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 lense of 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

2 Related Work

Price and Beaver

-

Wedell et al.

-

Lambiotte and Evans

-

3 Materials & Methods

3.1 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

3.2 Clustering Methods

3.2.1 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

3.2.2 New Clustering Method

Talk about Four-Stage KMP-Valid Pipeline briefly

The kmp-clustering pipeline takes as input a network N and values for parameters k and p. It then has 4 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

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

Proposed study We propose to begin with the output of kmp-clustering, run in two different ways, and follow by this greedy algorithm.

- Try $k = 10$ and $k = 20$, and don't even bother with periphery construction (drop Steps 3 and 4).
- Run two variants: one where we optimize MCD entirely (and so omit Step 2) and require maintenance of MCD, and the other where we just maintain kmp-validity, and so allow for Step 2.

Two stopping conditions to consider.

- For a specified node that is being processed, how many of the clusters to look at? We look at them all.
- For which nodes can be put in more than one cluster? Try the top 1% in terms of total degree.

4 Results & Discussion

4.1 Properties of the citation network

4.2 Results of clustering methods

4.2.1 Results of Overlapping KMP-Valid Clustering

4.2.2 Comparison between Disjoint and Overlapping Clustering Methods

4.3 Marker Node Analysis

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

5 Conclusions