Methods for Data Science

PATTERNS IN PAPERS

COMMUNITY DETECTION AND UNSUPERVISED LEARNING APPLIED TO SCIENTIFIC RESEARCH PUBLICATIONS

Imperial College London

Peter M N Hull

Department of Mathematics, Imperial College London

peter.hull16@imperial.ac.uk

Introduction

The 21st century has seen the introduction of high-dimensional data analysis as an essential tool in the statistician's toolbox. We seek to answer the question of whether data science methods can be used to reveal relationships about research papers in scientific disciplines.

Data

The data set used consists of a collection of 2485 text documents. Each data point in the collection corresponds to a paper in a scientific research area. The data point is summarised by a 1433 dimensional vector describing the written content of the paper. Non-directional citations existing between papers are contained in an adjacency matrix, encoding a citation network - nodes represent the documents, and edges represent citations.

The 2485 x 1433 feature matrix encoding the written content of papers is used for unsupervised learning, and the adjacency matrix is used for network analysis and community detection.

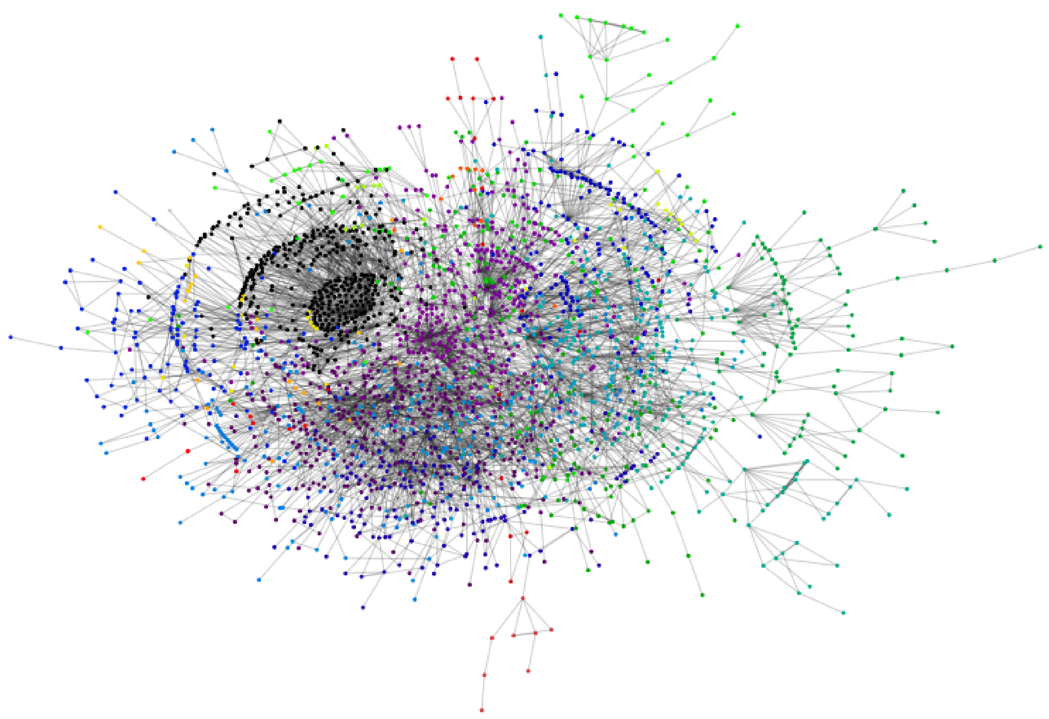

Figure 1: Citation network. 29 communities determined by Clauset-Newman-Moore Greedy Modularity Algorithm represented by node colouring.

Methods

Clustering based on Document Characteristics

K-Means Clustering algorithm was applied to the feature matrix to group the data set into different clusters. The optimal number of clusters to use was determined by internal clustering scoring criteria: Calinski Harabasz (CH), Silhouette Score, and Davies-Bouldin (DB). This was supported by quantitatively observing distances from points to centroids.

Community Detection

Communities in a network are groups of nodes more likely to be connected to each other than to those outside of the community. Clauset-Newman-Moore's Greedy Modularity Maximisation (CNM) algorithm (1) was implemented to find communities in the citation network.

Finding Central Documents using Citations

Three measurements of centrality were calculated for each node in the network: Degree, Betweenness, and PageRank. Degree and PageRank are both calculated based on the edges coming into nodes, and Betweenness is determined by the proportion of shortest paths that the node is a part of. Spearman's Rank Correlation Coefficient and Weighted Tau Coefficient were computed for pair-wise combinations of node rankings to determine the similarity in centrality measurements.

Results

- The internal clustering scores gave evidence for 20 clusters.
- CNM determined 29 communities as the optimal partitioning of the documents in the citation network, as seen in Figure 1.
- One document was the most central by far by all measurements, likely representing a foundational paper.
- The most central document was in the largest CNM community, represented by the black nodes in Figure 1.
- Degree and PageRank were found to produce the most similar complete ranking of central documents.

Measurement 1	Measurement 2	Weighted Tau	Spearman's Rank
Degree	Betweenness	0.820	0.739
Degree	PageRank	0.792	0.934
Betweenness	PageRank	0.653	0.771

Table 1: Similarity of Node Rankings of Centrality Measurements

Comparison of Results

- Adjusted Mutual Information Score and Adjusted Rand Index determined that the communities and clusters were significantly different.
- Figure 2 shows the very different visual clustering results produced by K-Means clustering .
- Without the information about citations, there are fewer visual trends within the citation network.
- Differences in results are likely due to differences in the information captured by text features and by citations.

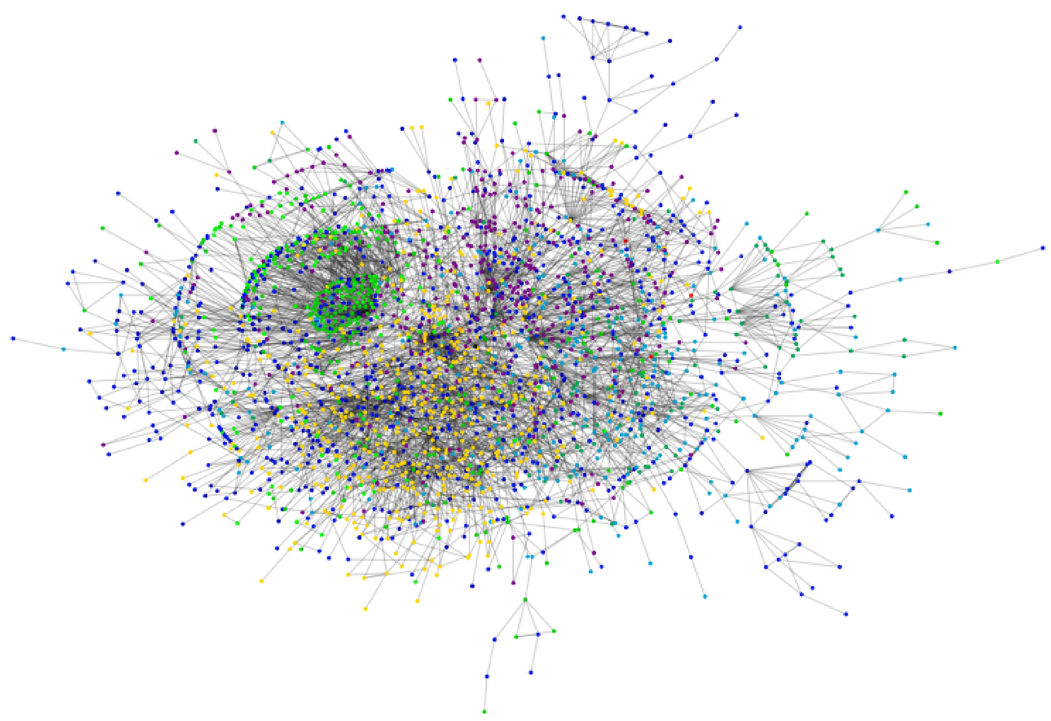

Figure 2: Citation network. 20 clusters determined by K-Means Clustering represented by node colouring.

Conclusions and Applications

- Unsupervised clustering and community detection provide two approaches to revealing information about how research papers could be grouped into sub-disciplines.
- Centrality analysis in citation networks provides a method to identify foundational or important papers in a certain field.
- Data science methods have profound applications for bringing structure to publications in research areas.

References

[1] Clauset A, Newman EK, Moore C. Finding community structure in very large networks. Phys. Rev. E; 2004.