Analyzing Overlapping Clustering Methods on Biological Communities

Akhil Jakatdar, Tandy Warnow, and George Chacko

April 25, 2022

Abstract

Community detection has become an important task in bibliometrics, where the collection of publications connected through citations and references through networks has become the primary data structure by which such analysis can be framed within. A variety of methods have been proposed in detecting such communities, with a focus on clustering algorithms, but there has yet to be an effort to utilize overlapping clustering methods to detect communities within large citation networks. This paper proposes a scalable, overlapping clustering method designed to detect communities, and compares this method to existing overlapping clustering methods as well as to the existing disjoint clustering method proposed in Wedell et al.

1 Introduction

Introduce the Problem of Clustering for Bibliometrics

- Trying to identify specific communities from a larger network of publications within an overarching field
- Attempts to solve this problem within the lensof citation networks which means through the connections of publications based on how they cite/reference other publications within their field
- Focal publications can be part of many separate communities like BLAST can and should be part of many smaller communities but disjoint clustering methods leave out this possibility
- Price and Beaver, Wedell et al mention the impact of a core and periphery structure, for which can we analyze the impact of overlapping clusters on the core structure

1.1 Center-periphery communities

Discuss Price & Beaver [1] and anything else that is related to center-periphery structure in communities.

1.2 kmp-clustering

The method from Wedell et al. [2] seeks to cluster a graph into disjoint clusters, each of which exhibits center-periphery structure, as specified by numeric parameters k and p that are provided by the user. Specifically, a cluster is said to be k-valid if... and p-valid if... and m-valid if...

The kmp-clustering pipeline developed in [2] performs this clustering, based on the user provided values for k and p, and has the following four steps:

- Step 1: IKC(N,k): producing a set of disjoint clusters, each of which is k-valid.
- Step 2 (optional). Divide into smaller clusters, Recursive Graclus
- Step 3 (optional): Augment for Periphery
- Step 4 (not optional): parse again, making sure to update for kmp-validity

1.3 Overlapping clustering methods

Here we summarize the existing methods for overlapping clusters.

Lambiotte and Evans

Others?

2 Study Design

2.1 Overview

We have developed a variant of the kmp-clustering method from Wedell et al. [2] that is designed to produce overlapping clusters. We refer to this as Overlapping kmp-clustering, or OKMP for short[1] Like kmp-clustering, the user provides values for the parameters k and p and the method is guaranteed to output clusters that are *kmp*-valid. However, unlike kmp-clustering, we allow nodes to be members of more than one cluster; as a result, the output is a clustering where the clusters can be overlapping.

We study OKMP in comparison to KMP-clustering (which by design produces disjoint clusters). We also compare OKMP to other methods that are designed to produce overlapping clusters. We evaluate methods on two citation networks that we construct, both developed for the Exosome biology research community. We report empirical statistics, such as node coverage and edge coverage, and also explore cluster size. We specifically explore the number of clusters that each publication belongs to, and evaluate the correlation between citation count and cluster membership. We examine pairs of overlapping clusters to understand how allowing for multiple memberships provides more insight into community structure.

2.2 Data

Talk about the Original Wedell et al. Dataset

- In order to better understand the impact of overlapping clustering methods compared to their disjoint counterparts, we decided to use the same Exosome Citation Network created and utilized in Wedell et al.
- Important to note the SABPQ expansion used to generate this large network which we can compare in size to Zachary's Karate Club dataset when talking about scalability

Mention Retraction Correction

- Data cleaning and normalization is important in removing unwanted characteristics of a dataset that may confound future results
- Removing publications from the network that had published retractions are important in removing publications that are no longer judged by the research community they were published to to be accurate or merit in their field
- Talk about impact of the retraction correction using RetractionWatch on the Exosome Citation Network

Mention High Referencing Correction (JC)

- Many publications pushed have extraordinarily large reference counts
- Mention some statistics regarding the reference count of the network
- Decided to prune all publications with greater than 250 references from the network, many publications with > 250 references have references misattributed and thus can be consider erroneous in the data generation step
- Talk about the impact of the Jakatdar Correction on the Exosome Citation Network

2.3 Clustering Methods

The main focus of this paper is the new clustering method that we have developed, Overlapping kmp-clustering, which we refer to as OKMP for short. We also compare OKMP to other methods, which we describe below.

[1] This is an awful name and we should change it!

2.3.1 Overlapping kmp-clustering

Talk about New Stage 5

What we propose to do is add a Step 5 that will achieve overlapping clusters. More generally, the Step 5 will be able to be used with any input clustering (even if not disjoint) of a network N. And if given values for k and p, will maintain kmp-validity, if desired. Furthermore, we have a variant of this Step 5 that will maintain MCD (minimum core degree).

Proposed greedy algorithm

Suppose we have as input a network N, values for k and p, and a clustering $\mathcal{C}$, and we want to now allow for nodes to be members in more than one cluster. Thus we want to enhance $\mathcal{C}$ to create a new clustering, but using $\mathcal{C}$ as a starting point. Here is a general technique:

- Sort the nodes of N according to some criterion (the studies mention in this paper will sort by total degree of the publication in the network)
- Process the nodes in order of this criterion, from best to worst, until a stopping condition applies (could be the size of the resulting clustering, or amount of time that has passed)
  - Given node v, add v to any cluster in $C \in \mathcal{C}$ where v has at least k neighbors (alternatively, v has at least $MCD(C)$ neighbors) among the core elements of C.

In the current version of this algorithm, when we add a node to a cluster, we do not add the node as a core member and thus we only need to iterate over the cluster once to add all nodes

2.3.2 Line Graph Method

From Evans and Lambiotte, a method of overlapping clustering was proposed that involves computing the line graph equivalent of a given network and running a disjoint clustering algorithm like Leiden on the resulting line graph.

- Step 1: LG(N): produce a network where edges represent nodes, nodes represents edges
- Step 2 Cluster(LG(N)). Run disjoint clustering algorithm on the line graph

Talk about scalability issues of Method

2.4 Experiments

Recall that OKMP depends on the values for k and p, but in this initial experiment we do not consider periphery membership, and so p is irrelevant. It also depends on which publications are allowed to be put in more than one cluster; for our algorithm, this is mainly based on modifying N where we order the nodes based on total degree and then only allow the top N nodes to be in multiple clusters. However, we also have a variant where we process a user-provided set of nodes, which could be all the nodes in a specific cluster, or all the marker nodes, etc. Finally, OKMP depends on the rule for allowing a node to join a new cluster: do we only enforce k-validity, or do we enforce MCD values (where MCD stands for Minimum Cluster Degree)? Here we note the MCD value of a cluster is always (by definition) at least k, so that enforcing MCD is a stronger requirement, and may result in a node being added to fewer clusters.

- Experiment 0: Characterize step 1 of kmp-processing (i.e., we only do IKC(k)) for k ranging from 10 to 50. These will be used in subsequent experiments.
- Experiment 1: Look at OKM-clustering for different values of k and for a specific set of nodes for processing. That set of nodes will be the top N nodes based on total degree or something else. For this, we only look at enforcing km-validity (not MCD). Probably we use the top1% of the nodes in terms of total degree for this set.
- Experiment 2: Based on experiment 1, we will fix the value for k (to at most two values), and now vary the set of nodes for processing. Here we make some discoveries about this kind of overlapping clustering.
- Experiment 3: vary MCD vs k-m-validity to decide differences in sights.
- Experiment 4: Take best settings so far, but break up the largest clusters (i.e., add back in the Stage 2 using Recursive Graclus).
- Experiment 5: Allow for periphery.

In each experiment, we report:

- Characteristics of the set of nodes that are only in singleton clusters
- Node coverage
- Edge coverage
- Cluster size distributions
- Overlap between clusters

3 Results & Discussion

Shown below are the results for experiment 0. Experiment 0 consists of running the original disjoint IKC clustering method for k-values in the range of $\{10, 20, 30, 40, 50\}\}$ as well as the statistics of the clustering that include the number of non-singleton clusters, the number of singleton clusters, the minimum, median and maximum cluster sizes, as well as node and edge coverage of the network as well as the marker node set.

Experiment 0										
Method	Number of Non-Singleton Clusters	Number of Singleton Clusters	Minimum Cluster Size	Median Cluster Size	Maximum Cluster Size	Node Coverage (%)	Edge Coverage (%)	Marker Node Coverage (%)	Marker Edge Coverage (%)	Elapsed Time
IKC (k=10)	**128**	13454271	14	79.0	**214877**	**3.83%**	**11.8%**	**89.5%**	**8.33%**	26:46.34
IKC (k=20)	44	13713538	60	246.5	170413	1.97%	8.24%	48.4%	6.62%	26:11.65
IKC (k=30)	22	13901091	73	361.5	36305	0.63%	3.33%	42.1%	6.35%	25:22.49
IKC (k=40)	6	13952702	124	1936.5	20076	0.26%	1.61%	41.1%	6.29%	26:01.65
IKC (k=50)	1	**13987432**	**2004**	**2004.0**	2004	0.01%	0.11%	0.00%	0.00%	24:46.88

Shown below are the results for experiment 1. Experiment 1 consists of running OKMP on the original disjoint IKC clustering method for k-values in the range of $\{10, 20, 30, 40, 50\}\}$ with a candidate selection method of including all publications that fall in the top one percent of nodes by total degree as well as the statistics of the clustering that include the number of non-singleton clusters, the number of singleton clusters, the minimum, median and maximum cluster sizes, as well as node and edge coverage of the network as well as the marker node set.

Experiment 1										
Method	Number of Non-Singleton Clusters	Number of Singleton Clusters	Minimum Cluster Size	Median Cluster Size	Maximum Cluster Size	Node Coverage (%)	Edge Coverage (%)	Marker Node Coverage (%)	Marker Edge Coverage (%)	Elapsed Time (s)
IKC (k=10)	**128**	13446836	14	85.0	**275688**	**3.88%**	**19.1%**	**89.5%**	**25.3%**	73.74
IKC (k=20)	44	13702335	61	278.0	194798	2.05%	11.3%	53.7%	14.0%	39.84
IKC (k=30)	22	13894563	73	388.5	38260	0.68%	3.80%	46.3%	7.66%	21.61
IKC (k=40)	6	13951394	130	2299.0	20488	0.27%	1.70%	44.2%	6.82%	8.05
IKC (k=50)	1	**13987362**	**2074**	**2074.0**	2074	0.01%	0.11%	0.00%	0.00%	1.67

3.1 Results of Overlapping KMP-clustering

Note that OKMP (Overlapping kmp-clustering) depends only on the values for k and p, and also on the stopping rule (i.e., which nodes we process and allow to be in multiple clusters). In this first experiment, we explore the impact of changing the value of k and the set of nodes to process. We also then consider the impact of allowing for periphery membership.

3.2 Comparison between Overlapping kmp- and Disjoint kmp-clustering

3.3 Comparison between Overlapping kmp-clustering and Line Graph Clustering

3.4 Marker Node Analysis

Question: Can we run an MDS analysis similar to what was done in Wedell et al?

4 Conclusions

References

[1] Derek de Solla Price and Donald DeB. Beaver. Collaboration in an invisible college. *American Psychologist*, 21(11):1011–1018, 1966.

[2] Eleanor Wedell, Minhyuk Park, Dmitriy Korobskiy, Tandy Warnow, and George Chacko. Center–periphery structure in research communities. *Quantitative Science Studies*, 3(1):289–314, 2022.