, AS94]. However, and this is the main motivation for this thesis, these methods are much more challenging in the graph domain, because the underlying structure of the data makes the interpretability of the mining results much more complex.

3 Subject

The subject of the thesis is about extending classical enumeration techniques for graph mining over different novel research axes. The candidate will tackle the problem of *"inter-*

estingness" of a pattern by means of statistical models. As a starting point, the general idea of uniform sampling of graphlets [RBH12] is to be studied in relation with large graphs. Another axis is to focus on spectral graph theory [Chu97] to discover strong bridges between pattern enumerations (such as pseudo-clique and triangle extraction) and the eigenvalues and eigenvectors of the graph adjacency matrix (or Laplacian). This type of enumeration is usually expensive to process unless accurate and efficient design is taken into account. This opens up the possibility for the candidate to develop solutions that are able to run on highly parallel and scalable frameworks such as the MAPREDUCE (HADOOP) environment [KMF11]. A final axis of research is the extraction or enumeration of graph patterns based on some *"preference constraints"* expressed by the application experts such as Pareto optimality, Choquet Integrals or even as a constraint satisfaction problem [SRPC11, RPK10, GLP13, UBLC12]. This novel type of extraction paradigm is already applied to set patterns and sequences, however no work focused on this particular and challenging problem for graph data.

The candidate who will work on this thesis subject will take advantage of the experience of the Orpailleur research team in knowledge discovery guided by domain knowledge. This thesis subject is theoretical but also practical. The different approaches developed by the candidate will be thoroughly tested in real applications that are linked to ongoing projects like the BioIntelligence project (intelligent systems for biology), the Hybride project (Data mining for the study of orphan diseases), and the PoQeMON project (pattern mining for network telemetry analysis).

References

[AS94] Rakesh Agrawal and Ramakrishnan Srikant. Fast algorithms for mining association rules in large databases. In Jorge B. Bocca, Matthias Jarke, and Carlo Zaniolo, editors, *VLDB*, pages 487–499. Morgan Kaufmann, 1994.

[AW10] Charu C. Aggarwal and Haixun Wang, editors. *Managing and Mining Graph Data*, volume 40 of *Advances in Database Systems*. Springer, 2010.

[Chu97] Fan R. K. Chung. *Spectral Graph Theory.* Regional Conference Series in Mathematics. 92. Providence, RI: American Mathematical Society (AMS). xi, 1997.

[dCRTB05] Luciano daF. Costa, Francisco A. Rodrigues, Gonzalo Travieso, and P. R. Villas Boas. Characterization of complex networks: A survey of measurements. *Advances in Physics*, 56(1):167–242, January 2005.

[GLP13] Lucie Galand, Julien Lesca, and Patrice Perny. Dominance rules for the choquet integral in multiobjective dynamic programming. In Francesca Rossi, editor, *IJCAI*. IJCAI/AAAI, 2013.

[KMF11] U. Kang, Brendan Meeder, and Christos Faloutsos. Spectral analysis for billion-scale graphs: Discoveries and implementation. In Joshua Zhexue Huang, Longbing Cao, and Jaideep Srivastava, editors, *PAKDD (2)*, volume 6635 of *Lecture Notes in Computer Science*, pages 13–25. Springer, 2011.

[PN09] Frédéric Pennerath and Amedeo Napoli. The model of most informative patterns and its application to knowledge extraction from graph databases. In Wray L. Buntine, Marko Grobelnik, Dunja Mladenic, and John Shawe-Taylor,

editors, *ECML/PKDD (2)*, volume 5782 of *Lecture Notes in Computer Science*, pages 205–220. Springer, 2009.

[RBH12] Mahmudur Rahman, Mansurul Bhuiyan, and Mohammad Al Hasan. Graft: an approximate graphlet counting algorithm for large graph analysis. In Xue wen Chen, Guy Lebanon, Haixun Wang, and Mohammed J. Zaki, editors, *CIKM*, pages 1467–1471. ACM, 2012.

[RPK10] Chedy Raïssi, Jian Pei, and Thomas Kister. Computing closed skycubes. *PVLDB*, 3(1):838–847, 2010.

[SRPC11] Arnaud Soulet, Chedy Raïssi, Marc Plantevit, and Bruno Crémilleux. Mining dominant patterns in the sky. In Diane J. Cook, Jian Pei, Wei Wang, Osmar R. Zaïane, and Xindong Wu, editors, *ICDM*, pages 655–664. IEEE, 2011.

[UBLC12] Willy Ugarte, Patrice Boizumault, Samir Loudni, and Bruno Crémilleux. Soft threshold constraints for pattern mining. In Jean-Gabriel Ganascia, Philippe Lenca, and Jean-Marc Petit, editors, *Discovery Science*, volume 7569 of *Lecture Notes in Computer Science*, pages 313–327. Springer, 2012.