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 on CPM

Palla et. al [3] developed an algorithm based on clique percolation (see section 3) that allows the investigation of overlapping communities over time. They uncovered basic relationships characterizing community evolution within a co-authorship network and a phone-call network. The following section describes the main steps of the algorithm. A short summary of their main findings can be found in section 4.4.

4.1   Algorithm

The starting point of the algorithm is a set of undirected and unweighted graphs for each timestep. How to produce those temporal graphs is explained later in section 4.2. In general the algorithm uses the clique percolation method to find communities in each temporal graph, and matches the communities of consecutive timesteps.

(1) Extract communities with CPM for each graph g_t at time step t.
(2) Match set of communities at consecutive time steps of graph g_t, g_{t+1}, as follows:
  (2.1) Construct joint graph $g_\cup = g_t \cup g_{t+1}$.
  (2.2) Extract communities V with CPM in joint graph $g_\cup$.
  (2.3) For each extracted community V_i:
      – Extract communities in g_t and g_{t+1} that are contained in V_i.
      – Calculate relative overlap for each pair.
      – Match communities in decending order.
(3) Gap filling.

   The following describe the steps (2) and (3) in detail. Step (1) was explained in section 3 already.

Matching Communities   For matching communities, the *relative node overlap* $C(A, B)$ between two nodes A and B, in a simple way, is defined as follows:

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

As overlapping communities are allowed the matching from consecutive time steps in descending order of their relative node overlap can lead to missmatching. For example when small communities gain a lot of members or vice versa. An example for this problem is given in figure X [TODO-PIC]. As a solution, for each time steps t and $t+1$ a joint graph $g_\cup$ is constructed, containing all links from both networks. Let D be the set of communities at time step t and E the set of communities at time step $t+1$. The set of communities from the joint graph $g_\cup$ are extracted using CPM again and are called V. For any community $D_i \in D$ or $E_j \in E$ exactly one community $V_k \in V$ can be found. For checking weather E_i or D_j is contained in V_k the links are compared instead of nodes. For each community $V_i \in V$ the set of communities $D_i^k \in D$ and $E_j^k \in E$ contained in V_i are extracted. Now the relative node overlap between every possible pair can be calculated as follows

$$C_{i,j}^k = \frac{\left|D_i^k \cap E_j^k\right|}{\left|D_i^k \cup E_j^k\right|}$$

and the pairs can be matched in descending order.

   In figures X three examples are given: figXa is a simple matching of a propagating community, figXb showing two merging communities with one community dying and figXc showning the splitting of a community into two communities with one community is new born.

Gap Filling In some cases a community which was disintegrated at a certain time step suddenly reappear in a later timestep, due to low publishing rates for example. That means a newborn community includes a formerly dead community. This problem is overcome by just fillig the gap with the last step of the almost disintegrated community.

4.2   Construction of the temporal co-authorship network

Events in the co-authorship are paper publications. The social connection between people writing a paper together usually starts before the event and last for some time after the event. The higher the fequency the closer the relationship [5].

The edge weight resulting from one paper is $n/(n-1)$ with n authors. 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 is 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 emporal graph. So for each timestep a graph can be constructed based on the collaboration strenght per time step.

The authors of the paper used $w^* = 1.0$ for the co-authorship network, they said nothing about λ at all. (maybe explain how w/k was chosen?)

The dataset containes 142 month of publications, but I could find in the paper how the data was aggregated, because in the figues it looks like 50 timesteps in total.

4.3   Summary of variables an measured statistics

For evaluating the quality the overall coverage of the community structure (ratio of nodes contained in at least one community) is measured. Also the distribution of community size is measured.

4.4   Main Findings of the paper

The paper summarized differences between large and small communities and their development over time. Small communities live longer if the members stay the same over time. If members in small communities change frequently, they only live for a short time. Large communities live longer if members are changed permanently, if members stay the same they die quickly.

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.
5. José J Ramasco and Steven A Morris. Social inertia in collaboration networks. *Physical Review E*, 73(1):016122, 2006.