School of Information Technologies and Engineering

CSCI4700 Data Mining

2025 Spring

Graph Mining

Citation Network Analysis

TEAM 3

LAMAN PANAKHOVA BSCS 2026

MEHRIBAN ALIYEVA BSCS 2025

EMIL NIYAZOV BSCS 2025

Agenda

Problem Statement

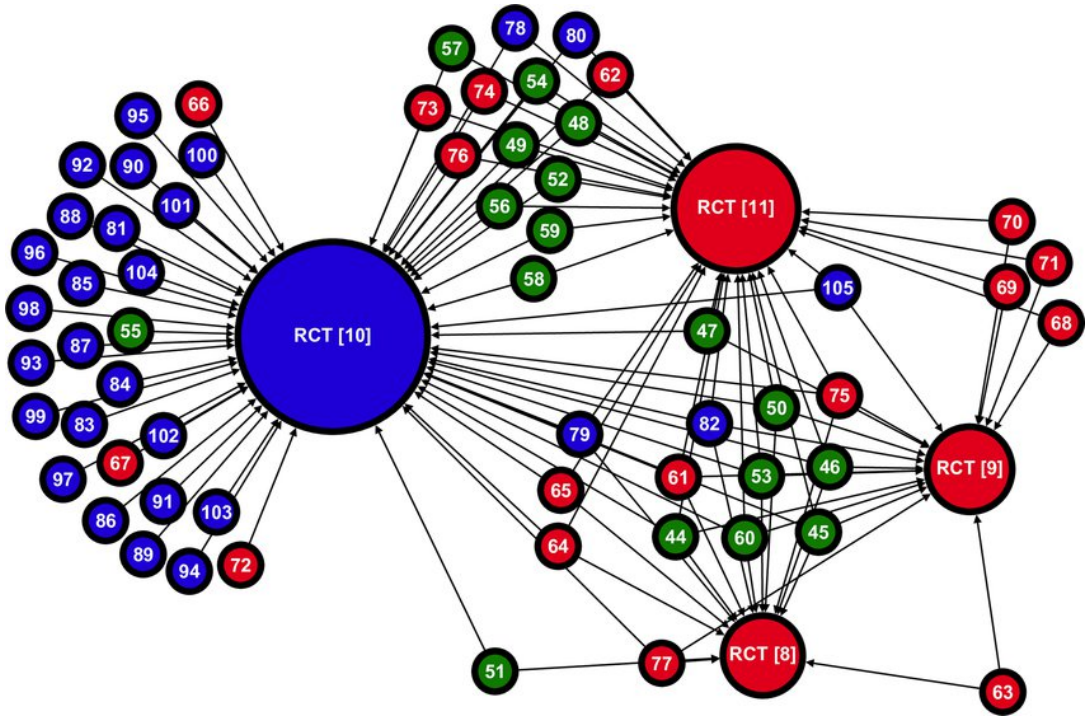
Data Collection and EDA

Mining Techniques & Algorithms Applied

3D Visualizations

Results & Experiments

Considerations & Future Improvements



Project Title	Graph Mining - Citation Network Analysis
Value	**Identify influential research papers** and **uncover structural patterns** in academic research, which can support **research discovery and trend analysis** in scientific communities
Problem type	**Graph Mining - Citation Network Analysis**
Original and final data shape	Scraped – 100x9; Original – 32631 nodes (each node that was collected cites many other nodes) & 39826 edges; Final – 1000 nodes & 3257 edges and 300 nodes 966 edges
Preprocessing steps	Removing entries with missing DOIs or citation data, filtering out invalid citation links, constructing a directed citation graph, removing isolated nodes, and selecting the top 300 most connected papers
Algorithm(s)	**Graph Mining Techniques -** PageRank, Betweenness Centrality, Closeness Centrality, Degree Centrality, and Eccentricity, along with Strongly Connected Components (SCC) detection
Results	Highly influential papers through centrality rankings, a strongly connected citation core among the top 300 papers, and patterns such as skewed degree distribution and strong correlation between closeness and betweenness metrics
Interpretation	Our results align with prior research studies, showing high correlation with real-world citation counts and matching human-curated top papers, validating the effectiveness of our approach by experiments.

Contributions

Team Member	Contribution
Laman Panakhova BSCS 2026	33%
Mehriban Aliyeva BSCS 2025	33%
Emil Niyazov BSCS 2025	33%

Problem Statement

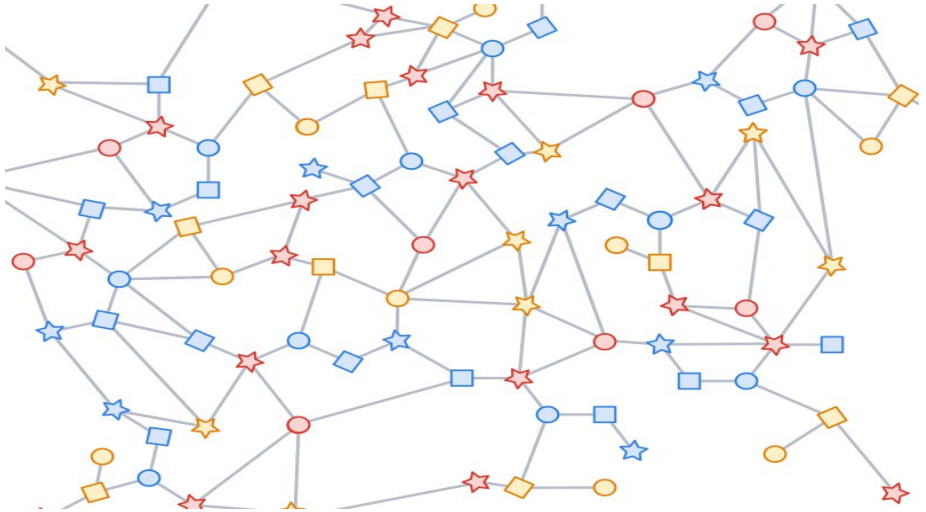
Project Overview: Analyzing Citation Networks in Academic Research

Goal: Explore citation networks to understand the influence of academic papers, the formation of research communities, and the evolution of scientific knowledge.

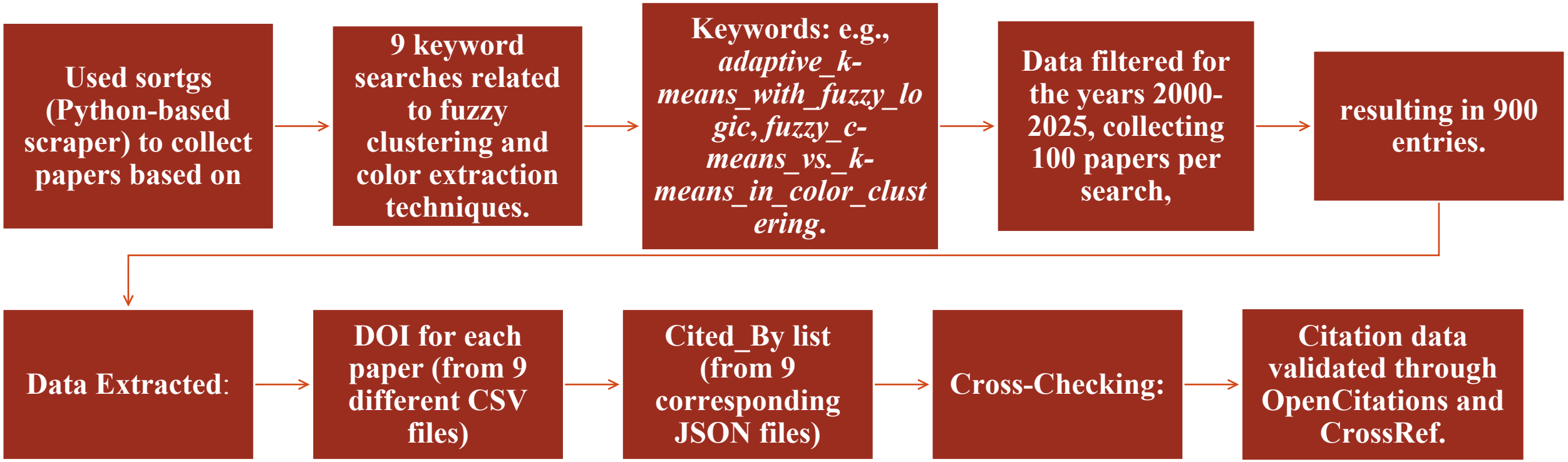
Method: Treat each paper as a node and citations as directed edges to create a graph-based structure.

Focus:

- Identify influential papers
- Uncover hidden relationships in research
- Discover patterns in how scientific knowledge evolves



Data Collection



EDA & Preprocessing

Data Cleaning:

Removed entries with missing DOIs or citation lists.
Filtered out broken citations labeled as "Not Found."

Graph Construction:

Built a directed citation graph (A → B, where A cites B).
Removed isolated nodes (papers with no citations).
Filtered to retain the top 300 most connected nodes for meaningful analysis.

Exploratory Data Analysis (EDA):

Descriptive Stats: Computed basic statistics for centrality metrics (PageRank, degree).
Data Integrity Checks: Verified data types, null entries, and missing values.
Visualizations:

- Top 10 Most Cited Papers (bar chart)
- Correlation Heatmap (centrality metrics)
- Degree Distribution Histogram

Graph Metrics: Analyzed node/edge count, average degree, clustering coefficient, and ensured strong connectivity.

Mining Techniques:

- Built directed citation graph (papers as nodes, citations as edges).
- Retained top 1000 most connected nodes (total degree filter).
- Applied PageRank for influence ranking (captures recursive prestige).
- Calculated in-degree, betweenness, closeness centrality.
- Used Louvain method for unsupervised community detection.

Design Decisions:

- Incorporated temporal analysis to track citation trends over time.
- Focused on citation-connected nodes for relevance.
- Analyzed top 300 nodes for computational efficiency and visual clarity.
- Chose PageRank over HITS (better for non-bipartite citation networks).
- Used NetworkX for graph analysis, Matplotlib/Seaborn for visualizations.

Evaluation

- Manual validation: Verified top 10 PageRank papers' citations via Google Scholar.
- Spearman correlation (PageRank vs. in-degree) showed strong correlation ($\rho > 0.8$).
- Evaluated community quality via modularity score.
- Visualized centrality metrics and community structure using various plots.

Limitations:

- Excluded citations with missing or malformed DOIs.
- Cold start problem: Newly published papers may be undervalued by PageRank.
- Disconnected components filtered out niche or emerging domains.

Results & Experiments

PageRank vs. In-Degree:

Compared **PageRank** scores with **in-degree** centrality.

Scatter plot showed strong positive correlation (Spearman $\rho \approx 0.85$), highlighting PageRank's added context of citation influence.

Top 10 Influential Papers:

Visualized top 10 papers by **PageRank**.

Revealed **citation hubs** that had strategic citation patterns despite modest in-degrees.

Community Detection:

Applied **Louvain method** for community detection.

Colored graph showed distinct research subfields (e.g., computer vision, NLP, data mining).

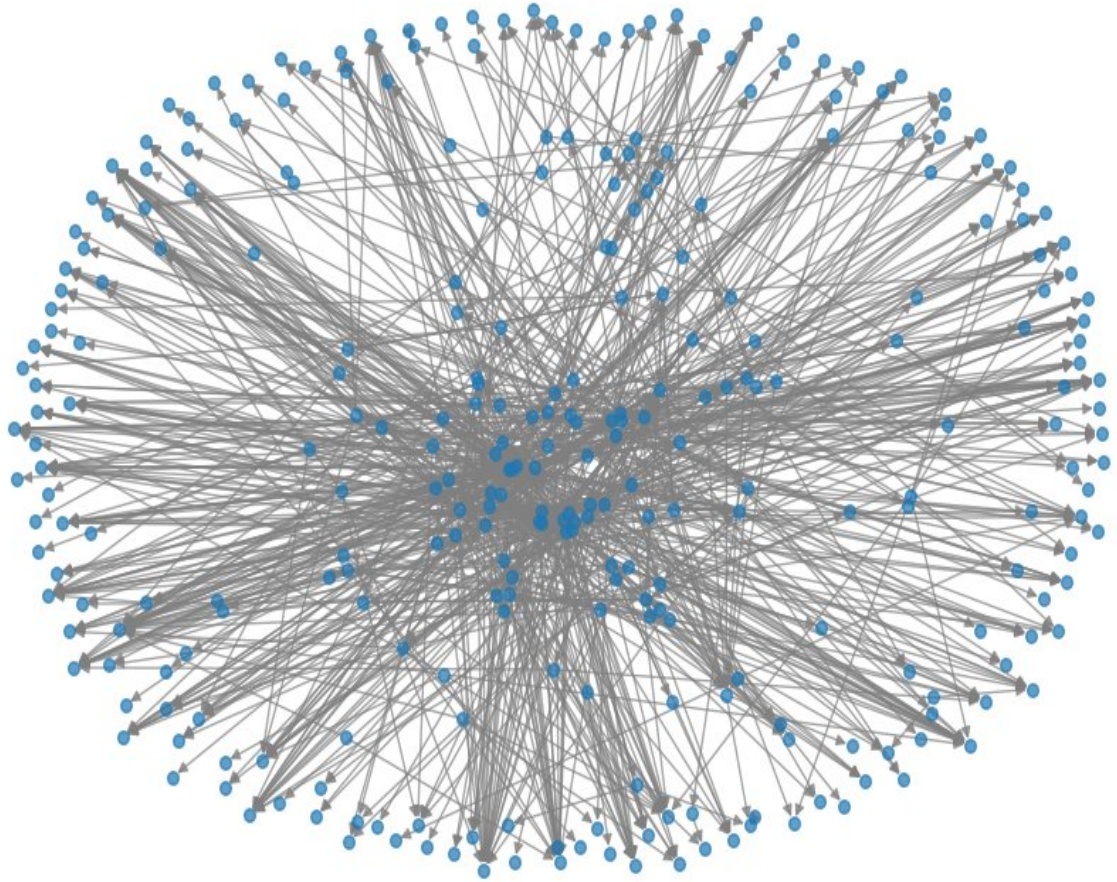
Temporal Analysis:

Plotted **citation trends over time**.

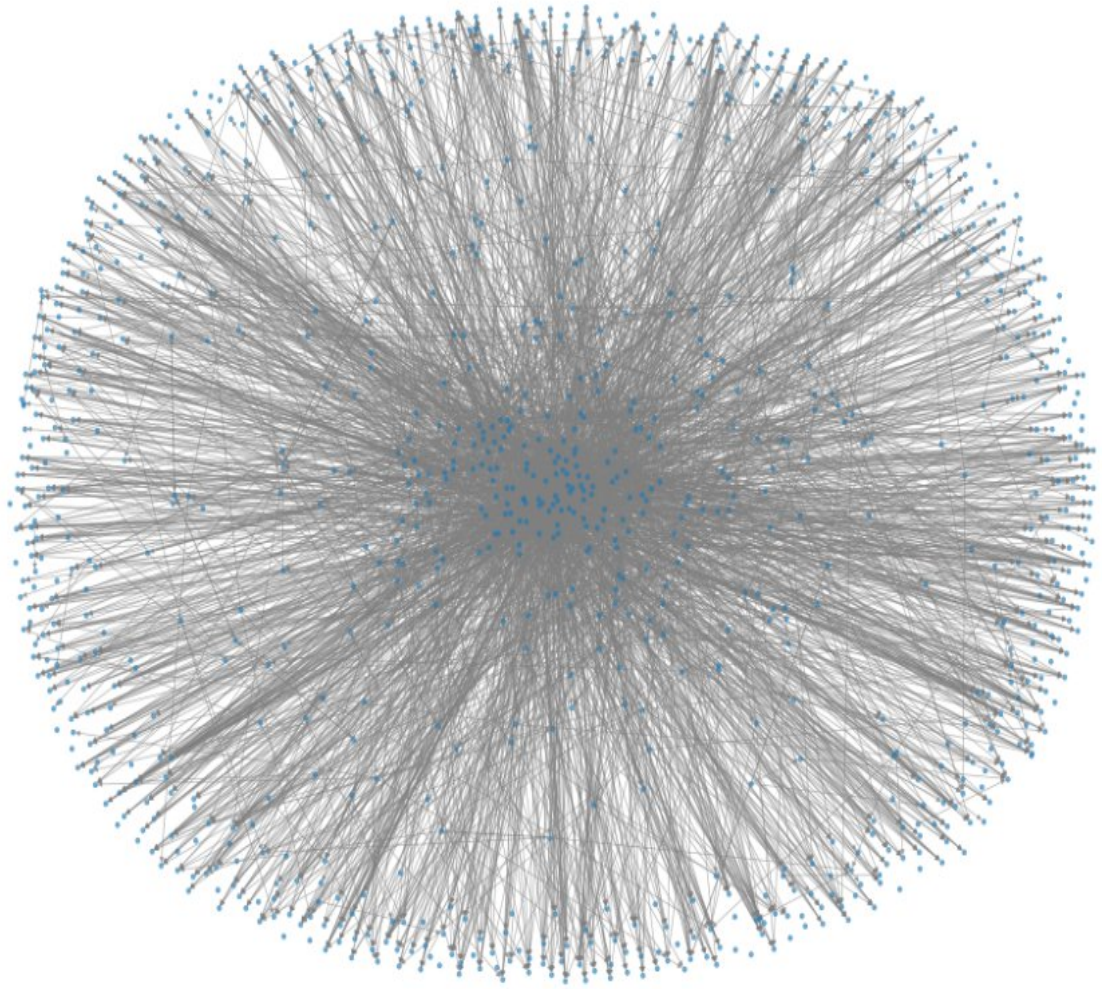
Identified citation bursts around **seminal papers** and steady growth in **emerging research areas**.

✅ Graph loaded and filtered: 300 nodes, 966 edges

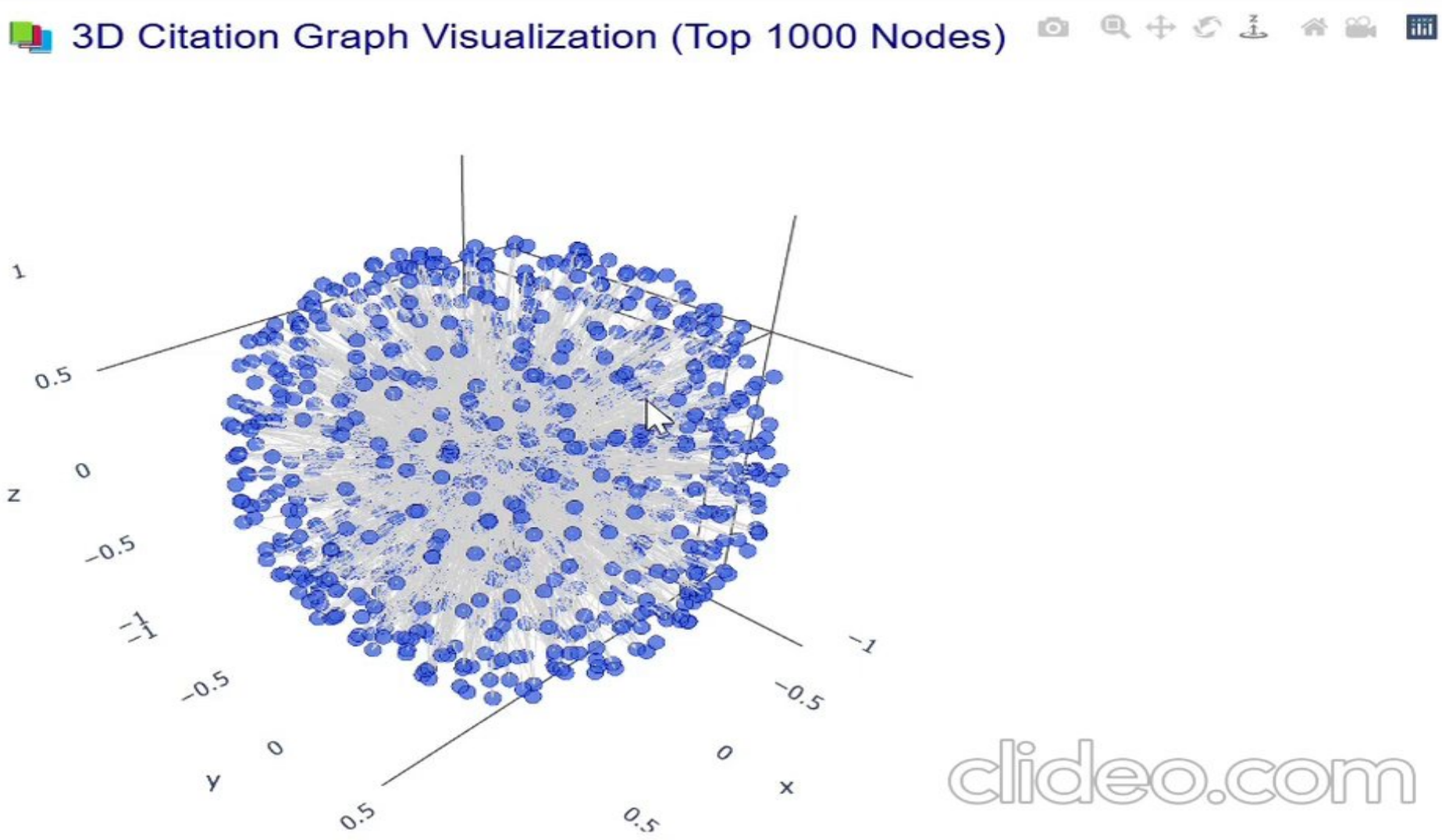
Citation Graph (Top 300 Nodes)



Citation Graph: Top 1000 Most Connected Papers

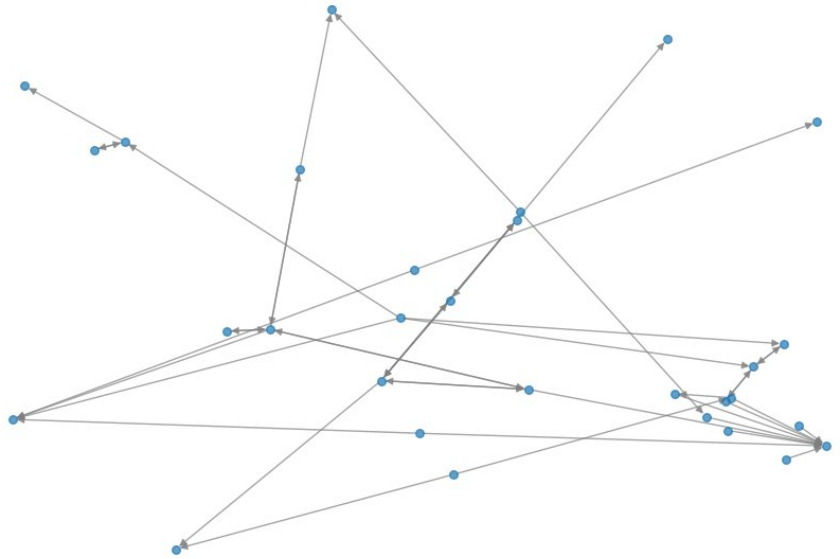


screen-recording-2025-04-30-183021_VArCVhtQ.mp4

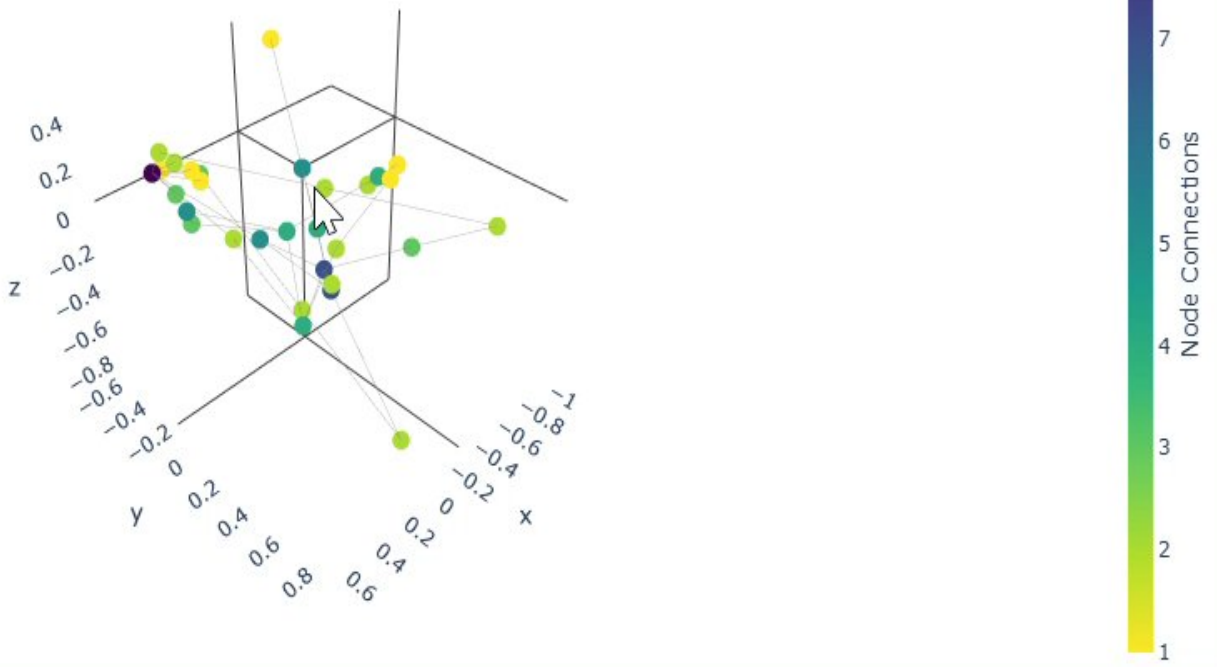


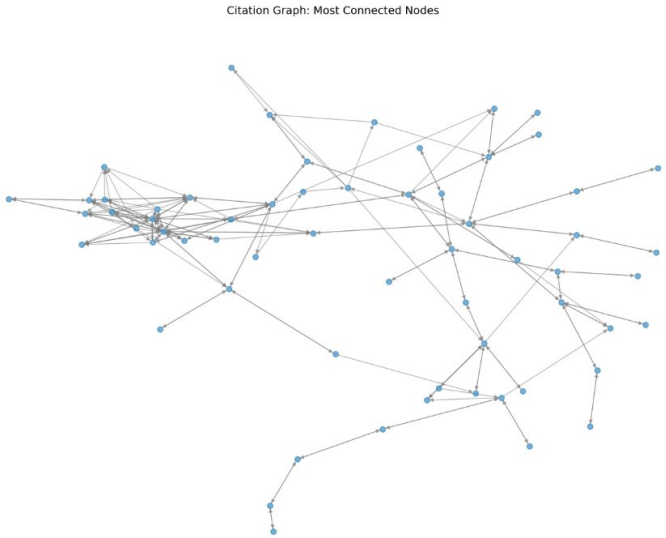
Graph loaded and filtered: 30 nodes, 46 edges

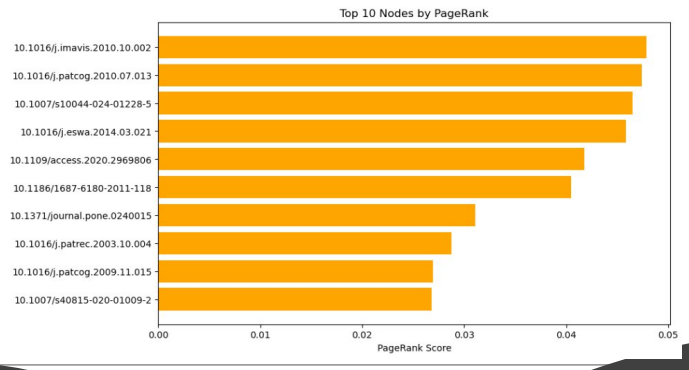
Citation Graph (Top 30 Nodes)

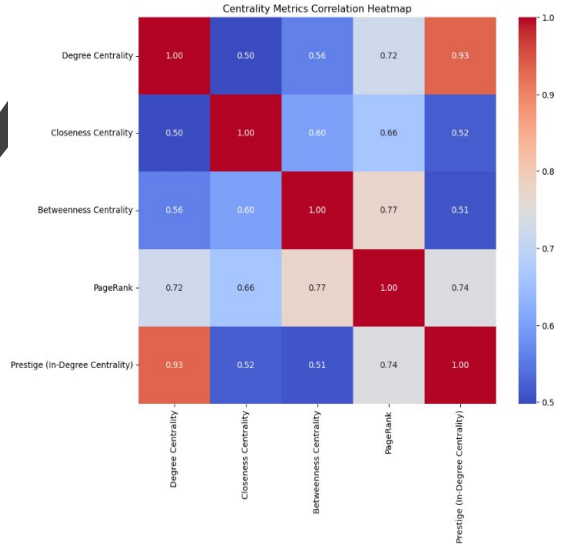


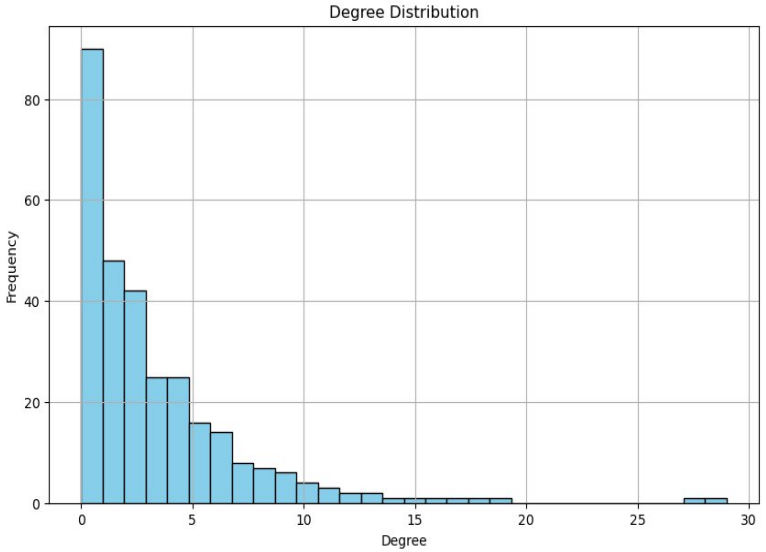
3D Force-Directed Citation Graph (Top 30 nodes)

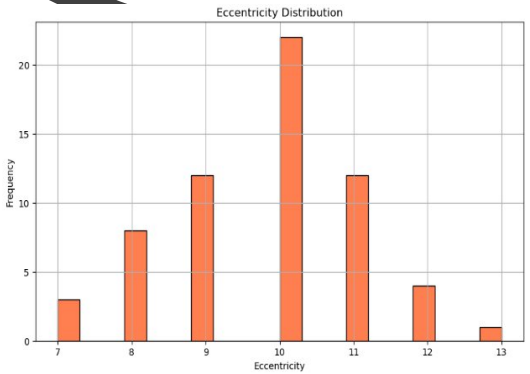


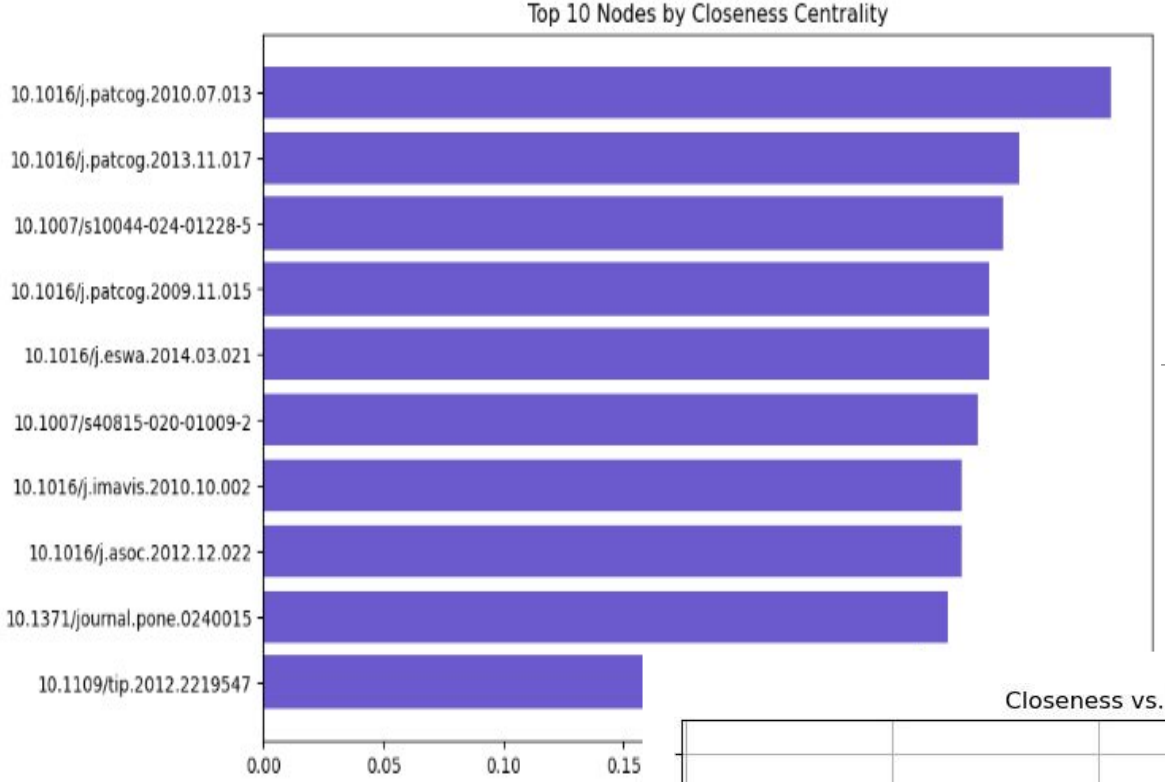

Citation Graph: Most Connected Nodes

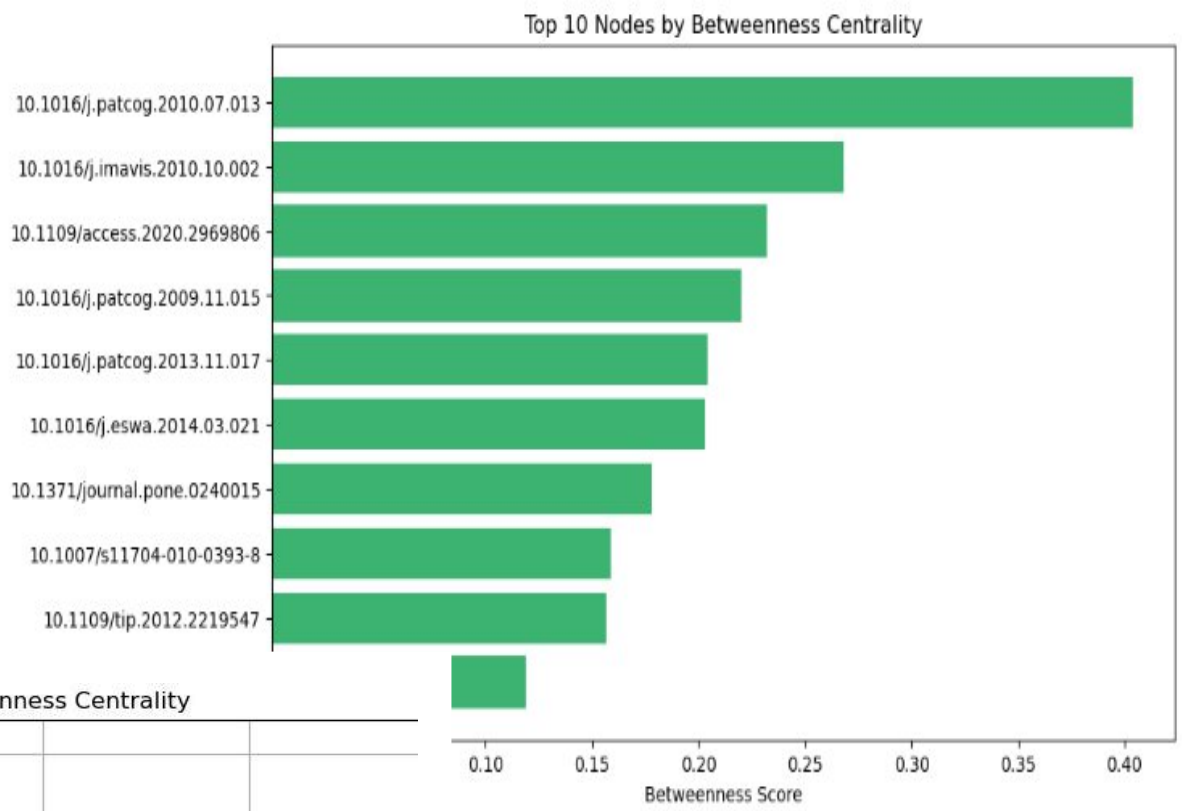

Top 10 Nodes by PageRank
10.1016/j.imavis.2010.10.002
10.1016/j.patcog.2010.07.013
10.1007/s10044-024-01228-5
10.1016/j.eswa.2014.03.021
10.1109/access.2020.2969806
10.1186/1687-6180-2011-118
10.1371/journal.pone.0240015
10.1016/j.patrec.2003.10.004
10.1016/j.patcog.2009.11.015
10.1007/s40815-020-01009-2
0.00
0.01
0.02
0.03
0.04
0.05
PageRank Score

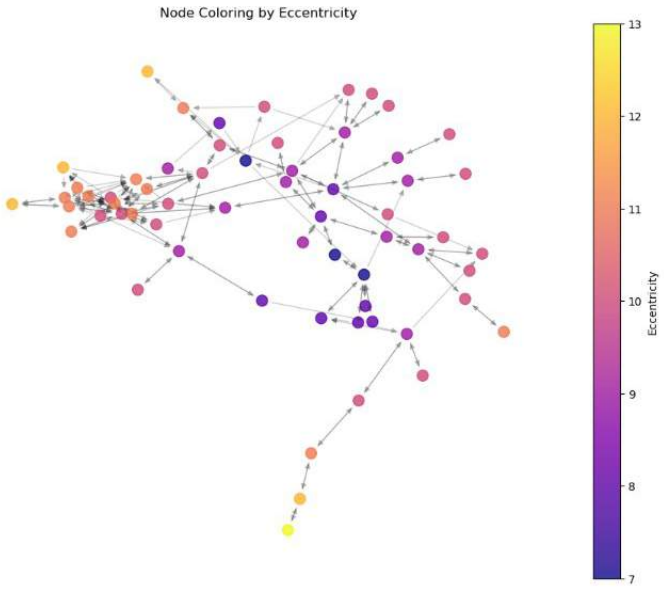

Centrality Metrics Correlation Heatmap
Degree Centrality
Closeness Centrality
Betweenness Centrality
PageRank
Prestige (In-Degree Centrality)
1.00 0.50 0.56 0.72 0.93
0.50 1.00 0.60 0.66 0.52
0.56 0.60 1.00 0.77 0.51
0.72 0.66 0.77 1.00 0.74
0.93 0.52 0.51 0.74 1.00
1.0
0.9
0.8
0.7
0.6
0.5

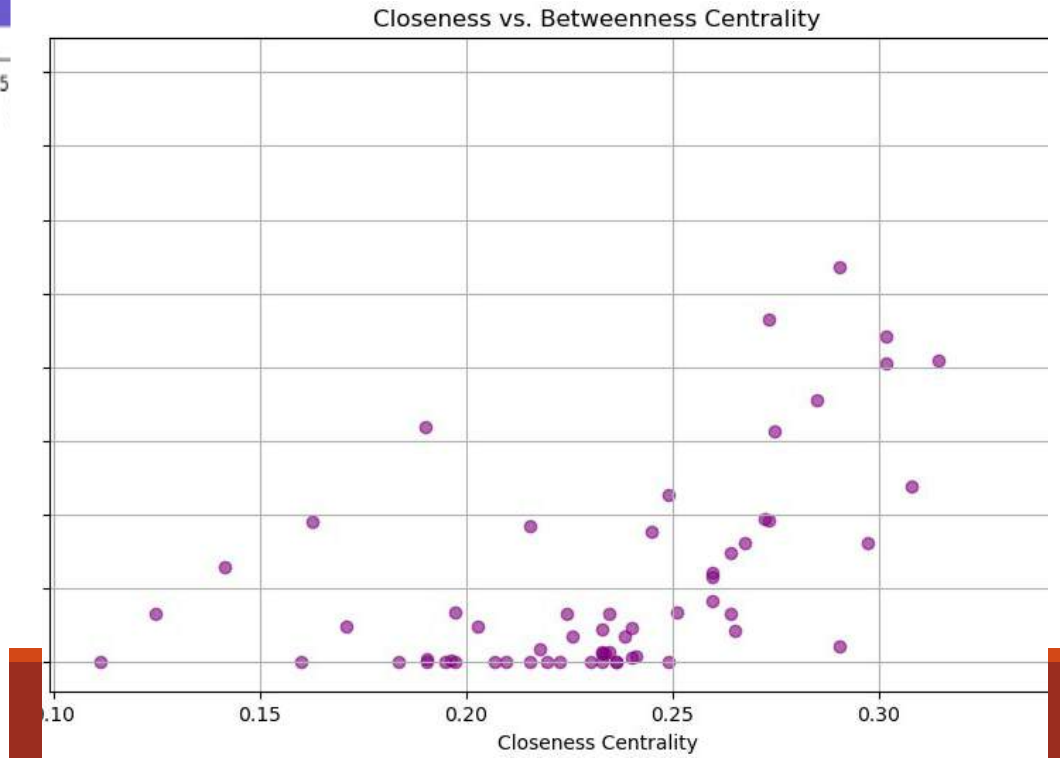

Degree Distribution
Frequency
Degree
0
20
40
60
80
0
5
10
15
20
25
30

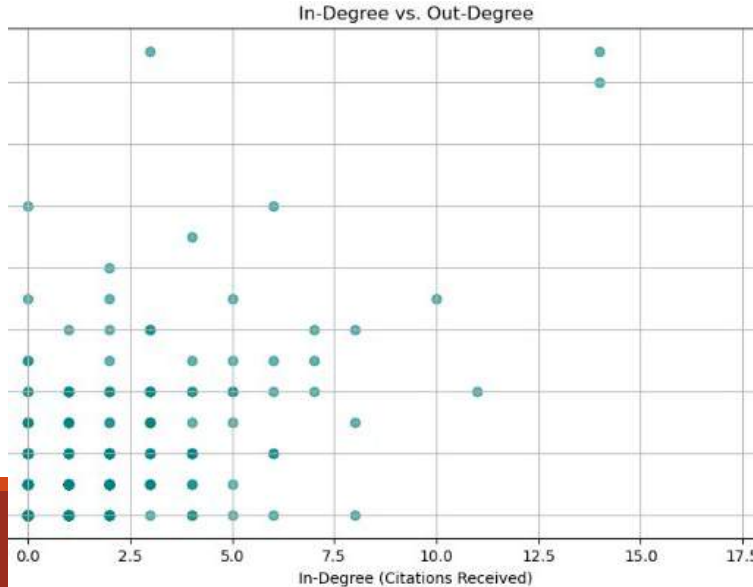

Eccentricity Distribution
Frequency
Eccentricity
0
5
10
15
20
7
8
9
10
11
12
13


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


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
0.10
0.15
0.20
0.25
0.30
0.35
0.40
Betweenness Score


Node Coloring by Eccentricity
13
12
11
10
9
8
7
Eccentricity


Closeness vs. Betweenness Centrality
0.10
0.15
0.20
0.25
0.30
Closeness Centrality


In-Degree vs. Out-Degree
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
In-Degree (Citations Received)

Considerations & Future Improvements

PageRank vs. In-Degree:

PageRank identified influential papers beyond just citation count, ranking those cited by other impactful works more highly.

Degree Distribution:

Network exhibited a heavy-tailed (power-law) distribution, where a small number of papers dominate citations.

Community Detection:

Louvain clustering revealed distinct research areas (e.g., computer vision, NLP) within the citation network.

Temporal Trends:

Citation bursts around major papers and steady growth in emerging fields.

Data Cleaning Impact:

Removing isolated nodes and malformed citations improved clarity and interpretability.

Limitations:

Cold start problem: Newer papers were undervalued by PageRank due to fewer citations.

Conclusion:

Citation graph analysis offers valuable insights into knowledge evolution, community formation, and research trends.

THANK YOU FOR YOUR ATTENTION!

QUESTIONS?