Analyzing Overlapping Clustering Methods on Biological Communities

Akhil Jakatdar, Tandy Warnow, and George Chacko

May 5, 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.

Akhil - does Stage 1 check for positive modularity? When we do Experiment 0, are we checking for positive modularity? Ditto for Experiment 1.

- 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

We show various empirical statistics of the clustering, including the number of non-singleton clusters, the number of singleton clusters, the minimum, median and maximum cluster sizes, node and edge coverage of the network, and node and edge coverage of the marker node subnetwork.

3 Results & Discussion

3.1 Explaining the Quantitative Metrics

We can begin by giving a description of all metrics evaluated in the following two tables. The Number of Non-Singleton Clusters represents the number of clusters generated with at least 2 nodes in the cluster. The number of Singleton Clusters represents the number of nodes not present in any of the non-singleton clusters. The minimum cluster size represents the size of smallest non-singleton cluster generated. Similarly, the median and maximum cluster size represent the size of the median and largest non-singleton cluster generated respectively. Node Coverage is calculated by finding the percentage of nodes found in a non-singleton cluster compared to the total number of nodes in the network. Similarly, edge coverage is calculated by finding the total number of edges found in non-singleton clusters (both endpoint nodes are found in the cluster) compared to the total number of edges found in the network. We can use the same process to define marker node coverage and marker edge coverage. Marker node coverage represents the percentage of marker nodes found in non-singleton clusters compared to the total amount of marker nodes (n=95). Marker edge coverage represents the percentage of all edges that can be found in non-singleton clusters where both endpoint nodes of the edge are in the cluster (with at least one endpoint being a marker node) compared to the total edges of marker nodes.

Results for Experiment 0 are shown in Table 1. This experiment runs the original disjoint IKC clustering method for k-values in the range of $\{10, 20, 30, 40, 50\}$. We show various empirical statistics of the clustering, including the number of non-singleton clusters, the number of singleton clusters, the minimum, median and maximum cluster sizes, node and edge coverage of the network, and node and edge coverage of the marker node subnetwork.

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 (min:s)
IKC (k=10)	**128**	13454271	14	79.0	**214877**	**3.83%**	**11.8%**	**89.5%**	**8.33%**	26:46
IKC (k=20)	44	13713538	60	246.5	170413	1.97%	8.24%	48.4%	6.62%	26:11
IKC (k=30)	22	13901091	73	361.5	36305	0.63%	3.33%	42.1%	6.35%	25:22
IKC (k=40)	6	13952702	124	1936.5	20076	0.26%	1.61%	41.1%	6.29%	26:01
IKC (k=50)	1	**13987432**	**2004**	**2004.0**	2004	0.01%	0.11%	0.00%	0.00%	24:46

Table 1: Results for Experiment 0. In this experiment we run IKC on different values for k between 10 and 50, and we report different empirical statistics. Note that this experiment produces disjoint clustering, and is equivalent to running Stage 1 of the *kmp*-clustering method of Wedell et al. (2022).

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) w/ Step 5	**128**	13446836	14	85.0	**275688**	**3.88%**	**19.1%**	**89.5%**	**25.3%**	39:43
IKC (k=20) w/ Step 5	44	13702335	61	278.0	194798	2.05%	11.3%	53.7%	14.0%	29:25
IKC (k=30) w/ Step 5	22	13894563	73	388.5	38260	0.68%	3.80%	46.3%	7.66%	28:26
IKC (k=40) w/ Step 5	6	13951394	130	2299.0	20488	0.27%	1.70%	44.2%	6.82%	26:59
IKC (k=50) w/ Step 5	1	**13987362**	**2074**	**2074.0**	2074	0.01%	0.11%	0.00%	0.00%	25:16

Table 2: Results for Experiment 1. This is the result of running IKC for different values of k and then allowing some nodes to be members in multiple clusters. The set of nodes that are allowed to belong to multiple clusters is the top 1% of nodes, by total degree. These are placed independently of each other, and so order does not matter. Membership in a new cluster requires that the added node be adjacent to at least k other nodes in the IKC cluster (thus, newly added nodes do not vote on membership for other new nodes)

Experiment 2:

- X is a random set of nodes found in clusters **done**
- X is set of publications with top 5% in terms of degree **done**
- X is the same number of publications at the 95th percentile (94.5-95.5
- X is the set of publications with top 1%in terms of in-degree **done**
- X is the set of publications with top 5% in terms of in-degree **done**
- X is the set of publications between 94.5-95.5% in terms of in-degree
- X is some set of marker nodes. (ALL marker nodes) All 1200 from Wedell et al. 2022 The more recent set of 94 marker nodes The set of S nodes

3.2 Observations

When looking at Experiment 0, which looks at disjoint IKC km-valid clusterings, we notice the trend that as we increase k-value, we restrict our clusters lesser nodes from the network, albeit potentially more informative and connected ones. The number of non-singleton nodes decreases as k increases, while the median and minimum cluster size increase and the maximum cluster size decreases. As the selectivity of the clustering algorithm increases, we notice that the distribution of cluster sizes tends to the median. Node coverage and edge coverage also decrease as the selectivity of the clustering algorithm increases as expected in the inverse relationship between selectivity and coverage. Interestingly, we notice that the marker node coverage is greater than node coverage (which represents all nodes in the graph), while marker edge coverage is actually less that the average node for k=10, k=20 byt bit for large values of k up until k=50. We can attribute this to the fact that increasing selectivity values papers with higher

Experiment 2 Indegree										
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 (%)
IKC (k=20) w/ Step 5 Indegree 1%	44	13702262	61	280.5	192690	2.05%	11.0%	52.6%	13.6%
IKC (k=40) w/ Step 5 Indegree 1%	6	13951603	131	2235.5	20462	0.27%	1.69%	42.1%	6.52%
IKC (k=20) w/ Step 5 Indegree 5%	44	13621558	61	450.0	229928	2.63%	14.5%	72.6%	17.7%
IKC (k=40) w/ Step 5 Indegree 5%	6	13949455	139	2645.0	21002	0.29%	1.80%	47.4%	7.29%

Table 3: Results for Experiment 2. We now decide to fix k=20, 40 and vary the input candidate set. We take a look at the following variations: changing the candidate criterion to in-degree instead of total degree and looking at the top 1% of nodes and top 5% of nodes respectively.

total degrees which in turn tend to be papers with greater significance in their field. Thus marker node abnd edge coverage being proportionally higher for more selective values of k supports this idea.

We can then look at Experiment 1, which looks at the OKMP valid clustering based on the same original disjoint IKC km-valid clusterings from Experiment 0. Given that our algorithm only adds candidate nodes to existing clusters from the input disjoint clusterings, the number of non-singleton clusters stays the same. However, we notice the general trends of increases in minimum, median and maximum cluster sizes, and decreases in singleton clusters. These trends continue to show in node and edge coverage which both increase for all k-values. Given that this stage takes the set of the top 1% of nodes by total degree to be considered when adding to existing clusters, we can notice that the maximum increase of node coverage is bounded at 1%. We also notice that unlike Experiment 0, the marker edge coverage exceeds the edge coverage for all k-values (except k=50), which gives evidence that points to OKMP valuing these marker nodes in its clustering coverage much greater than the disjoint clustering method did.

Experiment 2 Total Degree									
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 (%)
IKC (k=20) w/ Step 5 Total Degree 5%	44	13603978	62	480.5	232822	2.76%	15.7%	77.9%	21.7%
IKC (k=40) w/ Step 5 Total Degree 5%	6	13948073	144	2909.0	21345	0.30%	1.87%	48.4%	8.10%

Table 4: Results for Experiment 2. We now decide to fix k=20, 40 and vary the input candidate set. We take a look at the following variations: changing the candidate criterion to the top 5% of nodes by total degree.

Experiment 2 Random									
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 (%)
IKC (k=20) w/ Step 5 Random	44	13711657	60	250.0	171448	2.00%	8.32%	49.5%	6.72%
IKC (k=40) w/ Step 5 Random	6	13952652	124	1946.0	20090	0.26%	1.61%	41.1%	6.31%

Table 5: Results for Experiment 2. We now decide to fix k=20, 40 and vary the input candidate set. We take a look at the following variations: changing the candidate criterion to a random 1% of nodes from the network.

3.2.1 Experiment 2 Observations

Observations:

- Indegree in general gives lower edge coverage than total degree for both 1% and 5%
- Node coverage marginally increases for k=20 indegree however
- k = 20 delta from Experiment 0 to Experiment 1 node coverage: +0.08, edge coverage + 3.06, from

Experiment 2 95th Percentile									
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 (%)
IKC (k=20) w/ Step 5 Total Degree 95 Percentile	44	13684128	61	268.5	180166	2.18%	9.04%	49.5%	6.92%
IKC (k=40) w/ Step 5 Total Degree 95 Percentile	6	13952552	124	1940.0	20170	0.26%	1.62%	41.1%	6.29%
IKC (k=20) w/ Step 5 In-degree 95 Percentile	44	13700075	60	253.0	176775	2.07%	8.64%	49.5%	7.02%
IKC (k=40) w/ Step 5 In-degree 95 Percentile	6	13952523	125	1968.0	20128	0.26%	1.62%	41.1%	6.34%

Table 6: Results for Experiment 2. We now decide to fix k=20, 40 and vary the input candidate set. We take a look at the following variations: changing the candidate criterion to the 95th percentile (94.5-95.5) of nodes by either in-degree or total degree.

Experiment 1 to Experiment 2 node coverage: +0.71, edge coverage: + 4.40

- as we increase threshold to 5 percent, node coverage increases at higher rates than edge coverage
- 95th percentile of nodes sees greater node coverage than top 1 percent for k=20

3.3 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,

Experiment 2 Seed									
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 (%)
IKC (k=20) w/ Step 5 Seed	44	13526555	62	495.0	270561	3.31%	18.1%	77.9%	22.6%
IKC (k=40) w/ Step 5 Seed	6	13948032	144	2910.5	21371	0.30%	1.87%	48.4%	8.10%

Table 7: Results for Experiment 2 Seed. We now decide to fix k = 20, 40 and vary the input candidate set. We take a look at the following variations: changing the candidate criterion to the seed set and the marker nodes.

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.

3.4 Comparison between Overlapping kmp- and Disjoint kmp-clustering

3.5 Comparison between Overlapping kmp-clustering and Line Graph Clustering

3.6 Marker Node Analysis

Question: Can we run an MDS analysis similar to what was done in Wedell et al?
Plotting modularity increases and decreases, differences in size of cluster

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.