Citation Network Analysis

Graduate Karna Hai (ग्रेजुएट करना है)
Ashutosh Gera (2021026) Farhan Ali (2021045)
Ritisha Singh (2021089) Shruti Jha (2021289)

CSE558: Data Science Project Midsem Evaluation

November 3, 2024

INDRAPRASTHA INSTITUTE *of*
INFORMATION TECHNOLOGY **DELHI**

Original Dataset Overview

The dataset comprises a citation network designed for research purposes, combining data from multiple sources like DBLP, ACM, and Microsoft Academic Graph (MAG).

Dataset Composition:

- **Number of Papers**: 4,894,081 research papers.
- **Citation Relationships**: 45,564,149 citations linking these papers.
- **Attributes:**
  - **Paper Details**: Includes title, publication year, abstract, and venue.
  - **Authorship Information**: Lists authors associated with each paper, facilitating collaboration analysis.
  - **Citation Network**: Each paper includes references to other works, forming a citation graph.

Problem Statement

The enormous growth of research publications has made it challenging for academic search engines to bring the most relevant papers against the given search query. Numerous solutions have been proposed over the years to improve the effectiveness of academic search, including exploiting query expansion and citation analysis.

Community detection is a task in network analysis which aims to find sets of tightly connected nodes that are loosely connected with other nodes outside of those sets. We aim to implement algorithms to uncover hidden structures within a citation network dataset, which depicts relationships between cited papers and the papers which cite those papers.

Challenges with JSON Format

- **Overview:** Each paper is a node, with citations forming directed edges, allowing analysis of citation patterns, publication trends, and thematic clusters across research topics.
- **Large File Size:** The JSON dataset was ***12.52GB***, making it memory-intensive and difficult to process directly.
- **Complex Structure:** JSON's hierarchical format required extensive parsing for each field, increasing processing time.
- **Lack of Tabular Structure:** JSON is less suited for analysis; CSV format with rows and columns is more accessible for machine learning and statistical tools.

Solution: Convert JSON file to a structured ***9GB*** CSV file with 21 columns for simplifying access and enabling efficient analysis.

Processed Dataset Overview

The dataset is a structured CSV file derived from a citation network. It captures metadata on research papers, including authors, venues, citations, and keywords. Each row represents a unique paper with 21 features across columns.

Key Features:

- **id**: Unique identifier for each paper.
- **title**: Title of the paper.
- **author_name**: Names of authors, separated by semicolons.
- **n_citation**: Total citations received by the paper.
- **references**: List of cited paper IDs, representing directed edges in the citation network.

Exploratory Data Analysis-I

Phase 1: Initial Exploration and Analysis

- **Data Overview:**
  - **Shape:** *4.89M rows, 21 columns*
  - **Key Columns:** `id`, `title`, `year`, `n_citation`, `reference_count`, etc.
  - **Missing Data:** Major gaps in `volume`, `issue`, and citation-related fields.
  - **Outlier Detection**:
    - Extreme values in `n_citation` and `reference_count` highlighted using box plots.
    - Identified outliers in numerical columns using the Interquartile Range (IQR) method.

Exploratory Data Analysis-II

Phase 2: Data Cleaning and Noise Reduction

- **Handling Missing Values:**
  - Filled with placeholders: `n_citation`: 0, `doc_type`: "Unknown", `references`: empty lists, `venue_name`: "Not Specified", `keyword`: "No Keywords", `publisher`: "Unknown Publisher"
- **Dropping Columns with High Missing Data:** Removed `volume`, `issue`, and `weight` to reduce noise.
- **Outlier Handling:**
  - Flagged columns with >5% outliers for handling.
  - Applied log transformation on `id` and `n_citation` to reduce skewness.
- **Visual Analysis:** Trends and distributions to be explored in following slides.

Visual Analysis-I

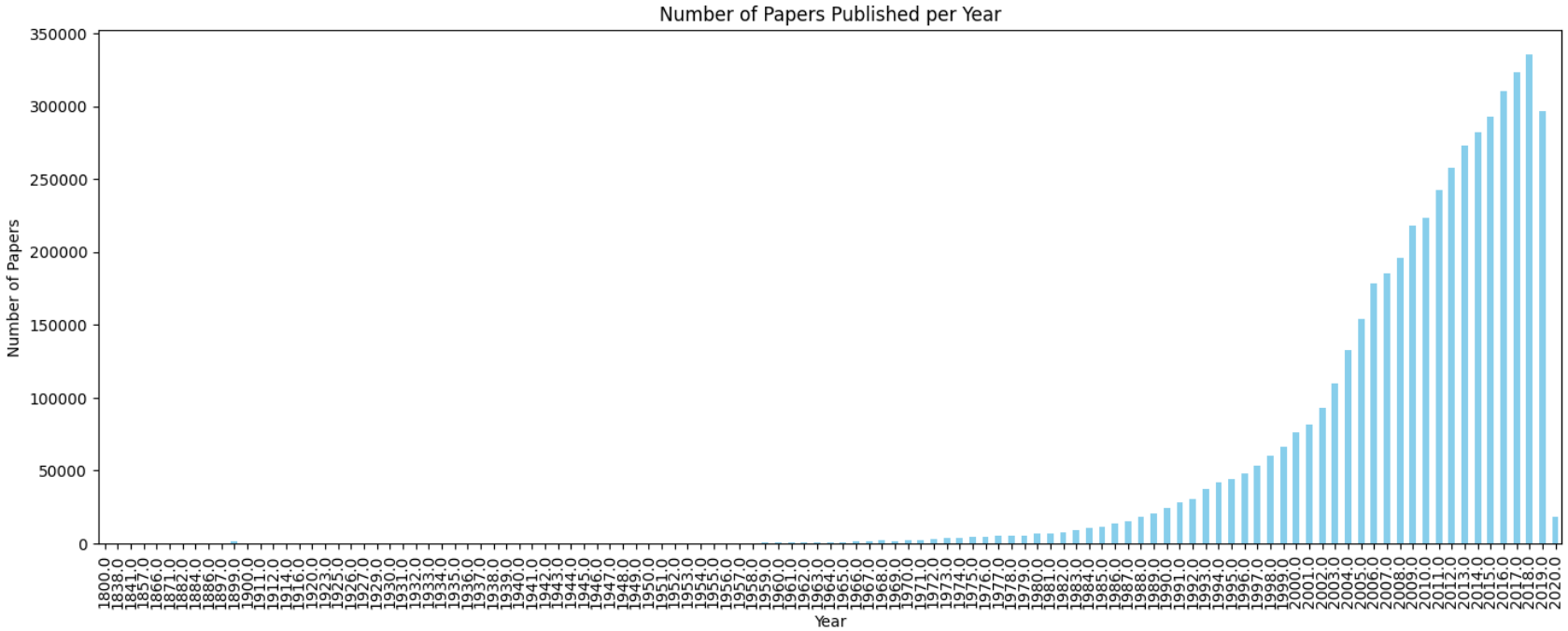

Figure 1: Number of Papers Published Per Year

Visual Analysis-II

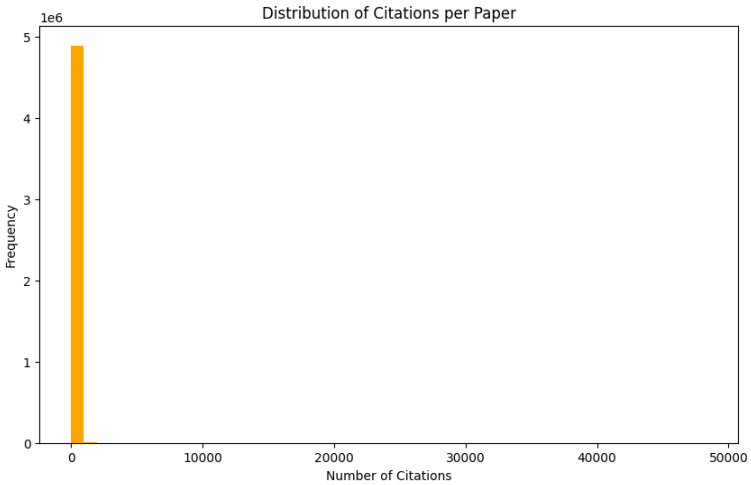

Figure 2: Citation Distribution Per Paper

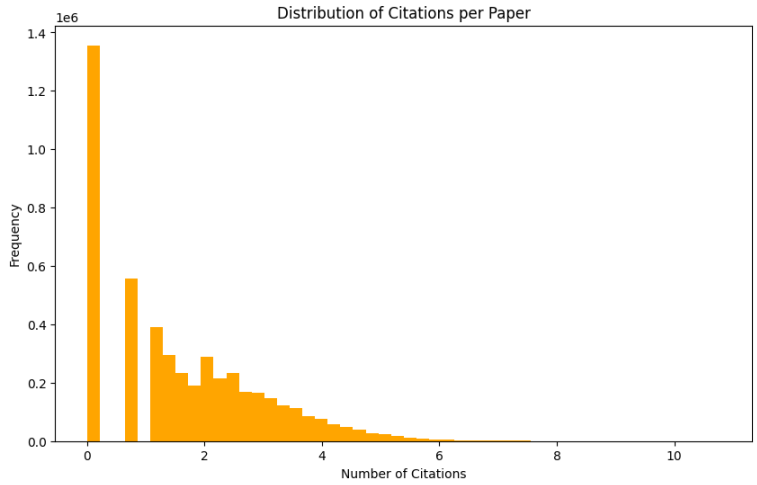

Figure 3: Citation Distribution Per Paper (log scale)

Visual Analysis-III

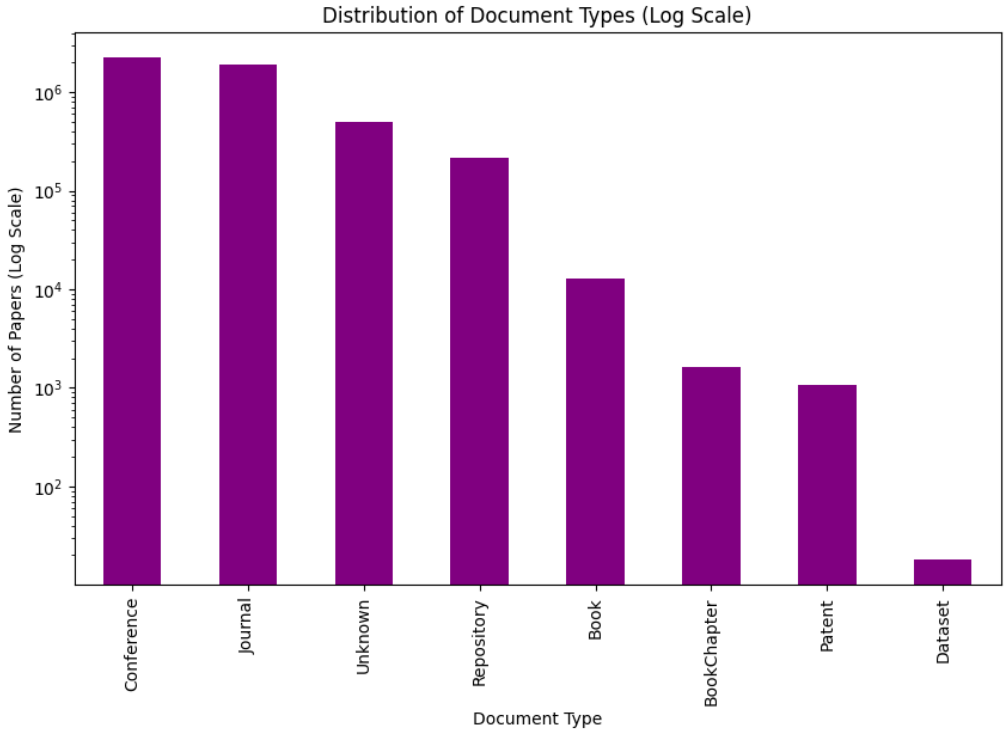

Figure 4: Distribution of Document Types on log scale

Visual Analysis-IV

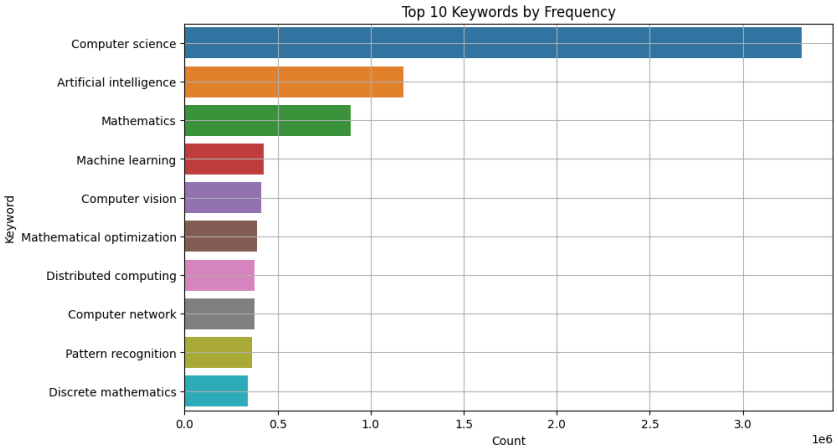

Figure 5: Most frequent Keywords

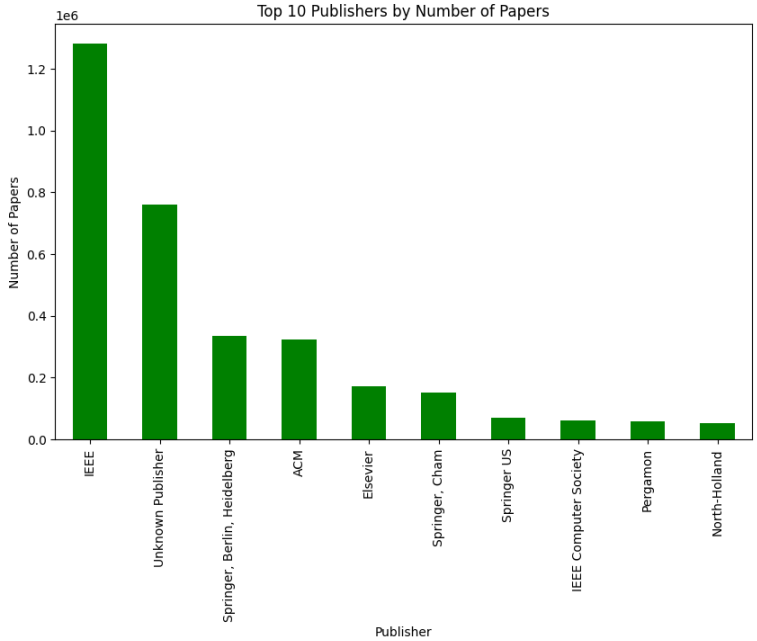

Figure 6: Most active Publishers

Visual Analysis-V

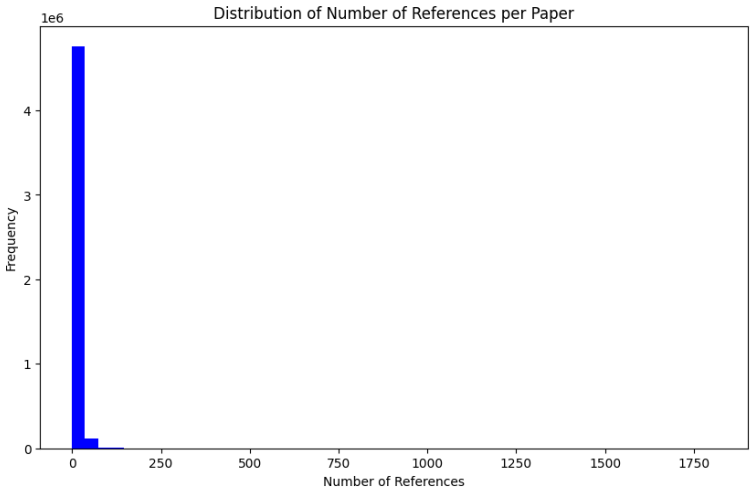

Figure 7: References Distribution Per Paper

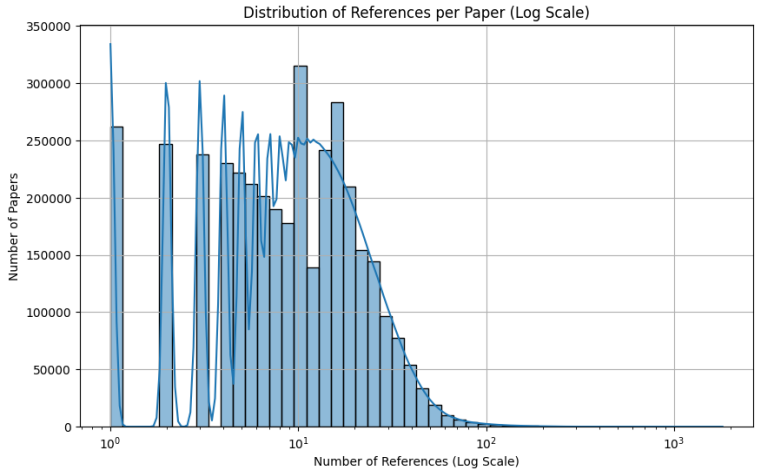

Figure 8: References Distribution Per Paper (log scale)

Correlation Analysis

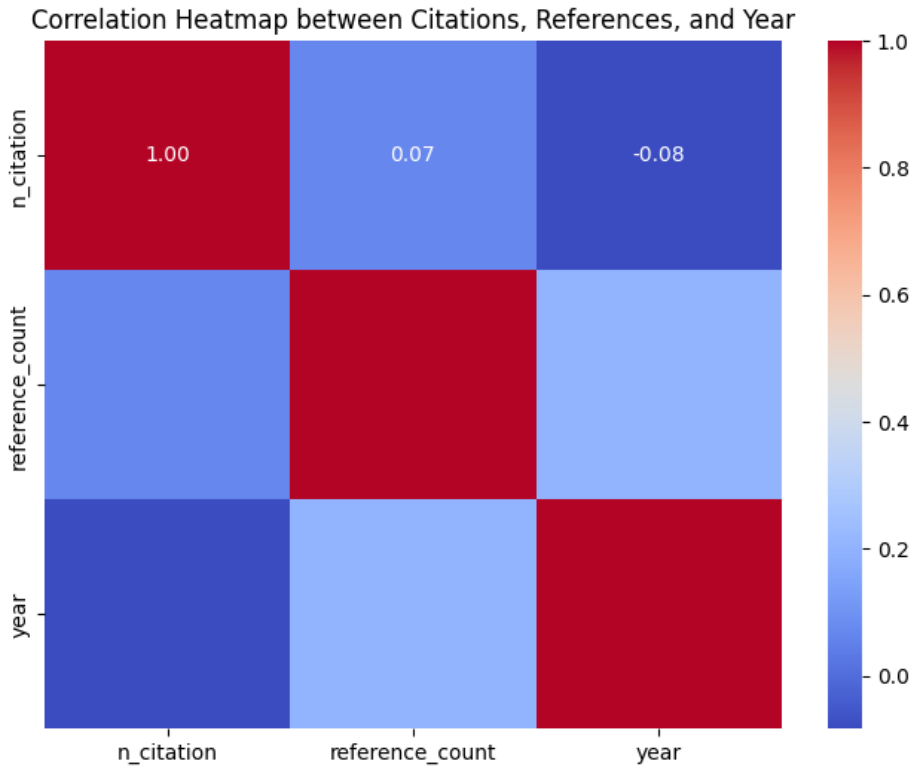

Figure 9: Correlation Heatmap for Key Features (`n_citation`, `reference_count`, and `year`)

Proposed Hypotheses

Based on our analysis, we propose the following hypotheses:

1. Papers that cite more references tend to receive more citations themselves.
2. Papers with more authors receive a higher number of citations compared to papers with fewer authors.
3. Older papers have accumulated more citations over time than newer papers.
4. There is a significant difference in citation counts across different research fields.
5. Papers published in high-impact journals receive more citations than those in lower-impact journals.

Hypothesis Testing-I

Papers that cite more references tend to receive more citations themselves.

Corresponding null hypothesis (H_0) and alternate hypothesis (H_1):

H_0: There is no correlation between the number of references cited by a paper and the number of citations it receives.

H_1: There is a positive correlation between the number of references cited by a paper and the number of citations it receives.

Hypothesis Testing-I

Papers that cite more references tend to receive more citations themselves.

Corresponding null hypothesis (H_0) and alternate hypothesis (H_1):

H_0: There is no correlation between the number of references cited by a paper and the number of citations it receives.

H_1: There is a positive correlation between the number of references cited by a paper and the number of citations it receives.

We use **Pearson's correlation coefficient test (r)**

For a small subset,

$$\implies r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}} = 0.0794 \qquad [\alpha = 0.05]$$

As p-value$= 2.9166 \times 10^{-9}$, we reject H_0 and conclude that there is a significant positive correlation b/w number of references and number of citations.

For the entire dataset,

$$\implies r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}} = 0.0597 \qquad [\alpha = 0.05]$$

As p-value≈ 0, we reject H_0 and conclude that there is a significant positive correlation b/w number of references and number of citations.

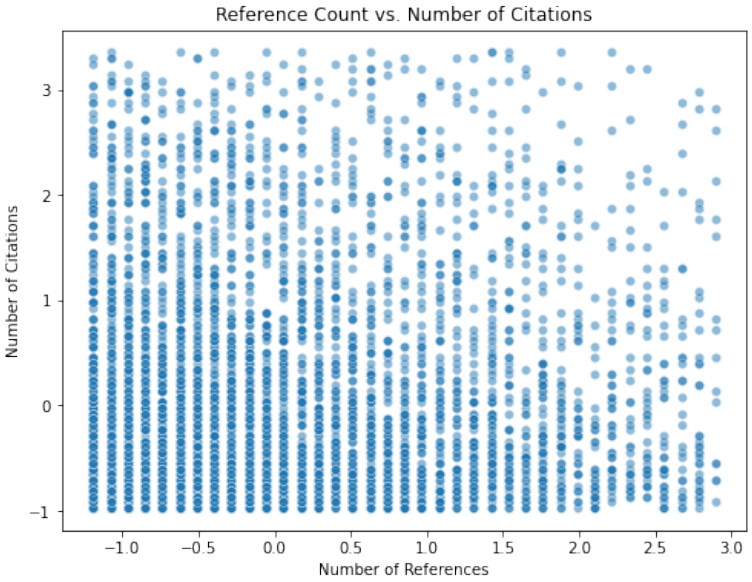

Figure 10: Reference Count vs. Number of Citations (after Normalization) in a small subset

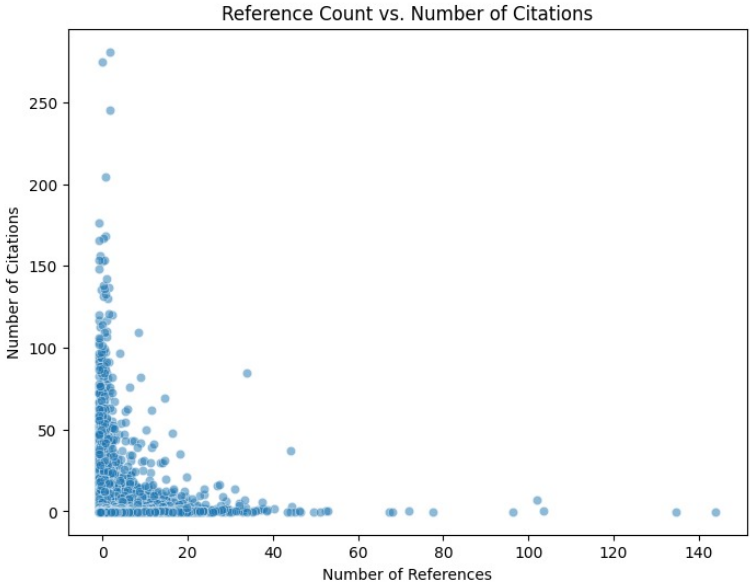

Figure 11: Reference Count vs. Number of Citations in entire dataset

Hypothesis Testing-II

Papers with more authors receive a higher number of citations compared to papers with fewer authors.
Corresponding null hypothesis (H_0) and alternate hypothesis (H_1):

H_0: There is no positive difference in the mean number of citations between papers with many authors and papers with few authors.
H_1: Papers with many authors have a higher mean number of citations than papers with few authors.

Hypothesis Testing-II

Papers with more authors receive a higher number of citations compared to papers with fewer authors.
Corresponding null hypothesis (H_0) and alternate hypothesis (H_1):

H_0: There is no positive difference in the mean number of citations between papers with many authors and papers with few authors.
H_1: Papers with many authors have a higher mean number of citations than papers with few authors.

We use **one-sided Independent samples t-test (t)**

For a small subset,

$$\implies t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = 3.9775 \qquad [\alpha = 0.05]$$

As $t_{\text{crit}} = 7.1329 \times 10^{-5}$, we reject H_0 and conclude that papers with more authors have significantly more citations.

For the entire dataset,

$$\implies t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = -16.4663 \qquad [\alpha = 0.05]$$

As $t_{\text{crit}} = 6.4461 \times 10^{-61}$, we reject H_0 and conclude that papers with more authors have significantly more citations.

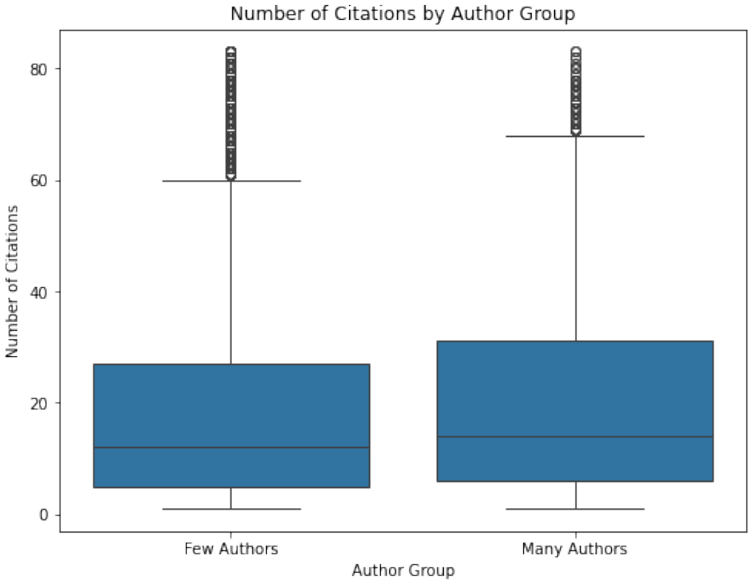

Figure 12: Box plot of number of citations based on number of authors ($\leq$ median or $>$ median) in a small subset

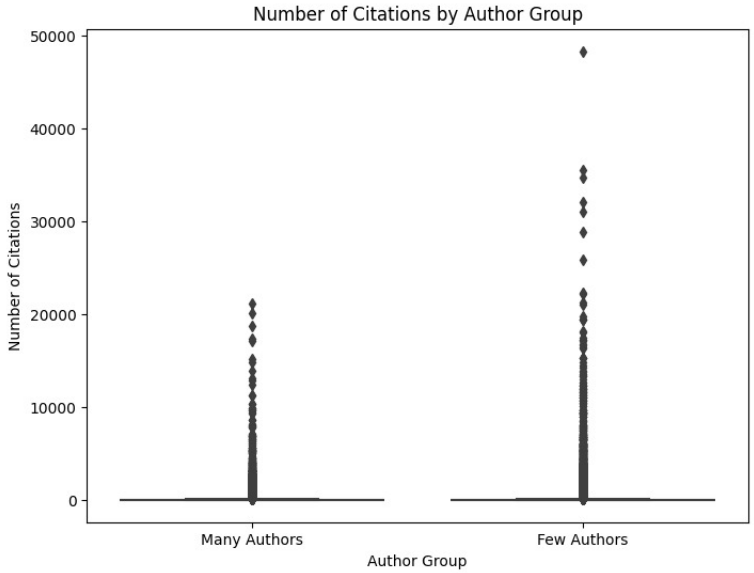

Figure 13: Box plot of number of citations based on number of authors ($\leq$ median or $>$ median) in entire dataset

Hypothesis Testing-III

Older papers have accumulated more citations over time than newer papers.

Corresponding null hypothesis (H_0) and alternate hypothesis (H_1):

H_0: There is no correlation between publication year and the number of citations.

H_1: There is a negative correlation between publication year and the number of citations.

Hypothesis Testing-III

Older papers have accumulated more citations over time than newer papers.

Corresponding null hypothesis (H_0) and alternate hypothesis (H_1):

H_0: There is no correlation between publication year and the number of citations.

H_1: There is a negative correlation between publication year and the number of citations.

We use **Spearman's rank Correlation test (ρ)**

For a small subset,

$$\implies \rho = 1 - \frac{6(\sum \mathrm{rk}(x_i) - \mathrm{rk}(y_i))^2}{n(n^2-1)} = -0.1734 \qquad [\alpha = 0.05]$$

As p-value$= 6.9893 \times 10^{-39}$, we reject H_0 and conclude that older papers have more citations.

For the entire dataset,

$$\implies \rho = 1 - \frac{6(\sum \mathrm{rk}(x_i) - \mathrm{rk}(y_i))^2}{n(n^2-1)} = -0.2882 \qquad [\alpha = 0.05]$$

As p-value≈ 0, we reject H_0 and conclude that older papers have more citations.

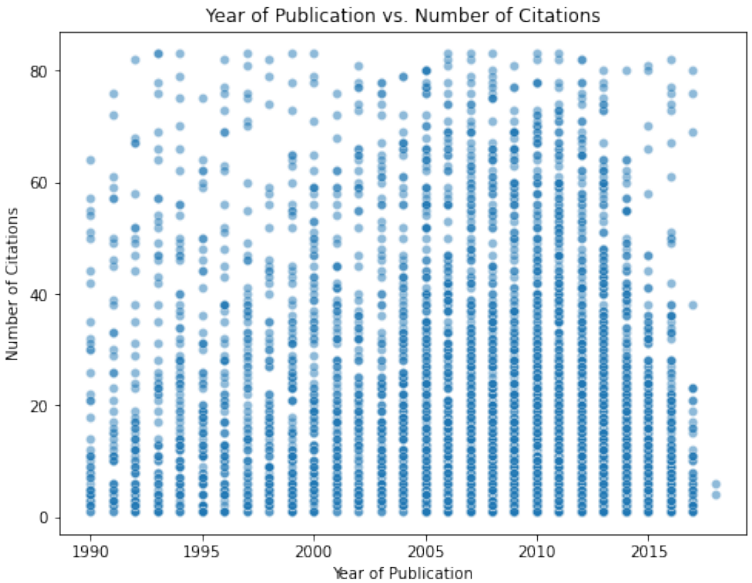

Figure 14: Year of Publication vs. Number of Citations in small subset

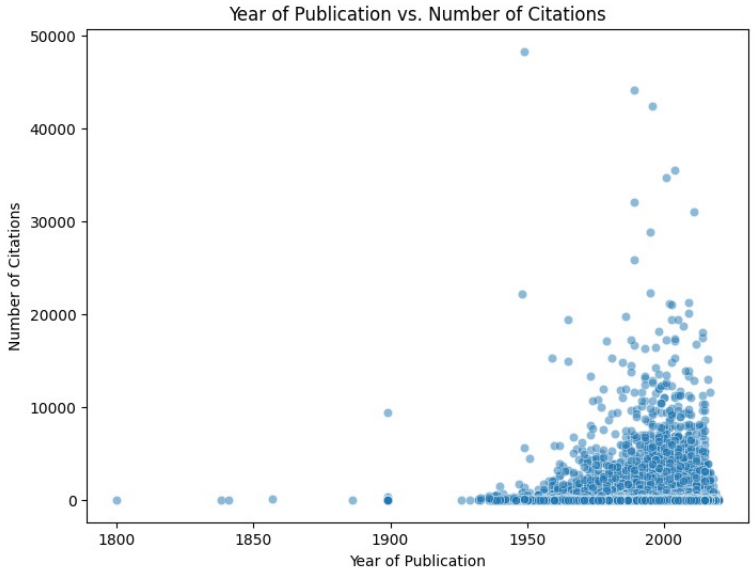

Figure 15: Year of Publication vs. Number of Citations in entire dataset

Hypothesis Testing-IV

There is a significant difference in citation counts across different research fields.

Corresponding null hypothesis (H_0) and alternate hypothesis (H_1):

H_0: Mean citation counts are equal across research fields.

H_1: At least one research field has a different mean citation count.

Hypothesis Testing-IV

There is a significant difference in citation counts across different research fields.

Corresponding null hypothesis (H_0) and alternate hypothesis (H_1):

H_0: Mean citation counts are equal across research fields.

H_1: At least one research field has a different mean citation count.

We use **one-way ANOVA test (F)**

For a small subset,

$$\implies F = \frac{\text{MS}_\text{B}}{\text{MS}_\text{W}} = 3.6356 \qquad [\alpha = 0.05]$$

As $F_{\text{crit}} = 5.9891 \times 10^{-3}$, we reject H_0 and conclude that significant differences exist among research fields.

For the entire dataset,

$$\implies F = \frac{\text{MS}_\text{B}}{\text{MS}_\text{W}} = 121.6890 \qquad [\alpha = 0.05]$$

As $F_{\text{crit}} = 5.3705 \times 10^{-104}$, we reject H_0 and conclude that significant differences exist among research fields.

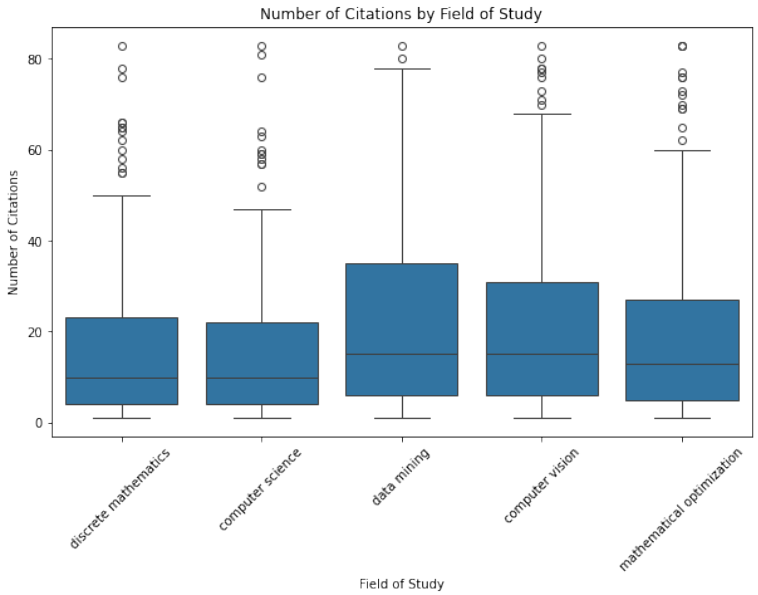

Figure 16: Box plot of number of citations based on field of study in small subset

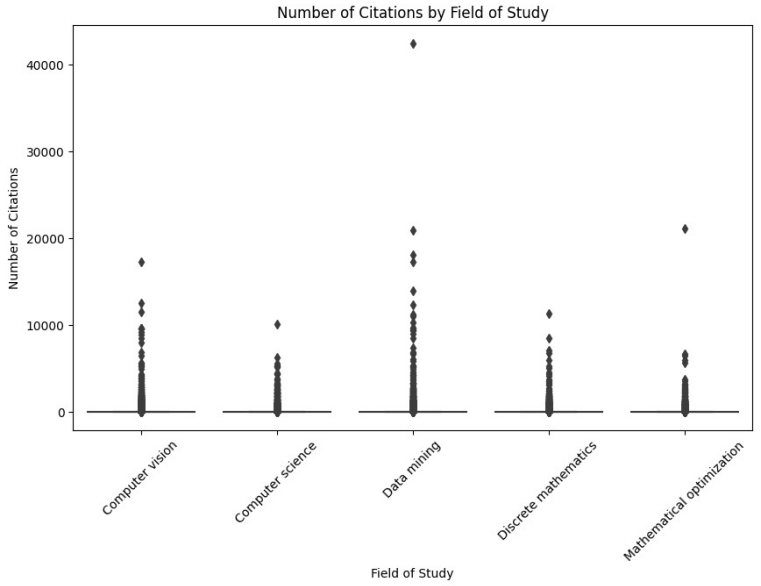

Figure 17: Box plot of number of citations based on field of study in entire dataset

Hypothesis Testing-V

Papers published in high-impact journals receive more citations than those in lower-impact journals.
Corresponding null hypothesis (H_0) and alternate hypothesis (H_1):

H_0: No positive difference in citation counts between high-impact and low-impact journals.
H_1: High-impact journals have higher citation counts.

Hypothesis Testing-V

Papers published in high-impact journals receive more citations than those in lower-impact journals.
Corresponding null hypothesis (H_0) and alternate hypothesis (H_1):

H_0: No positive difference in citation counts between high-impact and low-impact journals.
H_1: High-impact journals have higher citation counts.

We use **one-sided Independent samples t-test (t)**

For a small subset,

$$\implies t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = 32.7283 \qquad [\alpha = 0.05]$$

As $t_{\text{crit}} = 2.0389 \times 10^{-211}$, we reject H_0 and conclude that high-impact journals have significantly higher citation counts.

For the entire dataset,

$$\implies t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = 216.5702 \qquad [\alpha = 0.05]$$

As $t_{\text{crit}} \approx 0$, we reject H_0 and conclude that high-impact journals have significantly higher citation counts.

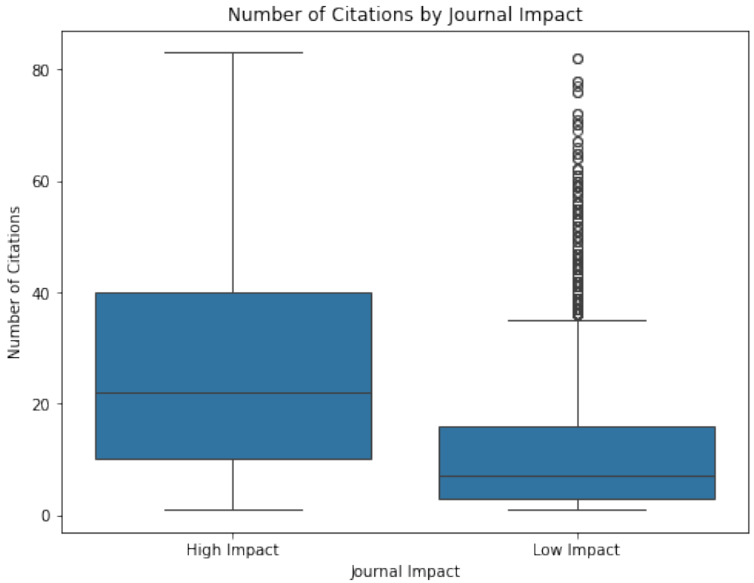

Figure 18: Box plot of number of citations based on average number of citations in journal ($\leq$ median or $>$ median) in small subset

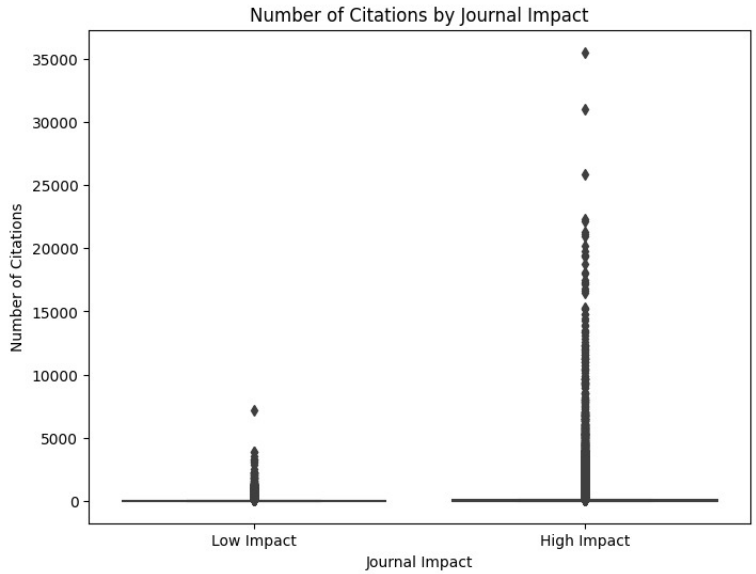

Figure 19: Box plot of number of citations based on average number of citations in journal ($\leq$ median or $>$ median) in entire dataset

Future Plans

We plan on:

1. Further analyzing properties of the dataset but with respect to its graph structure
2. Implementing algorithms to find communities of similar papers that span the entire dataset
3. Implementing a machine learning model to find a common community, giving a set of papers (If time permits)

Thanking you all for your time and attention

References I

[1] M. Aché, "Citation network dataset." [Online]. Available: https://www.kaggle.com/datasets/mathurinache/citation-network-dataset/data