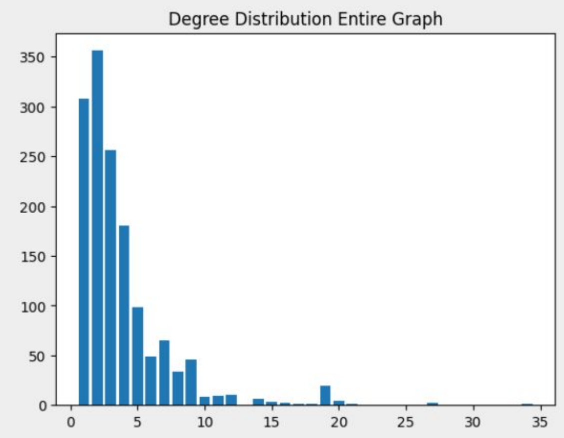
Spectral Community Detection

Group 3: Maarten van Sluijs; Roëlle Bänffer; Andrea Mangrella

Problem formalization

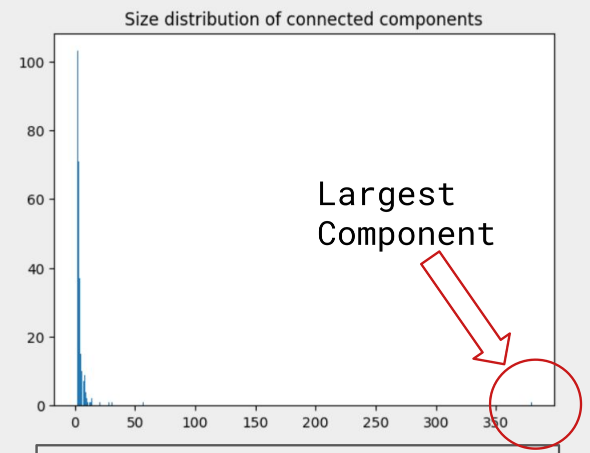
The dataset
The network science dataset consists of authors as nodes and collaborations between authors as edges. **1641 nodes, 2742 edges**. This is a ***sparse*** graph

Basic Algorithm used
- Pick some laplacian matrix L
- Compute k smallest eigenvalues and vectors: $U = (u_1, \dots, u_k)$
- Represent each node as a row of U
- Cluster the node representations using some clustering algorithm (e.g. k-means)



There is an unequal degree distribution:
Few nodes with a high degree

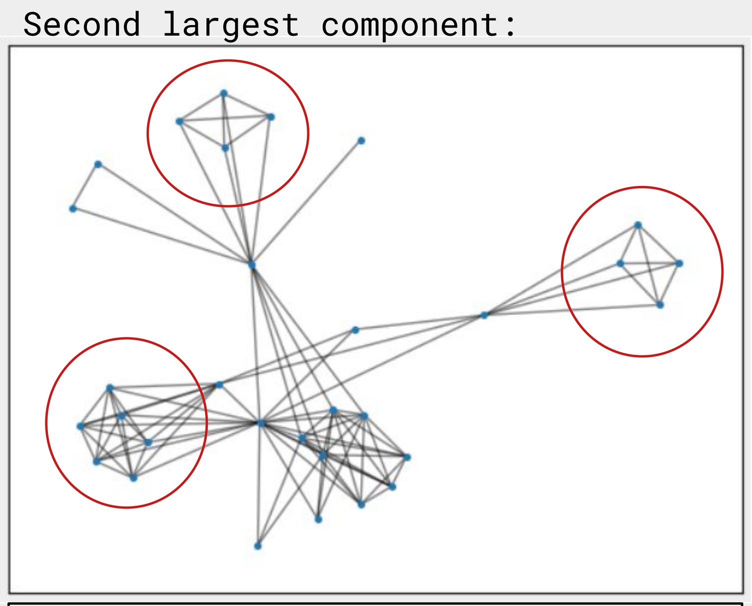
Many nodes with a low degree



Large number of connected components

Community detection only performed on the largest 4 components.

Second largest component:



Each component has many cliques connected to one another. These represent collaborations on a single paper

The goal
Partition the data in such a way that nodes within communities are similar, and nodes between are dissimilar

Find an assignment $c: V \rightarrow \{1, \dots, k\}$ where:

$$C_i = \{v \in V \mid c(v) = i\}$$

For each of the connected components we can define degree matrix D and adjacency matrix W:

$$D = \begin{Bmatrix} d1 & & \\ & \ddots & \\ & & dn \end{Bmatrix} \qquad W = \begin{Bmatrix} 0 & \cdots & \\ \vdots & 0 & \vdots \\ & \cdots & 0 \end{Bmatrix}$$

Possible Solutions

Possible Laplacians:

Unnormalized: $L = D - W$
This is used when we try to minimize RatioCut, given by the following formula:
$RatioCut(A_1, \dots, A_k) = \Sigma_{i=1}^{k} cut(A_i, \overline{A_i}) / |A_i|$

Symmetric: $L = I - D^{-0.5} W D^{0.5}$

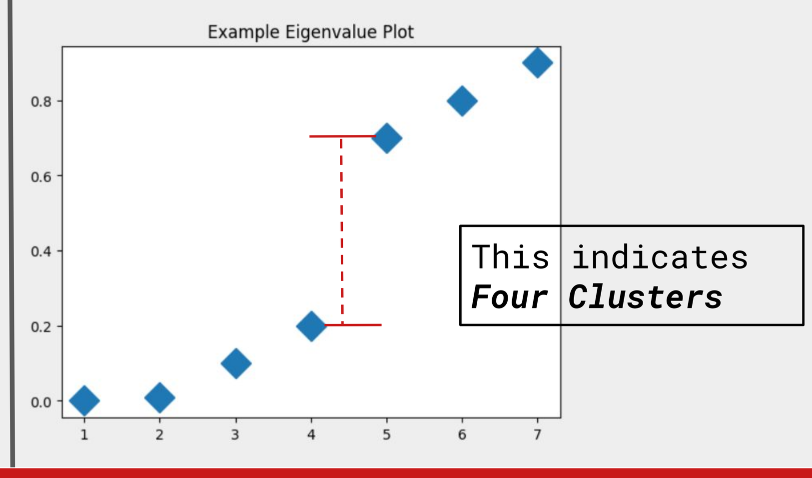
Random Walk: $L = I - D^{-1} W$

Both of these result from trying to optimize nCut:
$nCut(A_i, \dots, A_k) = \Sigma_{i=1}^{k} cut(A_i, \overline{A_i}) / vol(A_i)$

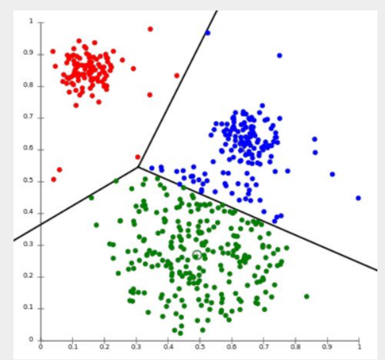
Choosing the right number of clusters:

Largest **gap** between the eigenvalues
Given the eigenvalues: $\lambda_1 \leq \dots \leq \lambda_n$
Find the **first largest gap between the eigenvalues**:

$$\lambda_i - \lambda_{i-1} \gg \lambda_j - \lambda_{j-1} \quad \forall\, 0 < j < i$$



Clustering algorithms
K-means:
- Each data point belongs to the cluster with the nearest center point (mean)
- Minimizes squared errors on the ability to reconstruct neighbors
- K-means assume clusters are round within k-radius from centroid



Discretized clustering:
Minimize the uncertainty of the class variable conditioned on the discretized feature variable
- Makes use of orthonormal transformations of eigenvectors
- Goal:find transform that leads to discretization

Column-pivotted QR factorization (CPQF):
- Extracts clusters from eigenvectors in spectral clustering
- Build for dealing with sparse SBM

Ranked Solutions

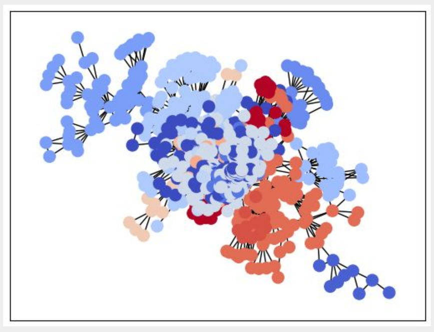
Number of communities is chosen based on the eigengap method

Pros and cons of using different types of Laplacians
Unnormalized Laplacian:
- Pro: Simple to compute and interpretable
- Con: Can create bulks because difference in node degree distributions

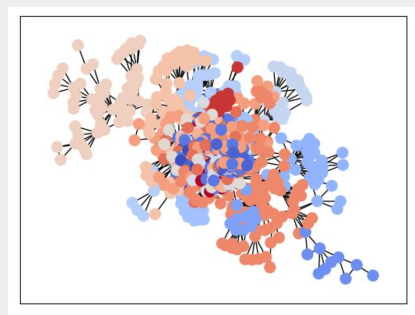
Symmetric Normalized Laplacian:
- Pro: Ensures symmetry and commonly used in spectral clustering
- Con: Computationally expensive compared to unnormalized

Based on these pros and cons, we decided to use Random-Walk Laplacian for each clustering technique:
- Can handle both directed and undirected graphs
- Performs better if we have clusters with irregular shapes
- Robust for sparse or incomplete data

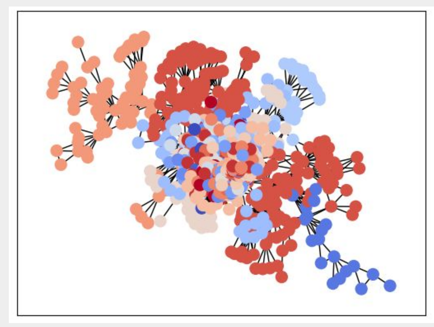
Pros and Cons of using different types of clustering algorithms



K-means
Pros:
- Simple and Intuitive
- Efficient and Scalable
- Guaranteed convergence
Cons:
- Output depends of input of k clusters
- Sensitive to outliers



Discretized clustering
Pros:
- Robust to random initialization
- Fast converges
- Interpretable
Cons:
- Loss of information to discretization



Column-pivotted QR factorization (CPQF)
Pros:
- Numerical stability
- Robustness
- No tuning parameters
Cons:
- Difficult to interpret

Applied Solution

To evaluate effectiveness of method we use:

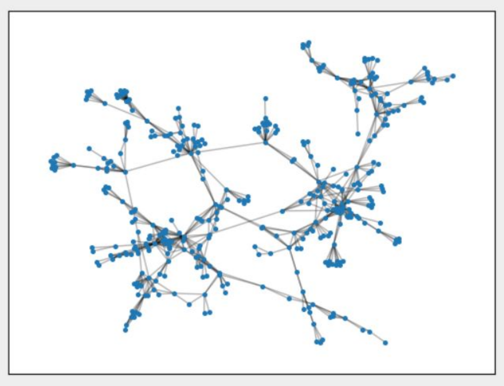
$$performance = \frac{\#\text{ of intra-comm edges} + \#\text{ of inter-comm non-edges}}{\text{Total }\#\text{ of possible edges}}$$

[2]

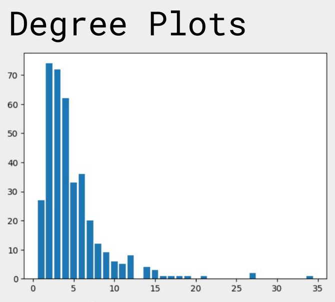
K determined using eigengap.
Random-walk Laplacians used for all networks given unequal degrees.

Only the best clustering is given below.

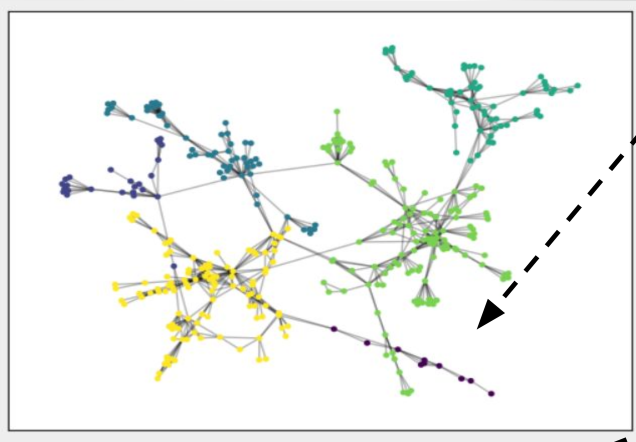
Component 1 | k = 6



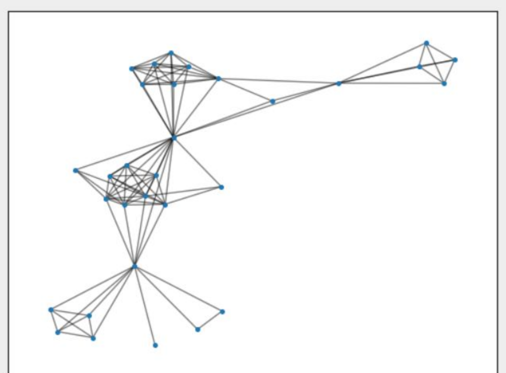
Degree Plots

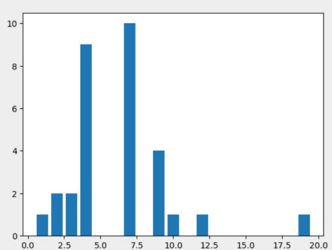


Performance
k-means: 0.77
discretized: 0.78
cluster QR: 0.78

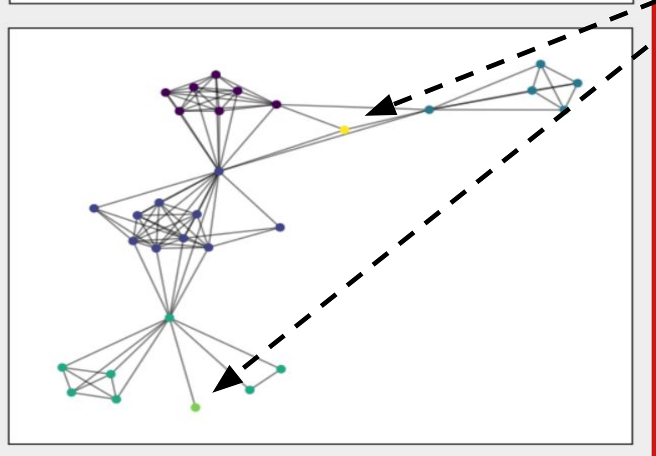


Component 2 | k = 6

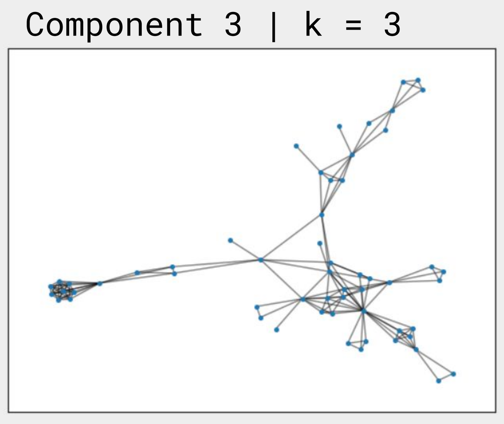


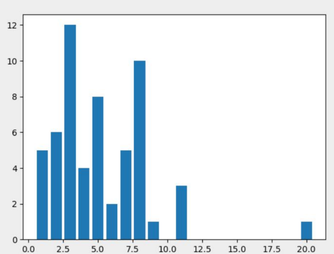


Performance
k-means: 0.82
discretized: 0.83
cluster QR: 0.82

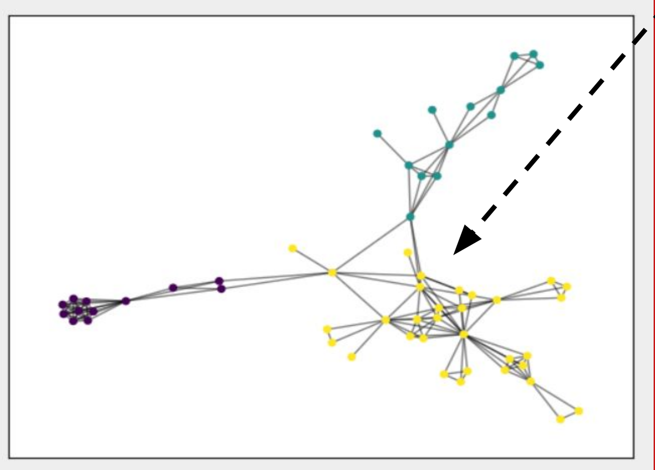


Component 3 | k = 3

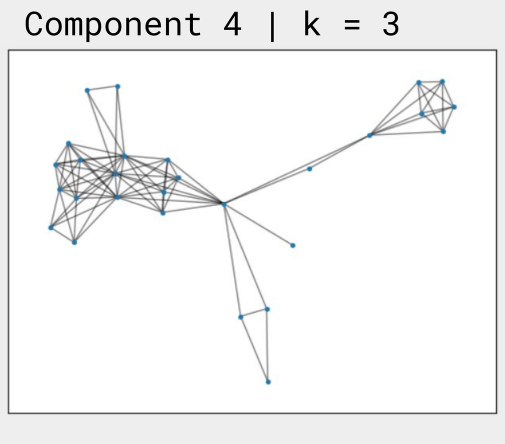


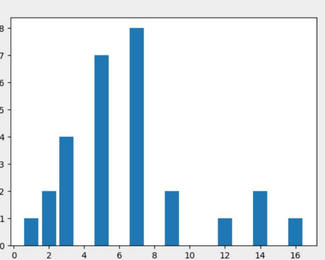


Performance
k-means: 0.69
discretized: 0.69
cluster QR: 0.69

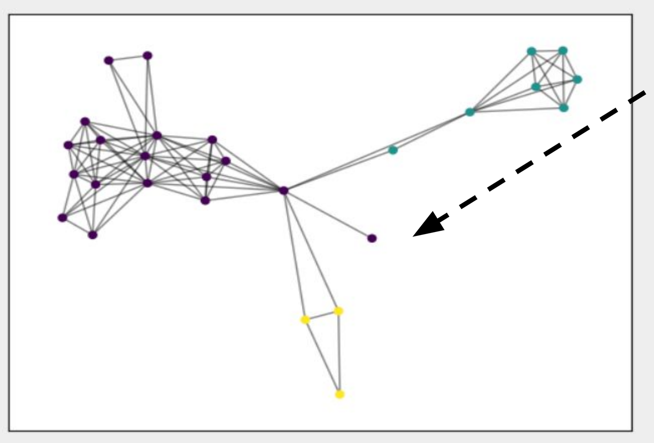


Component 4 | k = 3





Performance
k-means: 0.79
discretized: 0.79
cluster QR: 0.79



Solution Evaluation

Component 1:
- Neatly partitioned.
- Each cluster consists of a handful of cliques.
- Bottom right purple community seems more stretched than other communities.

Component 2:
- Eigengap method indicated 6 clusters.
- Seems unlikely as two of the detected communities have 1 node (yellow and light green).
- Light green node can clearly be clustered with other greens.
- Yellow node could belong to either of its bordering communities

Component 3:
- Very clean partition.
- Connections between communities are small cliques.
- While the communities themselves consist of bigger cliques.

Component 4:
- Component has one central node with three communities attached to it.
- It also claims a singular node for its community.

- Yellow component only consists of three nodes.
- Despite eigengap giving 3 clusters, yellow could be merged with other communities.

Additional Data:
- Additional meta-data for evaluation was available.
- However, hard to integrate into graph data.
- Therefore, evaluation done without this additional information.

Graph Meaning:
- Connected components are all collections of cliques.
- Shared nodes represent authors that have co-authored papers with the authors of its neighbouring cliques

Community Meaning:
- Strongly connected communities are authors with multiple distinct collaborations.
- Weakly connected cliques only have one or two collaborations on multiple papers.

Clustering algorithms:
- [1] says even simple clustering algorithms are sufficient.
- Similarity in our results supports this

Effect of sparsity:
- [1] indicates that spectral methods are less suited for sparse networks.
- This effect not evident from our experiment. Largest components are all quite neatly clustered into communities.

Overall conclusion:
- Spectral method performs well for this dataset.
- It detects both large and small communities, and segments components cleanly.
- Some errors are present. Likely caused by inaccuracy of eigengap method.

References

[1] Fortunato, S., & Hric, D. (2016). **Community detection in networks: A user guide.**
[2] Pattanayak, H. S., Verma, H. K., & Sangal, A. L. (2018, December). **Community detection metrics and algorithms in social networks.**
[3] Von Luxburg, U. (2007). **A tutorial on spectral clustering.**

TU/e EINDHOVEN UNIVERSITY OF TECHNOLOGY