Clustering of research papers using unsupervised learning

Juliette Limozin
Nature Publishing Group

nature publishing group npg

Introduction

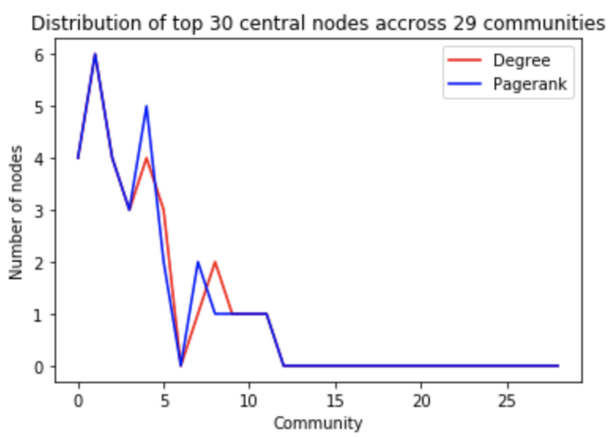
The aim of this data science project is to explore the structure of a collection of journal documents and citations between papers, through unsupervised learning, to attempt to group documents by their similarities as well as grouping documents that cite each other. We will also look at different rankings of importance of documents according to their citations.

Algorithmic grouping of the documents

- We first clustered the documents by their similarities between each other. For that we proceeded to perform a K-means algorithm to cluster the documents into 2 to 30 groups.
- Most optimal number of clusterings was 16
- Evaluated according to the Calinski-Harabasz, Davies-Bouldin and Silhouette scores

Graph-based grouping of documents

- Performed the Clauset-Newman-Moore greedy modularity maximisation algorithm to obtain an optimised partitioning of the papers, divided into 29 communities
- This is a graph-based clustering
- The top 30 most central nodes according to Degree Centrality and Pagerank were distributed similarly



Visualisation of the clusters

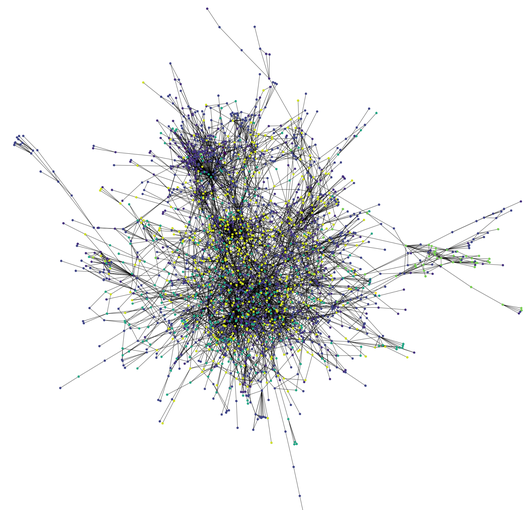


Fig. 2: Algorithmic clustering

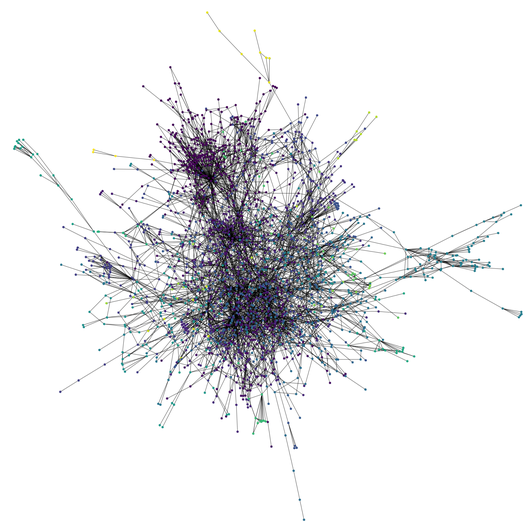


Fig. 3: Graph-based clustering

Which document was the most "important"?

- We ranked the documents according to degree, betweenness centrality and pagerank
- Best document according to citation paths: document 1245
- Rankings across measures are completely uncorrelated
- Despite the fact that these rankings are uncorrelated to each other, they give the same best node, suggesting that using the combination of these three measures is a good tool to evaluate the most central paper for future datasets

	Degree	Betweenness	Pagerank
0	1245	1245	1245
1	271	1846	1563
2	1563	1894	1846
3	1846	1563	271
4	1672	271	1672

Fig. 4: Top5 most central nodes according to Degree, Betweenness centrality and Pagerank

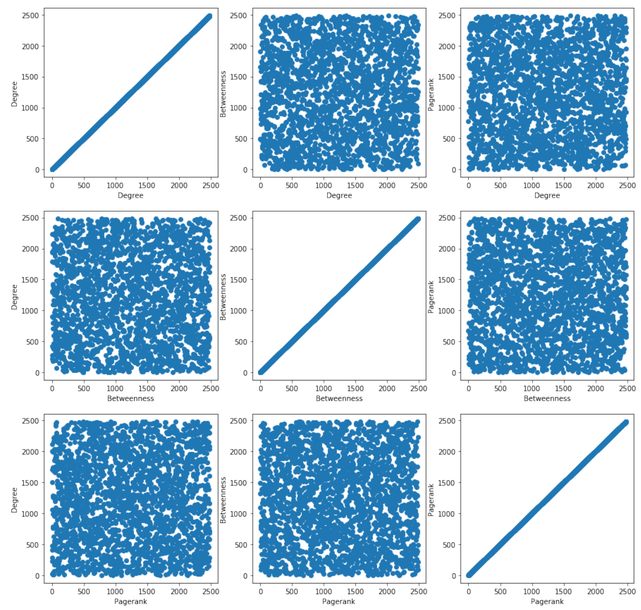


Fig. 5: Correlation plots of different rankings