Mining Large Networks under Homomorphism

Mostafa Haghir Chehreghani, Jan Ramon, Thomas Fannes
Department of Computer Science, KU Leuven, Celestijnenlaan 200A, 3001 Leuven, Belgium
{Mostafa.HaghirChehreghani, Jan.Ramon, Thomas.Fannes}@cs.kuleuven.be

The problem of finding frequent patterns from a database of graphs or from a single network has several important applications in different areas like bioinformatics, molecular interaction networks, web mining and social and information network analysis. It is also a fundamental problem in many other data mining tasks such as association rule mining, classification and clustering. In this work, we focus on the single network setting.

Existing algorithms for finding frequent patterns from a single large network mainly use *subgraph isomorphism*. However, subgraph isomorphism is expensive to compute: deciding whether a graph P is subgraph isomorphic to a graph G is NP-complete. Therefore, in massive networks consisting of millions of vertices subgraph isomorphism is intractable outside a very restricted class of patterns. It is also known that listing closed patterns under isomorphism is intractable even for simple classes of graphs like trees [2]. For these reasons, we argue that it might be interesting to use the subgraph homomorphism matching operator, since in this setting for patterns of bounded treewidth, the complexity of the matching operator only depends polynomially on the size of the network.

Only a few algorithms have been proposed for finding frequent patterns under subgraph homomorphism. Dries and Nijssen [1] presented the first algorithm that finds frequent rooted trees under homomorphism. They developed an efficient method for generating non-homomorphic tree patterns based on a canonical string representation. However, under homomorphism, the process of generating patterns is more difficult. In particular, a pattern may be homomorphic to a smaller pattern. This makes the ordered search more complicated than the case of subgraph isomorphism. Moreover, a pattern P might be generated from a smaller pattern P' by adding an arbitrary number of edges, rather than a fixed number of edges. This makes traversing the search space nontrivial.

In our work, we aim to go beyond tree patterns and address the aforementioned problems for graph patterns. In particular: 1) we introduce (redefine) a new class of graphs, called *rooted graphs*, and motivate and formulate the problem of finding rooted graph patterns. We discuss how the proposed mining task can tackle some existing problems in single network mining. 2) we propose a new method for generating rooted graph patterns and show its completeness. 3) we introduce a new notion for compactly representing all frequent patterns. Compact representations are useful since the number of all frequent patterns is usually huge. We propose a new closure operator for patterns which are frequent under homomorphism and investigate its properties.

Acknowledgements This work was supported by ERC Starting Grant 240186 "MiGraNT: Mining Graphs and Networks: a Theory-based approach".

References

[1] Anton Dries and Siegfried Nijssen, Mining patterns in networks using homomorphism, SDM, 260-271, 2012.

[2] Jan Ramon and Mostafa Haghir Chehreghani, On the complexity of listing closed frequent subgraph patterns, Proceedings of the 8th French Conference on Combinatorics (8FCC), Paris, France, 2010.