Subspace Search for Community Detection and Community Outlier Mining in Attributed Graphs

Emmanuel Müller
emmanuel.mueller@kit.edu
emmanuel.mueller@ua.ac.be

Karlsruhe Institute of Technology, Germany
University of Antwerp, Belgium

Attributed graphs are widely used for the representation of social networks, gene and protein interactions, communication networks, or product co-purchase in web stores. Each object is represented by its relationships to other objects (*edge structure*) and its individual properties (*node attributes*). For instance, social networks store friendship relations as edges and age, income, and other properties as attributes. These relationships and properties seem to be dependent on each other and exploiting these dependencies is beneficial, e.g. for community detection and community outlier mining. However, state-of-the-art techniques highly rely on this dependency assumption. In particular, *community outlier mining* [2] is able to detect an outlier node if and only if connected nodes have similar values in all attributes. Such assumptions are generally known as homophily [4] and are widely used. However, looking at multivariate spaces, one can observe that not all given attributes have high dependencies with the graph structure. For example, social properties such as income or age have strong dependencies with the graph structure of social networks [4]. In contrast, properties such as gender are rather independent from it. Consequently, recent graph mining algorithms degenerate for multivariate attribute spaces that lack dependency with the graph structure in some of the attributes. This calls for a general pre-processing step that selects subspaces, i.e. subsets of the attributes, showing dependencies with the graph.

This talk covers several methods for the selection of such relevant subspaces in attributed graphs:

As first method, *ConSub* [3] proposes the statistical selection of *congruent subspaces*, i.e. subsets of attributes showing a dependency with the graph structure. A core challenge in selecting these subspaces lies in the modeling of dependence between graph structure and attribute values. Further, one has to ensure that congruent subspaces are selected only if there is sufficient evidence on this dependence. *ConSub* addresses all those problems by: (1) a novel measure for the degree of congruence between a set of node attributes and a graph by means

In: T. Seidl, M. Hassani, C. Beecks (Eds.): Proceedings of the LWA 2014 Workshops: KDML, IR, FGWM, Aachen, Germany, 8-10 September 2014, published at http://ceur-ws.org

of edge counts and attribute values; and (2) a comparison of edge counts in subgraphs constrained by attribute value ranges in a Monte Carlo processing. The congruence measure exploits these dependencies between random subgraphs and their attribute subspaces and *ConSub* selects attribute subsets featuring those dependencies in multivariate attribute spaces. This selection can serve as general pre-processing step for algorithms that rely on the homophily assumption on attributed graphs.

As second method, *FocusCO* [1] incorporates the user preference into the selection of relevant subspaces in attributed graphs. *FocusCO* considers communities and community outliers based on user preference. This *focused* mining is of particular interest in attributed graphs, where users might not be concerned with all but a few available attributes. As different attributes induce different clusters and outliers in the graph, the user should be able to steer the subpace selection accordingly. As such, the user controls the mining by providing a set of exemplar nodes (perceived similar by the user) from which *FocusCO* infers *attribute weights* of relevance that capture the user-perceived similarity. The essence of user preference is captured by those attributes with large weights, i.e. the *focus attributes*, which form the basis for the discovery of focused clusters and outliers.

To illustrate the applicability of common graph mining tasks and in order to evaluate these selection schemes, community detection and community outlier mining is used. The methods are evaluated on several synthetic and real world graphs, in particual on a novel benchmark graph for attributed graphs that has been derived from a case study on the Amazon co-purchase network [5]. The selection of congruent subspaces clearly enhances outlier detection by measuring outlierness scores in selected subspaces only. Furthermore, focused attributes enable a more user-oriented mining of community structures. Experiments show that both approaches outperform traditional full space approaches and as general pre-processing steps they enhance the later data mining steps on attributed graphs.

References

1. Bryan, P., Akoglu, L., Iglesias, P., Müller, E.: Focused clustering and outlier detection in large attributed graphs. In: ACM SIGKDD (2014)
2. Gao, J., Liang, F., Fan, W., Wang, C., Sun, Y., Han, J.: On community outliers and their efficient detection in information networks. In: ACM SIGKDD. pp. 813–822 (2010)
3. Iglesias, P., Müller, E., Laforet, F., Keller, F., Böhm, K.: Statistical selection of congruent subspaces for mining attributed graphs. In: IEEE ICDM. pp. 647–656 (2013)
4. McPherson, M., Lovin, L.S., Cook, J.M.: Birds of a Feather: Homophily in Social Networks. Annual Review of Sociology 27(1), 415–444 (2001)
5. Müller, E., Iglesias, P., Mülle, Y., Böhm, K.: Ranking outlier nodes in subspaces of attributed graphs. In: Workshop on Graph Data Management at IEEE ICDE (2013)