

Description-Oriented Community Detection using Exhaustive Subgroup Discovery

Martin Atzmueller, Stephan Doerfel, Folke Mitzlaff

Summarised by Laura Rodríguez Navas

January 9, 2020

The main contribution of the article

This article is focuses on the description-oriented community detection, using exhaustive subgroup discovery to provide an structurally valid and interpretable description for communities (i. e., a set of nodes), with a graph's structures and descriptive features of a graph's nodes. In addition, presents the COMODO algorithm that use an efficient branch and bound method with appropriate pruning techniques to provide the exhaustive subgroup discovery. The article as well shows an implementation of COMODO algorithm that is evaluated by a selectable community quality measures using real world datasets, and analysed the results.

As a personal opinion, I consider the approach of this article so interesting and introduces an innovation because can solve problems that in contrast to global approaches is usually not achieved by classical community mining methods, that consider the nodes of a network as mere strings or ids. Also the discovered COMODO algorithm it is very relevant since can easily be incorporated into practical applications and is not limited to any system and can be applied to any kind of graph-structured data.

Another thing to comment, is the evaluation of the COMODO algorithm during the experiments. It is also very relevant, because reveals that the COMODO algorithm it is an efficient and scalable algorithm for large datasets. Thus, that is not encountered in the typical mining methods. Additionally, the results of the experiments are statistically proved, which makes them more valid and significant.

The secondary contributions of the article

As a relevant secondary contribution of the article, it should be mentioned the suitable optimistic estimates for a range of standard community quality measures or also called community quality functions.

This secondary contribution it is relevant because the optimistic estimates:

- Are efficient to compute and enable an effective approach.
- Are used on the experiments and they are an improvement for efficient mining methods.
- The COMODO algorithm can identify the communities, providing a characteristics description, referring the main contribution of the article, and their pruning method is more efficient.

Another secondary contribution that should be mentioned of the article is the method proposed for detection of overlapping communities. In my opinion, the detection of overlapping communities should be more delved deeper.

Because contrary to the methods commonly used for detection of overlapping communities, the method proposed in this article focuses on local and non-global patterns, as is common with methods more used.

And it is important to note too, that the COMODO algorithm directly uses this method to search overlapping communities according to standard community quality measures to add the description for the communities, referring the main contribution of the article.

The structure of the article

The article has the structure that an article should have as a general scheme: the abstract, the introduction, the methodology, the results of the research, the conclusions and a list of references. As well, includes other sections like: the preliminaries, the related work and the acknowledgements.

In the structure, I found it is very curious that the article has 66 references. Is the first article that I read with this big number of references.

Specifically, the article is structured as follows: first introduce the description-oriented community detection and present the COMODO algorithm. Then, summarizes basics of subgroup discovery, and provides general notions of graphs and community quality measures. After that, considers the three standard community quality measures: The segregation index, the inverse average ODF (out degree fraction) and the local modularity.

For this three standard community quality measures, that there are introduced in the introduction section, I would like to comment that initially it seems to be an optional process, which would be applied after the COMODO algorithm and remember that the experiments are based on these standards. Maybe, it should be described in the related work section. Because the related work section contains the main information of this article in grater detail.

The remainder of the article contains the related work and provides experiments using five data sets and discusses their results in the context of three real-world applications. Finally, the article concludes with a summary and directions for future research.

Besides from what I have commented previously, on the community quality measures, I think it is a well structured article. Maybe it is a bit long and complicated. Complicated because I have found it is difficult to read, specially the part of the experiments. For the big number of examples and their simultaneous comparison. I think that with fewer examples it would be easier to understand and clearer.

Contents for a presentation

If I had to prepare a presentation, the presentation will have between 10-12 transparencies and I would follow the following structure:

- The first slide would contain the title of the article, my name as the presenter, and the submission date. Additionally, information could be added about the place where it will be presented and/or information about my affiliation.
- The second slide, maybe named process, would summarize the procedures, from a dataset to a graph, the application of the COMODO algorithm and finally the processing of the quality measures.

For example, in horizontal view: Dataset – > Graph – > COMODO algorithm – > Quality Measures.

- The third slide would contain the definition of the description-oriented community detection using subgroup discovery. Basically to answer the question: How to identify a community?
- The fourth slide, maybe named example, would contain a basic example of community detection using subgroup discovery. Perhaps this example would consist of an image of a graph, where some communities would be drawn with different colours.
- In the following three transparencies, I would provide an overview on the data sets, obtained from three different social media applications, which are used in the experiments. I not mentioned before that the social media applications datasets are: BibSonomy, delicious and last.fm.

Therefore, the first transparency would refer to BibSonomy dataset. The second slide would refer to the delicious dataset and, finally, the third slide would refer to the last.fm dataset. In addition, transparencies would include three figures that would shown the results of the application of COMODO algorithm, for each standard quality measures in the presented datasets.

Each dataset contains a high number of tags, users, resources, etc. Maybe it would be good to put an extra slide to show this data, before the last slides.

- Finally, the last slides would contain a list of main conclusions in the antepenultimate slide, another list for future work, chosen the more important elements, and in the penultimate slide the typical thank-you slide at the end, where it could be added my e-mail address.