

Description-Oriented Community Detection using Exhaustive Subgroup Discovery

Martin Atzmuellera, Stephan Doerfela, Folke Mitzlaffa

Summarised: January 9, 2020

1 Abstract

For mining of communities, defining communities as subsets of nodes of a graph with a dense structure in the corresponding sub graph, usually only the structural aspects of the communities are taken into account. Typically, no concise nor easily interpretable for each community description is provided.

This paper focuses on description-oriented community detection using subgroup discovery to provide both structurally valid and interpretable communities, with the graph structure and descriptive features of the graph's nodes. A descriptive community pattern, created with the descriptive features, describes and identifies each community, i. e., a set of nodes, 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 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.

Finally, the approach is evaluated using five real world datasets, obtained from different social media applications.

2 Introduction

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

As a novelty, this work focuses on the task of description-oriented community detection based on additional descriptive functions of the nodes contained in the network, to identify communities such as sets of nodes, along with a description. Specifically, with 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.

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

The great interest for this article is that the method is discovered can resolve problems that typically are not addressed by standard approaches for community detection. And the community patterns can easily be incorporated into a practical application, for example, for recommendations in social bookmarking systems.

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

3 Main Points

First, is introduced the approach for description-driven community detection using exhaustive subgroup discovery and is mentioned the COMODO algorithm for the discovery of community patters through a community quality measure. The COMODO algorithm is a branch-and-bound algorithm that use an efficient branch-and-bound method with appropriate pruning techniques based on exhaustive subgroup discovery using optimistic estimates.

The proposed optimistic estimates for a range of standard community quality functions, which are efficient to compute faster the description-oriented community detection using the COMODO algorithm. Specifically, are consider a number of standard community quality functions: The segregation index, the inverse average ODF (out degree fraction) and the modularity. In addition, the different community quality measures are discussed and extended the optimistic estimates accordingly.

In the following, summarizes basics of subgroup discovery, and provides general notions of graphs and community quality measures.

Next, introduce the proposed approach for description-oriented community detection and presents a number of optimistic estimates for standard community evaluation functions.

After that, discusses related work.

For demonstrating the effectiveness and validity of the presented approach, for the next section that provides experiments using five data sets and discusses their results in the context of the three real-world applications.

Finally, concludes the paper with a summary and directions for future research.