Interrelationship of Journals from Various ASJC Domains

Motunrayo Ibiyo

1. Introduction

Over 64 million academic papers have been published since 1996[1]. Academic papers are critically structured documentation that focuses on factual arguments, and logical reasoning as regards a defined subject matter. These documentations allow scientists and researchers share their processes enabling other scientists or researchers in similar fields to evaluate the research or learn from it. References are a noteworthy component of an article; since they identify previous publications that have had some influence on the research presented in the current paper. By doing so, they establish a connection between the current study and the worldwide network of research publications. [2]

To understand the interconnection between research works in various fields, authors, or articles, citation analysis is usually conducted. Citation analysis is a technique used in library and information science to understand the impact of articles, or authors, knowledge flow, the diffusion of ideas intellectual structures of science, relevance of information resources.[2]

This report investigates the interrelationship of various research domains carried out at Massachusetts Institute of Technology (MIT). This was conducted via citation analysis of the artifacts published and cited in of those domains. The goal of this project is to understand which domains have had the high impact on other domains.

The research will be carried out following the six steps highlighted by [2]

1. Delineation of research field being studied.
2. Selection of core sets of objects representing the selected research field.
3. Measure of connectedness between objects in the core sets.
4. Multivariate statistical analysis
5. Network analysis and visualization
6. Interpretation of results and Validation

The target of this report is to understand the impact of computer science research on research work in other domains as well as find how other domains have impacted research carried out in computer science.

The Networkx library will be the main library used for the analysis. Other python libraries such as numpy and pandas for data preprocessing, matplotlib and seaborn for visualization and sklearn for similarity measurement.

The remaining part of this report will be discussed in three sections. The methodology section explains the methods used to implement the steps listed above. The results section shows the outcome of each step, and the conclusion summaries the finding of these project.

2. Methodology

In this section we will be discussing the method used to implement each step highlighted by [2]. To analyse the impact of computer science on other domains, this project analysed journals published by the research community of MIT alone within 1950 and 2018. The dataset was obtained from a citation database maintained by the Lens and indexed in I³ Open Innovation Dataset Index.

2.1. Delineation of research field being studied.

This step involves the clear definition of the research field, or scholarly communities that is to be considered in the project. Other form of delineation involves specifying the year of publication, publisher, or country the research was done.

The dataset from the citation database was pre-processed to extract nodes, edges as well as the attributes of the nodes. The nodes represent the lens-id assigned to the article and the edges shows citation relationship. For example, if paper A is in the reference list of paper B and paper C cites paper B, then two relationship for three nodes are extracted as (B, A) and (C, B). Other information such as the publication year, publisher, the All-Science Journal Classification (ASJC) code [3], list of affiliations, and paper type were also extracted for each node. The extracted data were cleaned, pre-processed, and analysed.

The publication year was also grouped into 5 groups.

Year range	Group
1950 - 1965	1
1966 - 1980	2
1981 - 1995	3
1996 - 2010	4
2010 2018	5

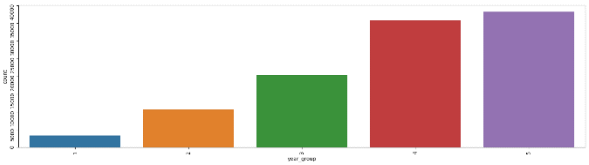

Figure 1:Distribution of dataset by year group

2.2. Selection of core sets of objects representing the selected research field.

To select a sample from the dataset that represent the different domain with the most relevant research works. Citedness-based selection technique was applied. Citedness techniques filters out articles that are below a set threshold. Article with less than 5 citation counts were removed from the dataset. The number of articles from each domain was also reduced to the top 2000 most cited article in each domain.

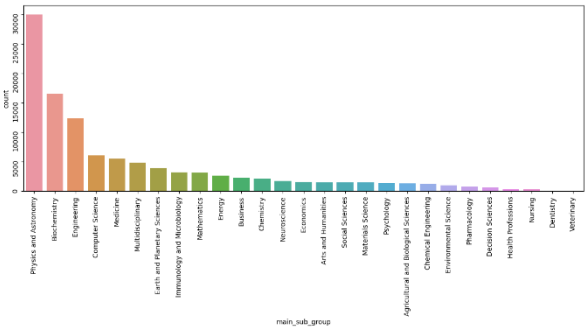

Figure 2: Distribution of ASJC domain before selection.

2.3. Measure of connectedness between objects in the core sets

To measure the connectedness between the different domains, a network of domains was generated from the graph of articles. Then a co- citation matrix was generated from the new network. A co-citation matrix shows objects that appear in the same reference list. In this project the matrix shows domains that are referenced together in another domain.

To generate a co-citation matrix, an adjacency matrix representing a the network with a $V \times V$ matrix $M = [f(i, j)]$ where each element $f(i , j)$ contains the attributes of the edge (i, j). [4]The element considered for this project was the weight which shows the number of times domain A referenced domain B in the dataset. To extract information from the co-citation matrix, the cosine similarity function was applied on it. Cosine similarity function determines the cosine of the angle between the vector representing co-citation coefficient of two domains.[5]

2.4. Multivariate Statistical Analysis

The co-citation matrix was analysed with the Factor analysis – Linear method allowing for the detection of underlying factors from relationships between numerous domains.

2.5. Network Analysis and Visualization

The network of the citation matrix was visualized using python networkx library. The dataset was aggregated to a graph of domains. This network was then analysed using graph visualization, and clustering. The Girvan newman algorithm was used to cluster the newtwork. Girvan Newman algorithm uses the iterative elimination of edges that have the highest number of shortest paths between nodes passing through them.

Four centrality measures were investigated.

1. Most connected node: this shows the node with the highest number of degrees.

2. Closeness centrality: this is the average distance from all other nodes.

3. Betweenness centrality: this determines the amount of influence a node has over the flow of information. Used to decipher the node that can server as a bridge to other domains.

4. Eigenvector Centrality: this measures the transitive influence of the node.

3. Results

This section shows the output of each step highlighted in the methodology.

3.1. Delineation of research field being studied.

Information on 194,506 published work was extracted from citation database. From these papers with 15,347,188 edges were generated from the reference and citation list. 86,188 articles with no ASJC code were removed from the dataset leaving 108,318 articles. This invariantly removed publication types like 'reference entry, 'dataset', 'dissertation'.

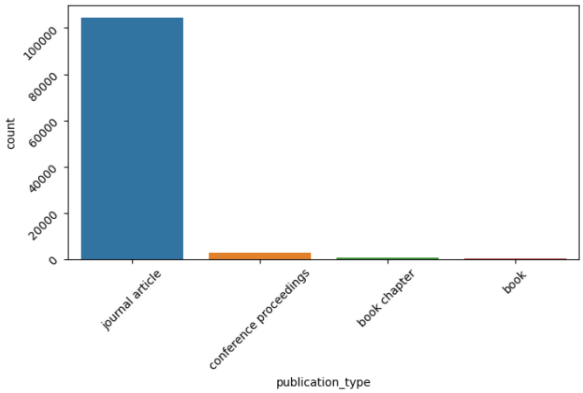

Figure 3: Distribution of publication type

3.2. Selection of core sets of objects representing the selected research field.

Following this step, the number of articles decreased to 39,295; however, the dataset was not skewed to a particular domain as seen in the full dataset. This did not affect the distribution of the articles in the year groups.

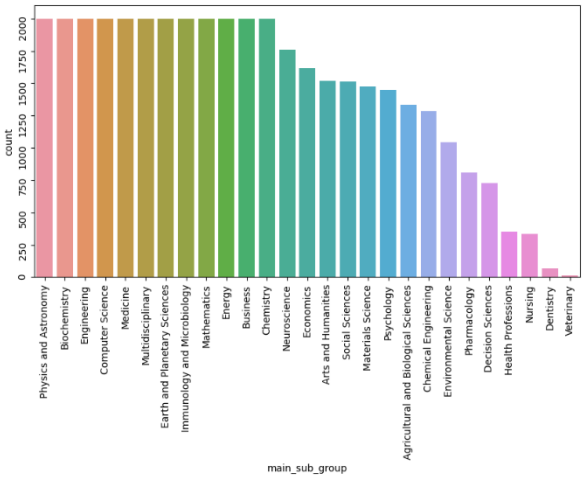

Figure 4: Distribution of ASJC domain after step 2

3.3. Measure of connectedness between objects in the core sets

The process to measure connectedness involved the generation of citation matrix and a co-citation matrix. Fig 5. shows the heat map of the cosine similarity matrix. From the figure, it can be observed that:

- no pair had a value below 0, hence there is little dissimilarity between all the domains.
- Domain pairs (Material Science and chemistry), (Dentistry and Veterinary) had the highest similarity, hence they usually cited in similar domains.
- Computer science had a similarity above 0.5 for all the domains.

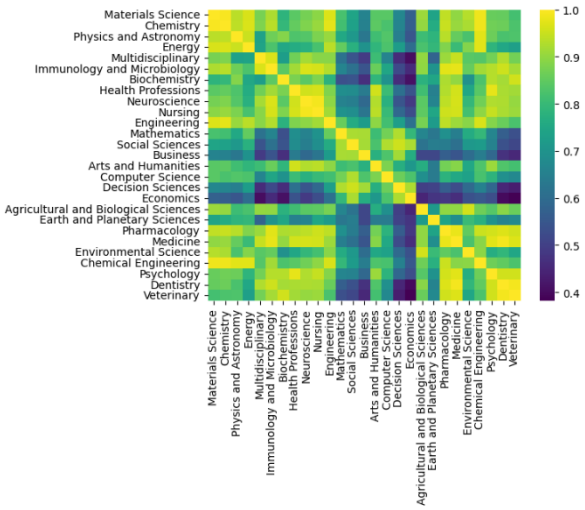

Figure 5: Heat map of cosine_similarity of co-citation matrix

3.4. Multivariate Statistical Analysis

The result of the factor analysis showed that only 5 domains had all positive value for all three factors selected. Of these 5, computer science had the highest communality.

Domains	Factor1	Factor2	Factor3	communalities
Energy	0.24	0.02	0.92	0.90
Neuroscience	0.78	0.11	0.48	0.85
Computer Science	0.59	0.84	0.24	1.12
Agricultural and Biological Sciences	0.90	0.04	0.29	0.89
Chemical Engineering	0.23	0.92	0.25	0.97

3.5. Network Analysis and Visualization

The graphs below show the relationship between the domains. The citation graphs in Fig 6. shows all the domains were analysed. From the graph it is easily observed that the articles with the highest citation counts.

- Computer science articles were not cited by articles in dentistry, veterinary, and pharmacology.
- Fig. 7 shows a strong interrelationship between all domains in physical sciences category. Especially the relationship between computer science and mathematics.
- From Fig. 8 and 9, it is observed that computer science articles were not cited by articles in dentistry, veterinary, and pharmacology. Showing a weak relationship between these domains.
- Art and humanities and business had a strong relationship with computer science. Also, there was a strong relationship also between articles categorised as multidisciplinary and psychology.

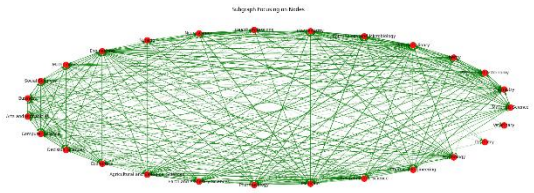

Figure 6: Full network of all domains

Figure 7: Network of physical sciences domains

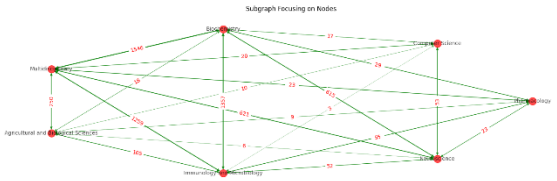

Figure 8: Network of Life Sciences and Computer Science

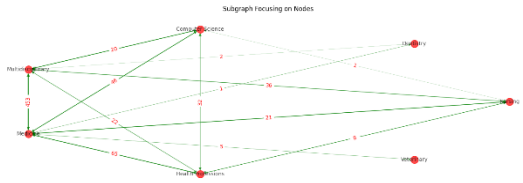

Figure 9:Network of Health Sciences and Computer Science

The Girvan Newman clustering algorithms divided the network into two clusters, with veterinary medicine occupying a single cluster alone white the other domains were placed in the first cluster.

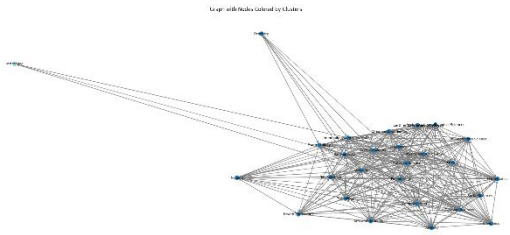

Figure 10: Cluster result of Girvan Newman clustering algorithm

The table below show the result of various centrality measures. From the table it is seen that articles in medicine and multidisciplinary domains are the most central.

Centrality Measure	Domain
Most connected node	Medicine
Closeness centrality	Multidisciplinary
Betweenness centrality	Medicine
Eigen vector centrality	Multidisciplinary

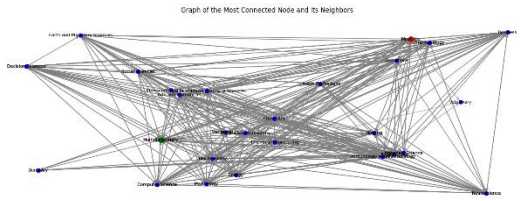

Figure 11: Most connected nodes and their connections.

4. Conclusion

This project reports on the analysis of scholarly articles published by researchers at MIT across various domain. From the analysis it can be observed that most references used by the researchers in MIT are articles published by researchers from other institutions. This was noted by the number high reduction in the dataset when the source and target it were limited to articles available in the dataset.
From the result of the project, the strength of the connection between various domains can be identified easily. Also, the influence of domains like computer science and engineering can be deciphered.
With the co-citation matrix, it is also possible to group domains that usually appear together in other domains.

From the project, it was easy to download the dataset as Json files from the database URL. Downloaded data required several preprocessing for it to be useful.
The extracted database had a lot of nodes and edges; however, a sparse network was generated by the edges. This is a common problem in citation analysis, hence the need to carry out step 2: selection of core sets of objects representing the selected research field.
The sparseness of the plain network resulted in difficulty of clustering algorithm in finding a meaningful cluster.

The current project only considered research work from MIT. In future works, publications from other institutions could be explored. This will help reduce the sparseness encountered with the articles alone. Other databases that can be explored include Google Scholar, Scopus, and Web of Science.

The script for this project is available in my github repository: https://github.com/Motunrayo244/EDISS-MP/blob/f00555d62c1094a5c6b23aa8939740766175cc41/Data%20Science/Project_work/Project3/Mini_Project3.ipynb

5. References

[1] "Number of Academic Papers Published Per Year – WordsRated." Accessed: Nov. 11, 2023. [Online]. Available: https://wordsrated.com/number-of-academic-papers-published-per-year/

[2] D. Zhao and A. Strotmann, "Analysis and Visualization of Citation Networks," *Synthesis Lectures on Information Concepts, Retrieval, and Services*, vol. 7, no. 1, pp. 1–207, Feb. 2015, doi: 10.2200/S00624ED1V01Y201501ICR039.

[3] "[ASJC] All Science Journal Classification Codes -." Accessed: Nov. 12, 2023. [Online]. Available: https://scientificresearch.in/asjc-all-science-journal-classification-codes/

[4] Z. Budrikis, "Disorder, edge, and field protocol effects in athermal dynamics of artificial spin ice," *Solid State Physics - Advances in Research and Applications*, vol. 65, pp. 109–236, 2014, doi: 10.1016/B978-0-12-800175-2.00002-9.

[5] X. Bai *et al.*, "Recommendation Algorithms for Implicit Information," *Service Science, Management, and Engineering: Theory and Applications*, pp. 77–94, Apr. 2012, doi: 10.1016/B978-0-12-397037-4.00005-3.