Frequent Subgraph Mining in Outerplanar Graphs

Tamás Horváth ^a

Jan Ramon ^b

Stefan Wrobel ^a

^a Dept. of Computer Science III, University of Bonn, Germany
Fraunhofer AIS, Sankt Augustin, Germany

^b Dept. of Computer Science, Katholieke Universiteit Leuven, Belgium

Abstract

In recent years there has been an increased interest in frequent pattern discovery in large databases of graph structured objects. While the frequent connected subgraph mining problem for tree datasets can be solved in incremental polynomial time, it becomes intractable for arbitrary graph databases. Existing approaches have therefore resorted to various heuristic strategies and restrictions of the search space, but have not identified a practically relevant tractable graph class beyond trees.

In this paper, we define the class of so called *tenuous outerplanar graphs*, a strict generalization of trees, develop a frequent subgraph mining algorithm for tenuous outerplanar graphs that works in incremental polynomial time, and evaluate the algorithm empirically on the NCI molecular graph dataset.

The discovery of *frequent patterns* in a database is one of the central tasks considered in data mining. In addition to be interesting in their own right, frequent patterns can also be used as features for predictive data mining tasks (see, e.g., [2]). For a long time, work on frequent pattern discovery has concentrated on relatively simple notions of patterns and elements in the database as they are typically used for the discovery of association rules (simple sets of atomic items). In recent years, however, due to the significance of application areas such as the analysis of chemical molecules or graph structures in the WWW, there has been an increased interest in algorithms that can perform frequent pattern discovery in databases of structured objects such as *trees* or arbitrary *graphs*.

While the frequent pattern problem for trees can be solved in *incremental polynomial time*, i.e., in time polynomial in the *combined size* of the input and the set of frequent tree patterns *so far* computed, the frequent pattern problem for graph structured databases in the general case cannot be solved in *output polynomial time*, i.e., in time polynomial in the *combined size* of the input and the set of *all* frequent patterns. Indeed, this can easily be seen by considering a database consisting of two examples: one simple cycle of length n with identical vertex labels and one arbitrary graph of size n . One can then observe that the size of the output is at most $n + 1$ (n paths of length $0 \dots n - 1$ and one cycle), while the problem of deciding whether the cycle is a subgraph of (is isomorphic to a hamiltonian cycle in) the other example is NP-hard.

Existing approaches to frequent pattern discovery for graphs have therefore resorted to various heuristic strategies and restrictions of the search space (see, e.g., [1, 2, 4, 7]), but have not identified a practically relevant tractable graph class beyond trees.

In [3], we define the class of so called *tenuous outerplanar graphs*, which is the class of planar graphs that can be embedded in the plane in such a way that all of its vertices lie on the outer boundary, i.e. can be reached from the outside without crossing any edges, and which have a fixed limit on the number of inside diagonal edges (k -tenuous means at most k diagonals). This class of graphs is a strict generalization of trees, and is motivated by the kinds of graphs actually found in practical applications. In fact, in one of the popular graph mining data sets (the NCI data set), 94.3% of all elements are outerplanar graphs and 99.99% of these are 5-tenuous. At the same time, in the paper, we develop an algorithm for enumerating frequent tenuous outerplanar graph patterns which is guaranteed to work in incremental polynomial time.

Our approach is based on a canonical string representation of outerplanar graphs which may be of interest in itself, and further algorithmic components for mining frequent biconnected outerplanar graphs and candidate generation in an Apriori style algorithm. To map a pattern to graphs in the database, we define a special notion of *block and bridge preserving* (BBP) subgraph isomorphism, which does not allow tree components to be mapped onto cyclic (biconnected) components. This notion is motivated by application and complexity considerations, and we show that it is decidable in polynomial time for outerplanar graphs. Our algorithm is based on a transformation of the outerplanar graphs to a tree and the application of a tree subgraph isomorphism algorithm [6] and a subgraph isomorphism for outerplanar graphs [5].

We note that for trees, which form a special class of outerplanar graphs, BBP subgraph isomorphism is equivalent to subtree isomorphism. Thus, BBP subgraph isomorphism generalizes subtree isomorphism to graphs, but is at the same time more specific than subgraph isomorphism. Since in many applications, subgraph isomorphism is a non-adequate matching operator (e.g., when pattern matching is required to preserve certain type of fragments in molecules), by considering BBP subgraph isomorphism we take an important first step towards the direction of studying the frequent graph mining problem w.r.t. non-standard matching operators as well.

Beside complexity results including both an analysis of our own algorithm and a discussion of the complexity of related graph matching problems, we present also empirical results. These show that in the large NCI database of 250251 molecules our assumptions are satisfied by a vast majority of the instances, and that the favorable theoretical properties of the algorithm and pattern class also translate into efficient practical performance. In particular we observe that, when pattern matching operations share computations, the time needed to evaluate the frequency of patterns does not increase significantly with the pattern size. Even though our current implementation runs in main memory, the algorithm does not need to store results between passes over the database and therefore can run on large databases.

Besides working on optimization of the algorithm, in particular, on improving the time complexity of the coverage testing, it is natural to ask whether the positive result of this paper can be generalized to arbitrary outerplanar graphs. Our algorithm exploits the constant bound on the number of diagonals only in the computation of the set of frequent biconnected graphs in its first step. Therefore, to generalize the result of this paper to arbitrary outerplanar graphs, it is sufficient to consider the following special problem: Given a finite set of biconnected outerplanar graphs and a non-negative integer t , compute the set of t -frequent patterns in w.r.t. BBP subgraph isomorphism. We do not know whether this problem can be solved in output polynomial time.

References

- [1] D. J. Cook and L. B. Holder. Substructure discovery using minimum description length and background knowledge. *Journal of Artificial Intelligence Research*, 1:231–255, 1994.
- [2] M. Deshpande, M. Kuramochi, N. Wale, and G. Karypis. Frequent substructure-based approaches for classifying chemical compounds. *IEEE Transactions on Knowledge and Data Engineering*, 17(8):1036–1050, 2005.
- [3] Tamas Horvath, Jan Ramon, and Stefan Wrobel. Frequent subgraph mining in outerplanar graphs. In *Proceedings of the Twelfth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Philadelphia, PA, August 2006. To appear.
- [4] Akihiro Inokuchi, Takashi Washio, and Hiroshi Motoda. Complete mining of frequent patterns from graphs: Mining graph data. *Machine Learning*, 50(3):321–354, 2003.
- [5] A. Lingas. Subgraph isomorphism for biconnected outerplanar graphs in cubic time. *Theoretical Computer Science*, 63:295–302, 1989.
- [6] Ron Shamir and Dekel Tsur. Faster subtree isomorphism. *Journal of Algorithms*, 33(2):267–280, November 1992.
- [7] Xifeng Yan and Jiawei Han. gspan: Graph-based substructure pattern mining. In *Proceedings of the 2002 IEEE International Conference on Data Mining (ICDM 2002)*, Japan, 2002. IEEE Computer Society.