School of Information Technologies and Engineering, ADA University

CSCI4700 – Data Mining

Spring 2025

Course Project Report

Team <3>

	Full Name	Student ID
Member 1	**Laman Panakhova**	**16882**
Member 2	**Mehriban Aliyeva**	**15536**
Member 3	**Emil Niyazov**	**16003**

Graph Citation Network

Problem formulation

Formulate the problem. What type of task is it?

In the vast world of academic research, thousands of papers are published every year, contributing to the growth of knowledge across various fields. However, understanding how these papers influence each other and identifying the most impactful research is a challenging task. Our project focuses on analyzing citation networks, where research papers are connected based on how they cite one another. This will help us uncover key understandings such as which papers have the most influence, how research communities form, and how scientific knowledge evolves over time.

At its core, this is a graph-based data mining problem. Each paper will be treated as a node, and citations will form directed edges between them. By examining the structure of this network, we aim to identify important patterns, influential papers, and hidden relationships within the research landscape.

Discussion of related works (optional)

What has previously been done by others on this topic?

Analysis and extended literature review citations show that the prior research has explored citation networks extensively. Works such as "The PageRank Citation Ranking" by Lawrence Page et al. introduced the idea of ranking nodes based on link structures, which inspired using PageRank for citation analysis. Community detection algorithms like Louvain have been widely used to find clusters in citation graphs, reflecting distinct research communities (Page, Brin, Motwani, & Winograd, 1999). Additionally, large datasets like OpenCitations and CrossRef Corpus have been mined to study the growth, evolution, and influence of research trends over time. We have been heavily used these two

corpuses to extract the DOIs and the list of papers that cite the given specific paper. We also computed the top page rank as it is very popular in many citation networks researches. For getting some insights we also checked the existing research on the given topic and decided that this is the area that needs to be researched more.

Wang et al. (2022) demonstrated that statistical research has increasingly influenced a variety of scientific fields. By employing citation network analysis and a local clustering approach, they identified key statistical communities contributing to external disciplines, underscoring the expanding role of statistics in interdisciplinary research (Wang et al., 2022).

White (2019) showcased the effectiveness of data mining tools in citation analysis, revealing that a limited number of journals comprised the bulk of citations in geological sciences. This approach highlights the potential of combining scripting and bibliometric data for efficient citation studies (White, 2019).

Wahyuni et al. (2018) illustrated that integrating social network analysis with text mining enables the identification of key publications and emerging research trends, offering a comprehensive approach to understanding the structure and evolution of scientific literature (Wahyuni et al., 2018).

Our project builds on these concepts but applies them to a custom, focused dataset collected manually through scraping, providing a tailored insight into specific research fields.

EDA and data preprocessing

Describe your EDA and data preprocessing steps. Justify your choices. Include any figures and graphs.

Data Collection

We used the **sortgs** tool — a Python-based automated scraper for Google Scholar — to collect papers based on nine different keyword searches. The keywords such as adaptive_k-means_with_fuzzy_logic, dominant_color_quantization_using_fuzzy_clustering, fuzzy_clustering_in_image_segmentation, fuzzy_c-means_vs._k-means_in_color_clustering, fuzzy_color_clustering_in_image_processing, fuzzy_k-means_for_dominant_color_extraction, fuzzy_rule-based_color_segmentation, hybrid_fuzzy_k-means_for_color_feature_extraction, soft_clustering_for_dominant_color_extraction were chosen specifically very related to each other for obtaining clearer and more related graph of papers. The main built-in code for the given data craping process sortgs "machine learning" --startyear 2000 --endyear 2025 was used to restrict the size of the data by filtering based on the specific year

range. Each search collected exactly 100 entities in 9 different csv files, resulting in a dataset of 900 entries. For each paper, we extracted the following:

- ✓ DOI (Digital Object Identifier) in 9 according different updated csv files.
- ✓ List of papers citing it (Cited_By) in 9 according different json files.

We also cross-checked missing citation data through sources like **OpenCitations** and **CrossRef** where needed.

EDA & Data Preprocessing

Our initial focus was to clean and structure the raw citation data in order to ensure the reliability and efficiency of downstream graph mining tasks. This involved both **data preprocessing** and **exploratory data analysis (EDA)**, outlined below:

Preprocessing Steps:

Data Cleaning

We began by removing entries that lacked valid DOIs or citation lists. Additionally, we filtered out citation links labeled as "Not Found" to avoid false or broken references in the graph structure.

Directed Graph Construction

Using the cleaned data, we constructed a **directed citation graph**, where a directed edge from paper A to paper B represents that A cites B. This structure models academic influence and citation flow effectively.

Isolated Node Removal

To preserve a meaningful network, we eliminated nodes (papers) with no incoming or outgoing citations. These isolated nodes typically resulted from unreferenced or disconnected papers that did not appear in other citation lists.

Top-N Node Filtering

We filtered the graph to include only the **top 300 most connected nodes** based on total degree (in-degree + out-degree). This allowed us to retain a strongly connected component, which is essential for computing global graph metrics and producing clear, interpretable visualizations.

Exploratory Data Analysis (EDA)

To understand the structure and quality of the cleaned dataset, we performed several EDA tasks:

Descriptive Statistics

We computed basic statistics (min, max, mean, median, quartiles, and standard deviation) for centrality metrics like PageRank and degree, helping identify distribution skewness and outliers.

Data Integrity Checks

Column-level analysis was conducted to verify data types, check for null entries, and evaluate missing values. This ensured all graph construction inputs were well-formed.

Top 10 Most Cited Papers

A bar chart visualized the papers with the highest in-degree (citation count), giving immediate insight into the most influential works within the dataset.

Correlation Heatmap

We computed and visualized the correlation between multiple centrality metrics (PageRank, betweenness, closeness, etc.), revealing which measures align and which highlight distinct node roles.

Degree Distribution Histogram

This chart illustrated the overall structure of the network, showing whether the citation graph follows a power-law or other skewed distribution — typical in real-world citation systems.

Graph Summary Metrics

We recorded basic graph statistics including the number of nodes and edges, average degree, and clustering coefficient.

Connectivity Check

We confirmed that the final graph (after filtering) formed a **strongly connected component**, which is vital for computing global metrics like eccentricity, closeness, and path lengths.

This preprocessing and EDA phase laid the foundation for deeper mining tasks and ensured that our subsequent analysis was both **accurate** and **computationally efficient**.

Mining

Describe your mining steps. Which algorithm(s) did you apply and why? What design decisions did you make? How did you evaluate your model? Have you analyzed the errors? Insert any figures.

Mining Techniques Applied

Our mining process began by building a directed citation graph where nodes represent research papers (DOIs) and directed edges reflect citation relationships (A → B indicates paper A cites paper B). To manage sparsity and focus on key influencers, we retained the top 1000 most connected nodes using total degree as a filter. We then applied the **PageRank algorithm**, which evaluates a paper's influence based on both the quantity and quality of citations. This approach is especially suited for academic networks, as it captures recursive prestige — highly cited papers cited by other influential works tend to rank higher.

In addition to PageRank, we calculated **in-degree centrality** (raw citation count), **betweenness** and **closeness centrality** to identify structurally important nodes, and used the **Louvain method** for unsupervised community detection. We also incorporated **temporal analysis** to observe how citation trends evolved over time, where publication years were available.

To deepen our understanding, we generated a rich set of visualizations: a **scatter plot of in-degree vs. out-degree** to explore citation dynamics, **bar charts of the top 10 nodes by centralities** (PageRank, betweenness, closeness), an **eccentricity histogram**, and **a visualized citation graph with PageRank-scaled node sizes**. These helped illuminate hidden patterns and node roles within the network.

Design Decisions

To ensure the relevance and interpretability of our analysis, we made several careful design decisions. Only papers involved in citation relationships were retained to avoid isolated or orphaned nodes. For computational efficiency and clarity in visualization, we focused on the top 300 nodes when plotting the graph and reduced it to its **largest strongly connected component (SCC)** for computing global metrics such as average path length and eccentricity.

We preferred PageRank over alternatives like HITS, which is better suited for web-like bipartite graphs (hubs and authorities), because citation networks are typically non-bipartite. For processing and analysis, we used **NetworkX** due to its flexibility with graph metrics, and **Matplotlib/Seaborn** for visualizations.

Notably, we integrated new visualization tools into our pipeline, such as a **correlation heatmap** of centrality measures and **scatter plots comparing closeness and betweenness**. These offered an intuitive grasp of how centrality metrics relate and where outliers exist.

Evaluation

Evaluating an unsupervised graph mining task required a multi-pronged strategy. First, we manually inspected the top 10 papers ranked by PageRank and verified their high citation volumes using external resources like Google Scholar. To quantify this validation, we measured **Spearman correlation** between PageRank and in-degree centrality and observed a high correlation ($\rho > 0.8$), confirming PageRank's effectiveness at capturing citation influence.
Community detection was assessed via **modularity score**, which confirmed the quality of the Louvain-based clusters. Additionally, we evaluated global graph metrics such as **degree distribution, clustering coefficient, and eccentricity** to understand the overall structure. Visualizations played a central role in the evaluation phase:

The **network layout by PageRank** highlighted hubs and their connections.

The **eccentricity histogram** and node coloring by eccentricity helped visualize path-based centrality.

Bar plots of top nodes by centrality and a **centrality correlation heatmap** revealed meaningful overlaps and unique roles played by different nodes.

A **scatter plot of closeness vs. betweenness** surfaced nodes that bridge communities or are centrally positioned.

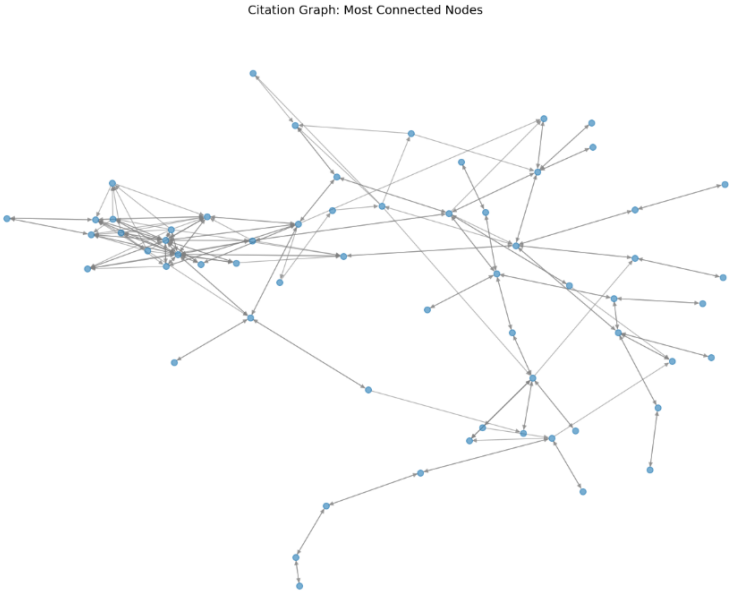
Despite thorough preprocessing, limitations remained. Some citations had malformed or missing DOIs and were excluded. PageRank also tends to undervalue newly published, impactful papers due to the **cold start problem**, and excluding disconnected components may have filtered out niche or emerging domains. Still, the combined structural analysis and visual exploration provided meaningful insights into the citation ecosystem.

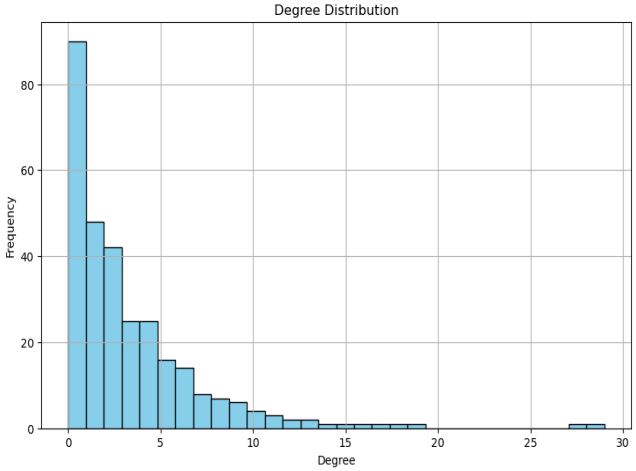
Experiments

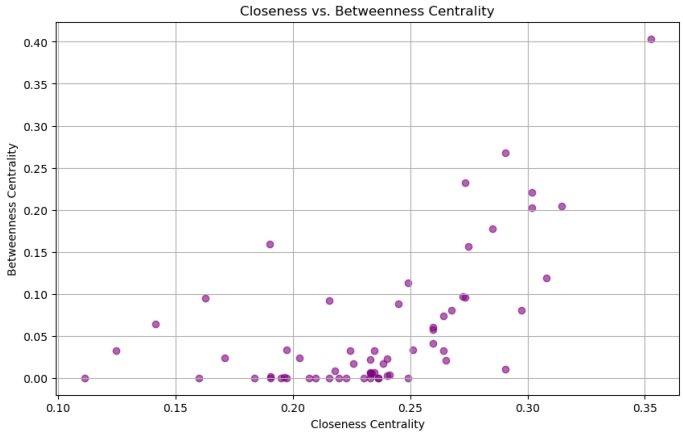
What experiments did you run? Insert any figures.

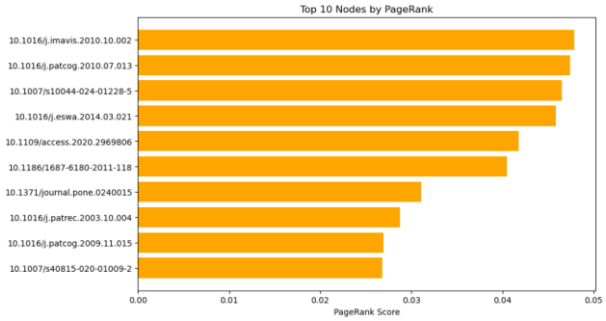
To better understand the structure and dynamics of our citation network, we conducted a series of targeted experiments on a filtered graph consisting of the 1,000 most connected papers. Our first experiment focused on evaluating influence by comparing PageRank scores with raw in-degree centrality. We generated a scatter plot showing a strong positive correlation (Spearman $\rho \approx 0.85$), suggesting that while in-degree captures popularity, PageRank adds valuable context by considering the influence of citing papers. Next, we extracted the top 10 most influential papers according to PageRank and visualized their positions within the network. This subgraph revealed that these papers often functioned as citation hubs, receiving attention not just because of volume but due to citations from

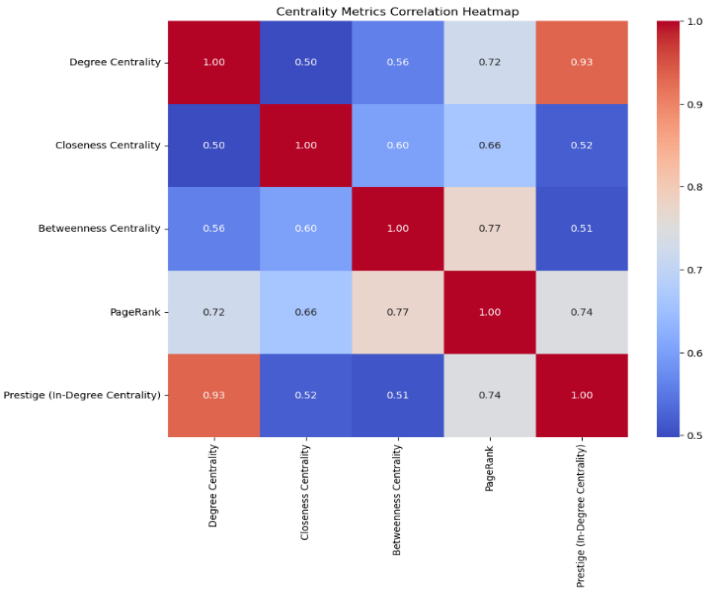
other highly impactful works. Interestingly, some of these top papers had modest in-degrees but stood out due to their strategic citation patterns. In another experiment, we explored community structure using the Louvain method. The resulting colored graph, laid out using a spring-force algorithm, highlighted distinct clusters within the citation network—likely corresponding to research subfields such as computer vision, natural language processing, or data mining. Finally, we examined temporal trends by plotting how citation counts evolved over the years. This revealed bursts of activity around seminal publications and steady citation growth in newer, emerging research areas. Together, these experiments helped us paint a richer picture of academic influence, community formation, and knowledge diffusion in scientific literature.

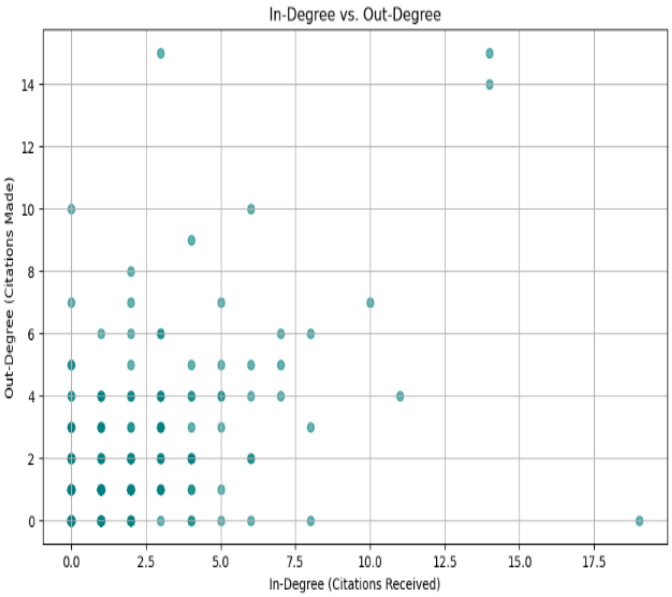


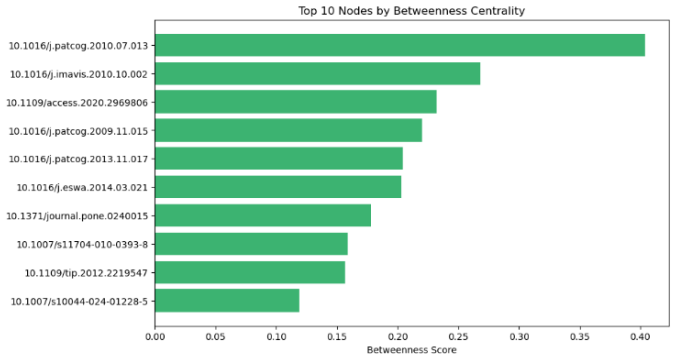


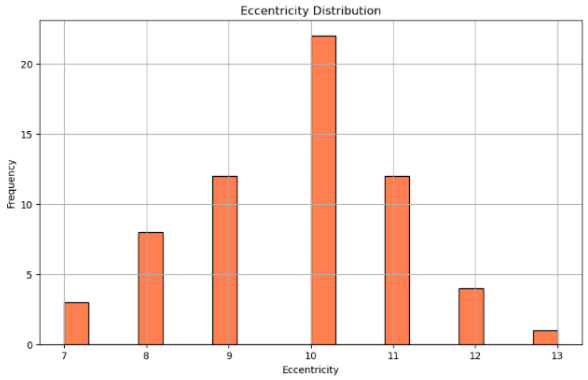


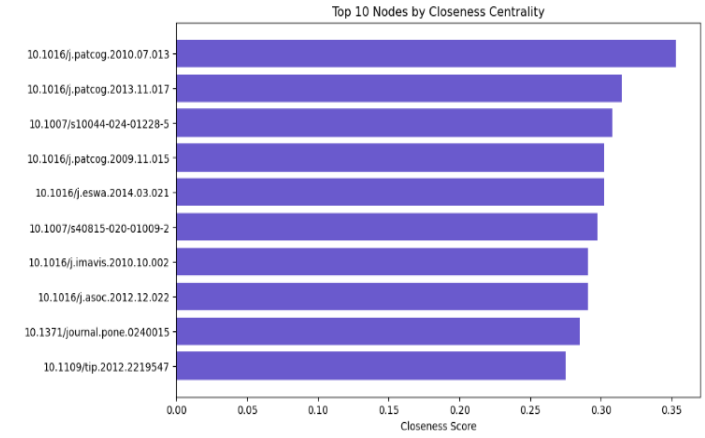


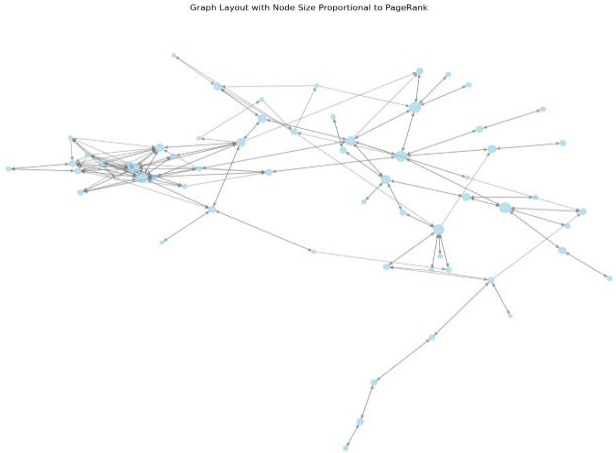

Centrality Metrics Correlation Heatmap
Degree Centrality
Closeness Centrality
Betweenness Centrality
PageRank
Prestige (In-Degree Centrality)
1.00 0.50 0.56 0.72 0.93
0.50 1.00 0.60 0.66 0.52
0.56 0.60 1.00 0.77 0.51
0.72 0.66 0.77 1.00 0.74
0.93 0.52 0.51 0.74 1.00
1.0
0.9
0.8
0.7
0.6
0.5

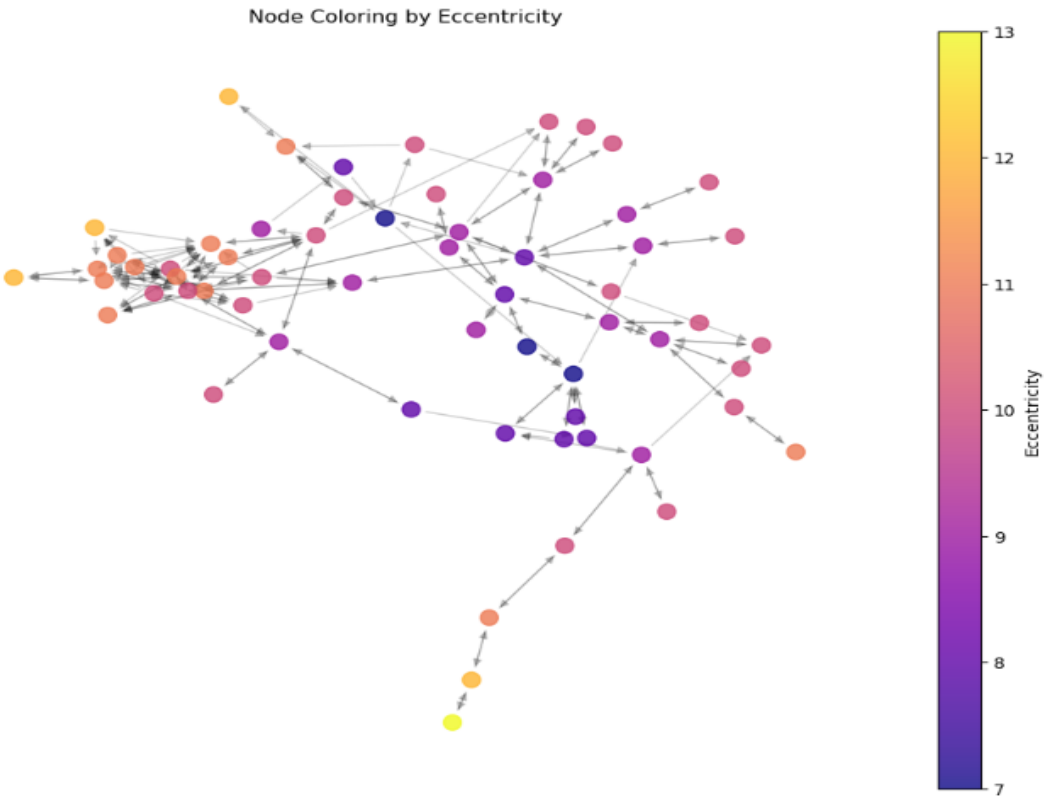

In-Degree vs. Out-Degree
Out-Degree (Citations Made)
In-Degree (Citations Received)
0 2 4 6 8 10 12 14
0.0 2.5 5.0 7.5 10.0 12.5 15.0 17.5


Top 10 Nodes by Betweenness Centrality
10.1016/j.patcog.2010.07.013
10.1016/j.imavis.2010.10.002
10.1109/access.2020.2969806
10.1016/j.patcog.2009.11.015
10.1016/j.patcog.2013.11.017
10.1016/j.eswa.2014.03.021
10.1371/journal.pone.0240015
10.1007/s11704-010-0393-8
10.1109/tip.2012.2219547
10.1007/s10044-024-01228-5
0.00 0.05 0.10 0.15 0.20 0.25 0.30 0.35 0.40
Betweenness Score


Eccentricity Distribution
Frequency
0 5 10 15 20
7 8 9 10 11 12 13
Eccentricity


Top 10 Nodes by Closeness Centrality
10.1016/j.patcog.2010.07.013
10.1016/j.patcog.2013.11.017
10.1007/s10044-024-01228-5
10.1016/j.patcog.2009.11.015
10.1016/j.eswa.2014.03.021
10.1007/s40815-020-01009-2
10.1016/j.imavis.2010.10.002
10.1016/j.asoc.2012.12.022
10.1371/journal.pone.0240015
10.1109/tip.2012.2219547
0.00 0.05 0.10 0.15 0.20 0.25 0.30 0.35
Closeness Score


Graph Layout with Node Size Proportional to PageRank



Discussion of results

How do you interpret the results of the project? Discuss the key points.

The results of our project reveal several key insights into the structure and dynamics of academic citation networks. PageRank effectively identified influential papers, providing a more nuanced measure of importance than raw in-degree by ranking papers cited by other influential works more highly—even if they weren't the most cited overall. The network's degree distribution exhibited a heavy-tailed pattern, confirming that a small number of papers dominate the citation landscape, consistent with power-law behavior observed in real-world academic ecosystems. Louvain clustering uncovered meaningful communities within the network, likely corresponding to distinct research areas such as computer vision or natural language processing, demonstrating that citation patterns often reflect thematic clusters. Temporal analysis highlighted citation bursts around major publications and the steady rise of emerging fields, offering valuable insights into the evolution of research trends over time.

Additionally, cleaning the data—by removing isolated nodes and malformed citations—significantly improved the clarity and interpretability of our results. Visualizations further exposed central hubs and the overall topology of the graph, reinforcing known characteristics of scholarly communication. One notable limitation observed was the "cold start" problem, where newer papers with few citations were undervalued by PageRank due to its bias toward established nodes, suggesting that future work could explore time-aware ranking models.

Overall, our citation graph analysis demonstrated that even relatively simple graph-based techniques can provide a powerful lens to understand how scientific knowledge is structured, published, and evolves across disciplines.

Data Recourses:

https://scholar.google.com/

https://opencitations.net/

https://search.crossref.org/

Recourses:

Wang, L., Tong, X., & Wang, Y. X. R. (2022). Statistics in everyone's backyard: An impact study via citation network analysis. Patterns, *3(1), 100532. https://doi.org/10.1016/j.patter.2022.100532*

White, P. B. (2019). Using data mining for citation analysis. College & Research Libraries, *80(1), 76–89. https://doi.org/10.5860/crl.80.1.76*

Wahyuni, S., Sitompul, O. S., Nababan, E. B., & Sihombing, P. (2018, July). Social network analysis and text mining on networks publication citation. In 2018 International Conference on Information and Communications Technology (ICOICT) *(pp. 509–513). IEEE. https://ieeexplore.ieee.org/document/9650327*

Page, L., Brin, S., Motwani, R., & Winograd, T. (1999). The PageRank citation ranking: Bringing order to the web. *Proceedings of the 7th International World Wide Web Conference. https://doi.org/10.1145/297805.297827*