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, scientific collaboration, co-authorship networks, social networks
motivation: dataset, already constructed co-authorship network, based on very simple modeling of collaboration, self constructed trashhold
research question: how do communities within co-authorship networks evolve over time? in geochemistry people often focus on one isotope system, and stick to this topic. do researchers in geochemistry really focus on one topic or do they change the field over time? why do they change field and were do they change to? by analysing the network over time this question could be answered maybe
goal of the paper of the paper is to find an appropriate method for analysing scientific collaboration networks over time

2 Related Work

community detection in complex and large graphs
different methods of community detection algorithms
communities in collaboration networks (co-authorship)
temporal aspects of communities
time slicing and dividing the network into snapshots
algorithms and implementations
focus not on different time slicing methods but more on community detection
references to temporal stuff [TODO cite moody]

2.1 Sientific collaboration

TODO definition [TODO cite]
based on who borad or narrow the definition of collaboration is, a function/measurement for edgeweight can be chosen

2.2 Co-Authorship Network

short summary networks, social networks, SNA, co-authorship networks
references to who studied those networks [Newmann and Barabasi] just references
nodes are authors, links are collaboration, wights for nodes and edges

2.3 Communities in Social Networks

Definition of communities, there is no proper definition yet
[2]
`http://www.ams.org/notices/200909/rtx090901082p.pdf`
[4]
[1]
`https://en.wikipedia.org/wiki/Community_structure` most real world networks contain parts in which nodes are more highly conected to each other than to the rest of the network, those sets are usually called clusters, communities, cohesive groups or modeuls [TODO cite], they have no wiedely accepted unique definition. the detection algorithms mainly define what a community is.

Community Detection Algorithms in General [TODO short summary with further readings] maybe short classification of algorithms from [1]
maybe there are so many aproaches
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 [4]

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 Construction of the co-authorship network

TODO, how are weights calculated in this network, what is a collaboration here. $n/(n-1)$ with n auhors for one publication

3.2 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]. The starting point for the algorithm is a undirected unweighted graph. In section 3.3 we talk about generating an unweighted collaboration, co-authorship graph using a treshhold.

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.

3.3 Construction of the network and Details on k & w^* for co-authorship networks

talk about choosing the right k and the rigth trashhold w^*. 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. all links smaller than w^* are removed.
how to chose the right k, usually between 3 and 6, then w^* is ajusted

3.4 Summary of variables an measured statistics

Maybe important what should I measure in my network. those measurement describe the quality of those detected communities
summary of variable

3.5 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.[TODO cite]

4 Community Evolution based in CPM

Palla et. al developed [3] in 2007 an algorithm based on clique percolation that allows us to investigate the time dependance of overlapping communities on a large scale, thus uncovering basic relationships characterizing community evolution.

4.1 Construction of the time dependant networks

Event in the network are paper publications. Social connection between people start before event and last for some time after the event, the higher the fequency the closer the relationship[TODO citeSocial intertia in collaboration networks, Ramosco and Morris 2006]

Weight is $n/(n-1)$ for each paper and its collaborators. The *link weight* between to nodes a and b at a certain time t is calculated as

$$w_{a,b}(t) = \sum_i w_i e^{\frac{-\lambda|t-t_i|}{w_i}}$$

The summation runs over all collaboration event in which a and b are involved. The event i occurs at time t_i and the corresponding edge weight at this time is called w_i. This function which looks like a decay functions kind of models the strenght of collaboration between authors over time considering all event ever occured in time.

A treshhold w^* is used to only include certain edges to the time dependant graph. So for each timestep a graph can be constructed based on the collaboration strenght in each time step.

here in the co-authorship network $k = 3, w^* = 1.0$ is used.

what are the time steps for co-author network [TODO find out read again]

4.2 Algorithm

The algorithm uses the clique percolation method to find communities in each time depandent graph. The starting point is set of an undirected and unweighted graphs for each timestep (snapshots usually called in other papers).

(1) Extract communities with CPM for each time step.
(2) Match the set of communities at succeeding time steps (far from trivial).

As (1) was already explained in section 3 we focus on (2) the matching of set of communities between consecutive time steps [is it really consecutive or only after one another?]

As overlapping communities are included, this step is not trivial. The relative node overlap $C(A, B)$ between two nodes A and B in a simple way can be defined as

$$C(A, B) = \frac{|A \cap B|}{|A \cup B|}$$

[TODO this method is not working for overlapping communities, paper is desribing an over complicated matching algorithm :(]

4.3 Summary of variables an measured statistics

the overall coverage of the community structure (ratio of nodes contained in at least one community) is measured
distribution of community size is measured too

4.4 Main Findings of the paper

difference between big and small communities and development over time
small communities live longer if members stay the same
if members change, they die
big communities live longer if members are changed permanently
if members stay the same they die

5 Implications regarding my dataset

what are problems with this method and what are implications regarding my dataset and network
I need a new function for calculating edge weights
find out what time slices are possible in my data[TODO]
try out CPM (there is an implementation in R) with static network
calculate measurement to see if quality of communities is good, find out right k and w^*, maybe use the one from paper but, evaluate if those are good
create snapshots of the paper and try if provided function for calculating weight at timesteps can be used
find an implementation of the community mapping for each time step

6 Future Work

find out what other methods are out there with implementations, I dont have the time to implement a cutting edge algorithm
compare my network/community structure to null-model(or other) or to other social network to see if everything is nice
maybe some time should be invested for finding a good representation for edge weights calculation, related to collaboration definition
next time choose a less complicated topic

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, Albert-László Barabási, and Tamás Vicsek. Quantifying social group evolution. *Nature*, 446(7136):664–667, 2007.
4. 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.