Description-Oriented Community Detection using Exhaustive Subgroup Discovery

Martin Atzmuellera, Stephan Doerfela, Folke Mitzlaffa

Summarised: January 9, 2020

1 Abstract

Usually, for mining of communities, defining communities as subsets of nodes of a graph with dense structure in the corresponding sub graph, only the structural aspects of this communities are taken into account. Typically, for each community, their description provided is no concise and easily interpretable.

This article is focuses on the description-oriented community detection, using subgroup discovery to provide an structurally valid and interpretable description for the communities, with a graph's structures and descriptive features of a graph's nodes. A descriptive community pattern, created with descriptive features of a graph, describes and identifies each community and vice versa. Essentially, patterns are looked up characterizing interesting set of nodes (i. e., subgroups) in the graph. The interestingness set of nodes that forms a community is evaluated by a selectable standard community's quality measure. The selection is based on several optimistic estimates of standard quality functions used for efficient pruning of the search space in an exhaustive branch and bound algorithm.

The approach of this article is evaluated using five real world datasets, obtained from different social media applications.

2 Introduction

Since 2010, the classical community detection in graphs, just identify subgroups of nodes of a graph with dense structure, lacking an interpretable description.

As a innovation, this article focuses on the task of description-oriented community detection, based on additional descriptive functions of the nodes of a graph contained in the network, to identify communities such as sets of nodes, along with a description. Specifically, for the identification is used a logical formula on the values of nodes, named community pattern, that describes characteristics of the nodes, providing and intuitive description of the community by and easily interpretable conjunction of attribute-value pairs.

The approach of this article is usually not achieved by classical community mining methods, that consider the nodes of a network as mere strings or ids.

The great interest of this article is that the discovered method can solve problems that are generally not addressed through standard approaches to community detection. Also because the community patterns can easily be incorporated into practical applications, for example, for recommendations in social bookmarking systems, like BibSonomy or delicious.

In contrast to global approaches, the method is exceptional. Because given community quality measures to local partitioning of a network , potentially overlapping communities can be considered. Also, is not limited to any systems and can be applied to any kind of graph-structured data for which additional descriptive features are available.

3 Main Points

The contribution of this paper is structured as follows:

First, is introduced an approach for description-oriented community detection using exhaustive subgroup discovery and then, is mentioned the COMODO algorithm for searching community patterns through a quality measure. The COMODO algorithm's is an algorithm that use an efficient branch and bound method with appropriate pruning techniques based on exhaustive subgroup discovery using optimistic estimates. The implemented COMODO algorithm's in this article contains a pruning scheme that makes the approach scalable for big data sets. Furthermore, the proposed optimistic estimates, for a range of standard community quality measure, are efficient to compute faster the description-oriented community detection using the COMODO algorithm. Specifically, the article consider a three standard community quality measures: The segregation index, the inverse average ODF (out degree fraction) and the local modularity. At the end, the different community quality measures are discussed and extended the optimistic estimates accordingly.

The article continues to summarize the basic concepts of subgroup discovery and provides general notions of graphs and community quality measures. Next, the approach to description-oriented community detection presents a series of optimistic estimates for standard community assessment functions. After that, is discussed the related work.

For demonstrating the efficiency of proposed optimistic estimates and validity of the obtained community patterns, are performed experiments using five different data sets. The experiments demonstrate the effectiveness of proposed descriptive mining approach applying the presented optimistic estimates. Furthermore, the implementation of COMODO's algorithm, confirm the validity of the comunity patterns with the structural properties of the patterns and the sub graphs induced by the respective patterns. As a consequence, the COMODO algorithm is compared to three baseline community detection algorithms.

Then, the article analisy the results of the experiments in the context of three real-world applications. The results indicate statisdically valid and significant results that de not exhibit typical problems and pathological cases such an small communities sizes that are often encountered when is using a typical community

mining method. This is facilitated by the COMODO algorithm is able to detect communities that are typically captured by a shorter descriptions leading to a lower description complexity, compared to the baseline community detection algorithms.

At the end, the article concludes with a summary and directions for future work with the intention to apply the presented method with more diverse data and extend the approach for community detection to dynamic networks.