Temporal Development of Overlapping Communities in Co-Authorship Networks

Alexa Schlegel

Freie Universität Berlin, Institute of Computer Science
alexa.schlegel@gmail.com

Abstract. The abstract should summarize the contents of the paper and should contain at least 70 and at most 150 words.

Keywords: co-authorship networks, scientific collaboration, overlapping community detection, temporal analysis, clique percolation

1 Introduction

introduction to the topic
why i am doing this
motivation
research question
limitations with my dataset
goal of the paper

2 Related Work

what areas belong to related work, what is covered here other papers doing the same as I want to do community detection and temporal aspects of network analysis

2.1 Sientific collaboration

TODO, why is this important

2.2 Co-Authorship Network

short summary networks, social networks, SNA, co-authorship networks, references to who studied those networks

2.3 Communities in Social Networks

Definition of communities
[2]
http://www.ams.org/notices/200909/rtx090901082p.pdf
[3]
[1]
https://en.wikipedia.org/wiki/Community_structure

Community Detection Algorithms in General [TODO short summary with further readings] maybe short classification of algorithms from [1]
divisive and agglomerative methos

Detecting Overlapping Communities Need for detecting overlapping communities in co-authorship networks [TODO - find citation]

[However, in real graphs vertices are often shared between communities (Section 2), and the issue of detecting overlapping communities has become quite popular in the last few years. We devote this section to the main techniques to detect overlapping communities. [1]]

One poplular method for detecting overlapping communities is the *clique percolation method* introduced by Palla et. al in 2005 [3]

Other methods are summarized by Fortunato [1] starting page 131 [TODO summary with references]

3 Clique Percolation Method

Clique percolation method (CPM) is used to identify overlapping communities in networks. The following section summarizes the main findings regarding co-authorship networks and the algorithm used in the paper *Uncovering the overlapping community structure of complex networks in nature and society* by Palla, Derenyi, Farkas and Vicsek.

The commmunity definition used in the paper relies on the fact that a community consists of fully connected subgraphs (*cliques*), that share many nodes. k-*cliques* are fully connected subgraphs with k nodes. A community in this context is called a k-*clique-community*, which is defined as a union of all k-cliques, which can be reached from each other through a number of *adjacent* k-*cliques*. Two k-cliques are called adjacent if they share $k-1$ nodes. An example can be seen in figure [TODO image k-clique and k-clique-community].

3.1 Algorithm

Based on the explained community definition the algorithm consists of the following steps, which will be explained in more detail. [TODO an example of a graph in each step of the algorithm can be seen in figure X].

1. Find all *maximal cliques*, these are cliques that are not part of larger cliques.
2. Prepare clique-clique overlap matrix.
3. Treshold the matrix.
4. All connected components represent a community.

Find all maximal cliques Maximal cliques cannot be subsets of larger cliques, that is why they are detected in decreasing order of their size. The largest possible clique size s_max is determinde by the maximal degree d_max found in the network.

(1) Determine $s = s_{max}$.
(2) Repeatedly choose a node v from the graph and
(3) extract all cliques of size s containing v then
(4) delete the node and its edges.
(5) When no nodes are left set $s = s - 1$ and start with (2) on the original graph.

The set of already found cliques do influence the found cliques in later steps, as the later found cliques are smaller. The detailed algorithm for step (3) finding cliques of size s of v can be looked up in supplemtary material to the paper on section 1.1.2, page 3. The result of this a set of all maximal cliques, this set contains n_c cliques.

Prepare clique-clique overlap matrix The dimension of the overlap matrix is $n_c \times n_c$. Each row and column represent a clique, the matrix element (not the diagonal entries) are the common nodes those cliques share. The diagonal entries represent the size of the cliques.

Treshold the matrix All off-diagonal entry smaller than $k - 1$ and diagonal entries smaller than k are set to 0, remaining elements are set to 1, resulting in a binary matrix, representing a network of cliques.

All connected components represent a community Looking at the binary matrix (or resulting graph) we just need to look for connected components, those represent the k-clique-communities.

3.2 Details on k & w^* for co-authorship networks

calculation of link weights $1/(n-1)$, with n number of authors per paper. Treshold w^* for link weights. This is how collaboration is weighted or defined.

3.3 Summary of variables an measured statistics

Maybe important what should I measure in my network.
[TODO what k to choose]

3.4 Main Findings of the paper

Overlaps in networks are significant. The distributions introduced in the paper (community size, community degree, overlap size, membership number) reveal universal features of networks. The network of communities has non-trivial correlations and specific scaling properties. Providing a tool with which to enterprete the inner organisation of large networks.

4 Community Evolution based in CPM

TODO

5 Limitations and Implications regarding my dataset

what are problems with this method and what are implications regarding my dataset and network

References

1. Santo Fortunato. Community detection in graphs. *Physics Reports*, 486(3):75–174, 2010.
2. Michelle Girvan and Mark EJ Newman. Community structure in social and biological networks. *Proceedings of the national academy of sciences*, 99(12):7821–7826, 2002.
3. Gergely Palla, Imre Derényi, Illés Farkas, and Tamás Vicsek. Uncovering the overlapping community structure of complex networks in nature and society. *Nature*, 435(7043):814–818, 2005.