Menu

Go to

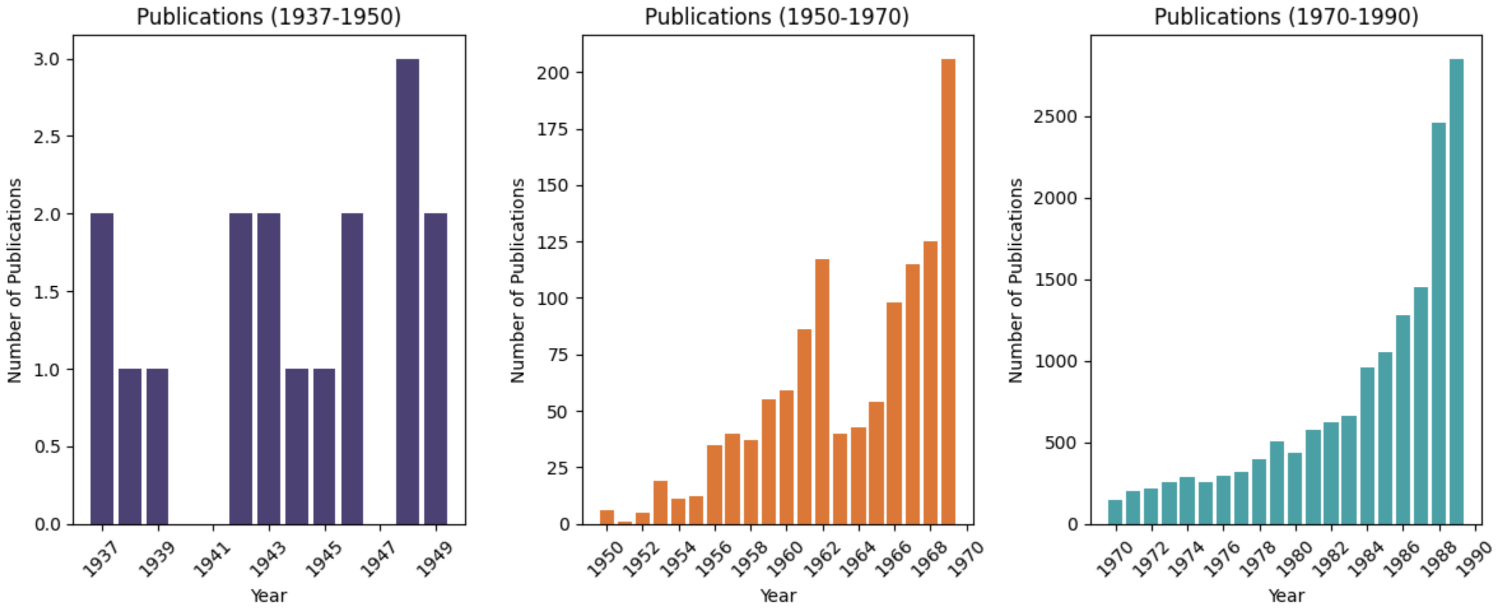
PART1

PART2

PART3

EDA

Create 3 bar charts. For each bar chart draw the number of publications in a given year range. Do this for the ranges 1937-1950, 1950-1970, 1970-1990. Compare the three



output_1.1

Publications (1937-1950)

The number of publications is relatively low.

There is some variation, with the highest peak reaching around 3 publications.

Some years have no publications.

Publications (1950-1970)

There is a clear upward trend.

The number of publications increases significantly, especially after 1960.

By 1970, the number of publications exceeds 200.

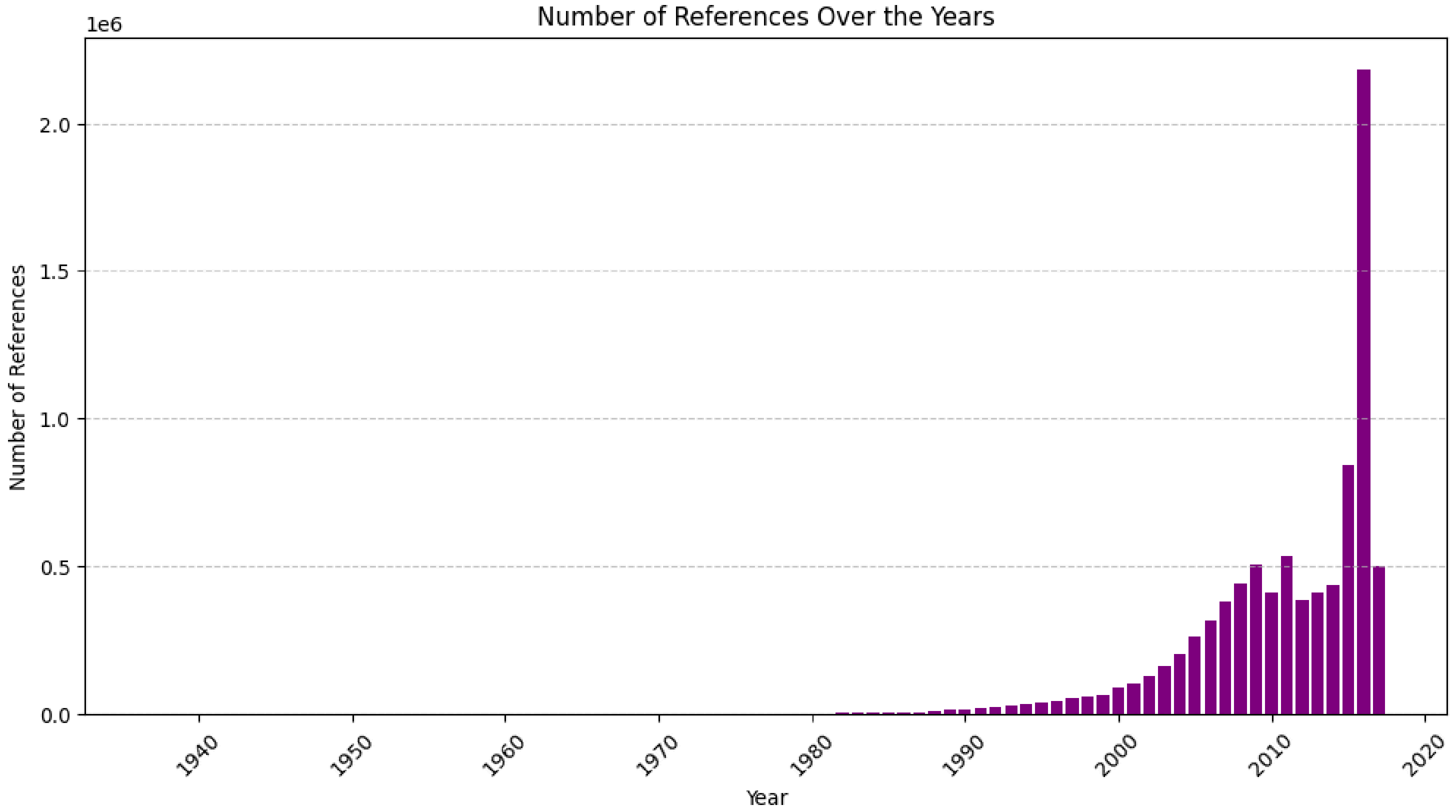
Publications (1970-1990)

A dramatic increase in the number of publications.

The number of publications continues to rise exponentially.

By 1990, the count reaches around 2700 publications.

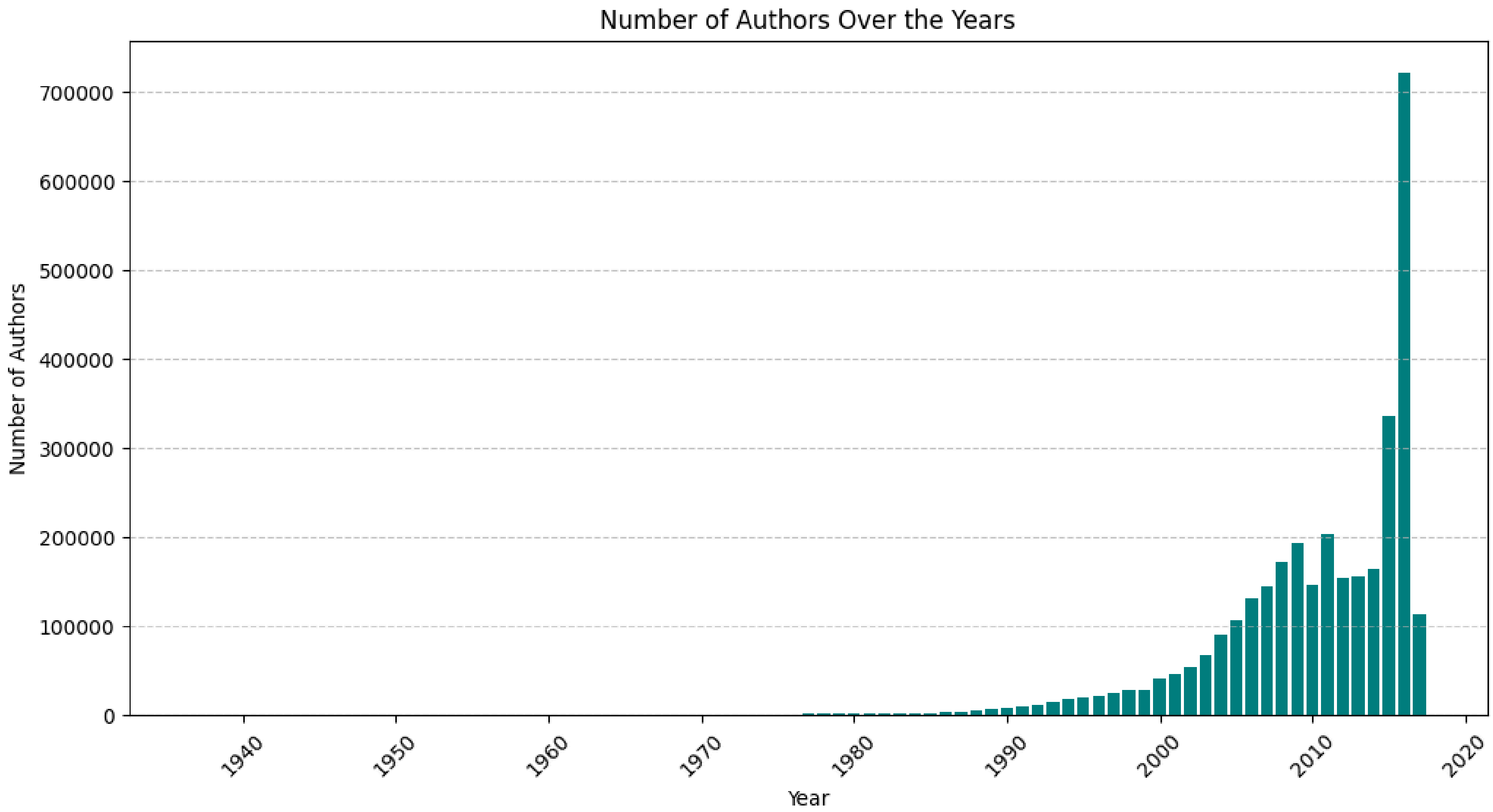
Create a bar chart of the number of references over the years.



output_1.1.2

The number of references significantly increases over time, particularly after the 1990s. Before 1980, the number of references per year is relatively low. A steady growth phase starts around the 1990s, increasing each year.

Create a bar chart of the number of authors over the years.



output_1.1.2

The number of authors has shown a steady increase over the decades, particularly after 1990.Before 1970, the number of authors was relatively low, which aligns with the historical trend of individual researchers publishing more than collaborative teams.Around the 1990s, there is a noticeable rise in the number of authors per year.The number of authors peaks dramatically in the 2010s and 2020.

Find the Pearson correlation coefficient and Spearman Rank correlation coeffbetween the number of authors and number of references.

Result is:

Pearson Coefficient is 0.056027878375202955 and Spearman Rank Correlation is 0.0872116655830498.

The Pearson correlation (0.056) suggests almost no linear relationship between the number of authors and references.

The Spearman correlation (0.087) indicates a slightly stronger but still weak monotonic relationship.

Find the Pearson correlation coefficient and Spearman Rank correlation coeffbetween the number of authors and number of references.

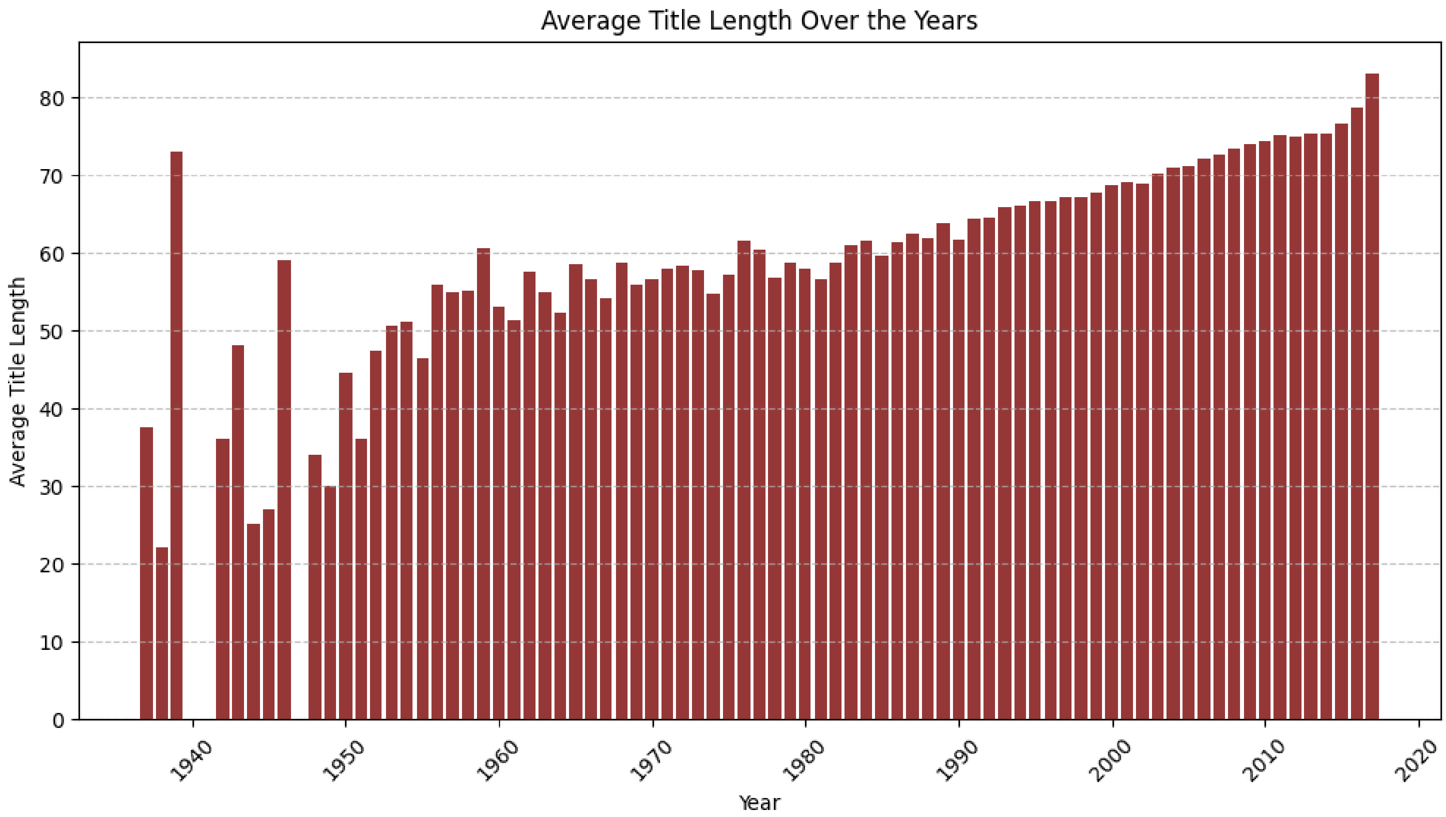
Result is:

Pearson Coefficient is -0.0027965270825141553 and Spearman Rank Correlation is -0.01661606049149943

The Pearson correlation is almost zero correlation between the number of authors and the number of citations.Suggests that there is no significant linear relationship between the two variables.

The Spearman correlation is slightly negative but still very weak correlation.Indicates that there is almost no monotonic (rank-based) relationship between the number of authors and citations.

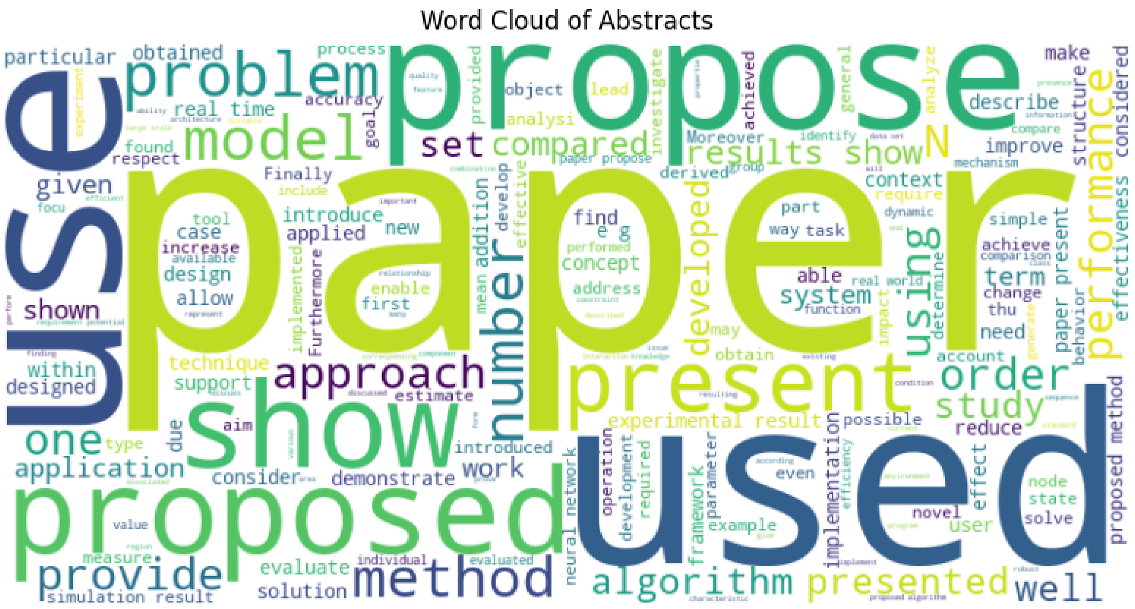
Draw a bar chart of the title length over the years.



output_1.1.2

From the 1940s to 2020, there is a clear upward trend in average title length.This suggests that research paper titles have gradually become longer over the decades.In earlier years (before 1950), title lengths were shorter and more varied, but after 1970, there is a steady increase.The 1930s to 1950s show large fluctuations in title length, with some years having very short average titles and others having extremely long ones.

Draw a bar chart of the title length over the years.



output_1.1.7

The largest and most prominent words include:"paper," "propose," "use," "present," "model," "problem," "proposed," "results."These words are commonly associated with academic research abstracts, reflecting the typical structure of research papers.

Pearson Correlation:

- Pearson correlation measures linear relationships.
- A value of 0.2718 indicates a weak positive correlation between title length and the length of referenced paper titles.
- This suggests that papers with longer titles tend to reference other papers with longer titles, but the relationship is not strong.
- The correlation is not close to 0, so there is some level of connection, but many other factors likely influence title length.

Spearman Correlation:

- Spearman correlation measures monotonic relationships, meaning it detects increasing/decreasing trends even if they are not perfectly linear.

- The Spearman coefficient is slightly higher (0.2737) than Pearson, suggesting a weak but slightly stronger monotonic relationship.
- This means that as one paper's title length increases, its referenced papers also tend to have longer titles, though not in a perfectly linear way.

Top Authors Based On Publications

	Author	Publications
0	Wei Wang	950
1	Wei Zhang	657
2	Yang Liu	629
3	Lei Zhang	579
4	Wei Li	559
5	Jun Wang	544
6	Lei Wang	519
7	Lajos Hanzo	458
8	Wei Liu	456
9	Jun Zhang	455

Top 10 Authors by Citations

	Author	Citations
0	David G. Lowe	65344
1	Hari Balakrishnan	55096
2	Scott Shenker	54164
3	Ian F. Akyildiz	53654
4	Michael I. Jordan	53448
5	Ion Stoica	52890
6	Chih-Jen Lin	52302
7	Takeo Kanade	50743

	Author	Citations
8	Deborah Estrin	49925
9	Vladimir Vapnik	49755

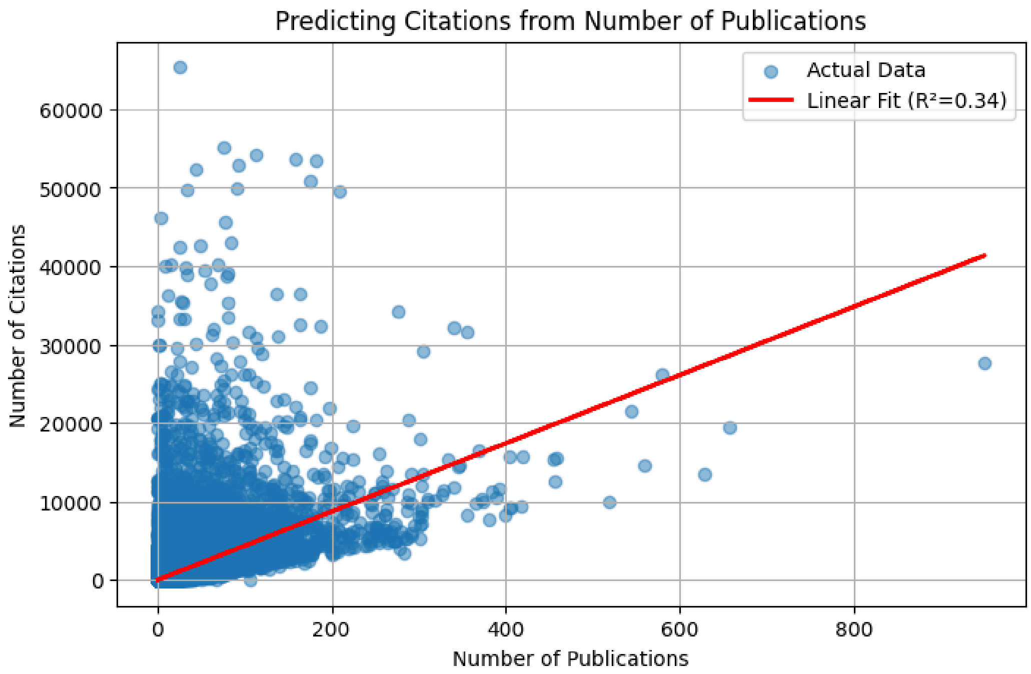
Top 10 Papers by References

	Paper ID	Title	References
0	371369	Comprehensive frequency-dependent substrate no...	759
1	780292	Time in Qualitative Simulation.	561
2	104143	Bibliography on cyclostationarity	412
3	214646	Fifty Years of MIMO Detection: The Road to Lar...	396
4	484969	An Exploration of Enterprise Architecture Rese...	394
5	223901	Structure and dynamics of molecular networks: ...	386
6	302124	The NP-completeness column: An ongoing guide	363
7	707510	Digital geometry	361
8	325083	Deep Learning: Methods and Applications	343
9	538381	Review: learning bayesian networks: Approaches...	326

Top 10 Papers by Citations

	Paper ID	Title	Citations
0	332760	Distinctive Image Features from Scale-Invarian...	42508
1	294527	Bowling alone: the collapse and revival of Ame...	34288
2	358174	LIBSVM: A library for support vector machines	33016
3	716671	Random Forests	28679
4	18485	Support-Vector Networks	26114
5	45248	MapReduce: simplified data processing on large...	24381
6	81801	A fast and elitist multiobjective genetic algo...	24245

	Paper ID	Title	Citations
7	150727	A theory for multiresolution signal decomposit...	24182
8	458466	ImageNet Classification with Deep Convolutiona...	22884
9	442067	Histograms of oriented gradients for human det...	22795

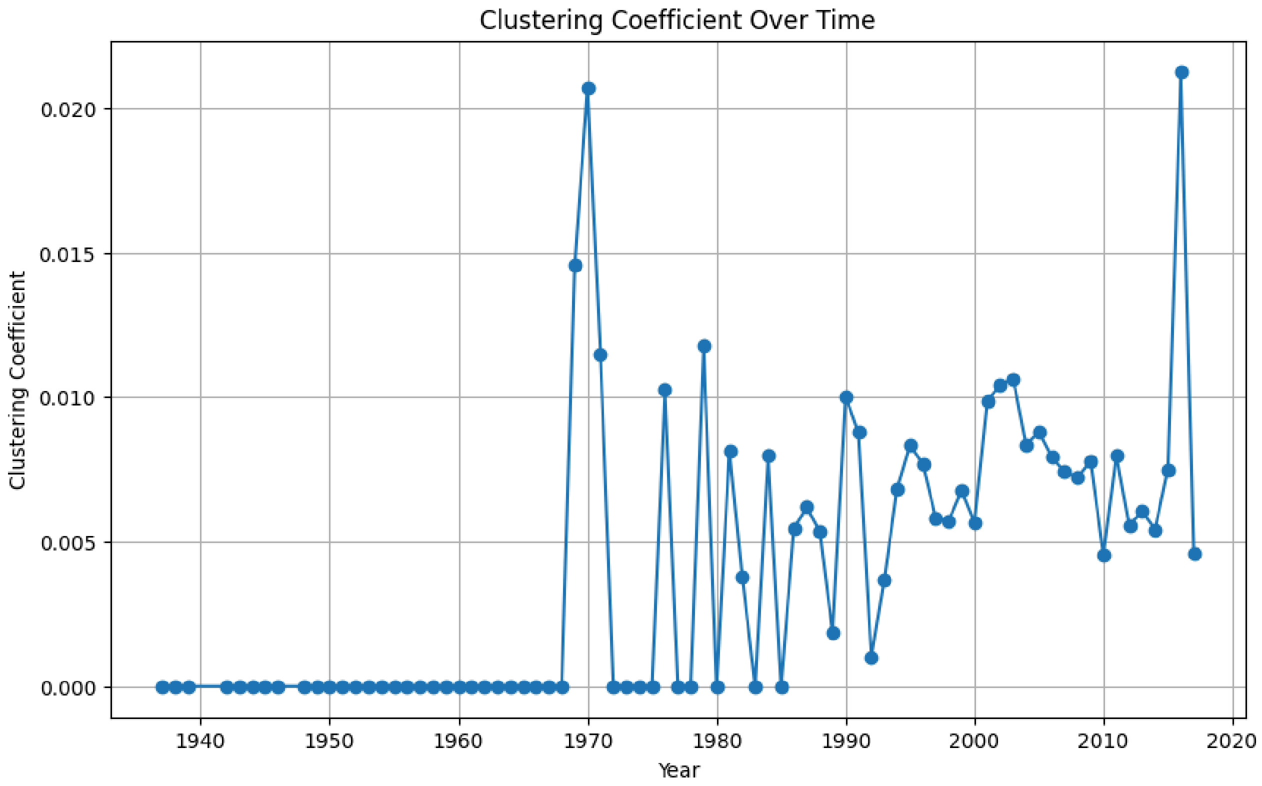


output_1.1.13

The blue dots represent actual data points, where each dot corresponds to an author’s number of publications (X-axis) and total number of citations (Y-axis).The red line represents the linear regression model's fit, attempting to predict the number of citations based on the number of publications.The spread of points indicates significant variation in citation counts, even for authors with a similar number of publications.

R^2 (Coefficient of Determination) = 0.34 means that 34 percent of the variance in citations can be explained by the number of publications.This suggests a moderate but not strong relationship
is some correlation between the number of publications and citations.However, 66 percent of citation variation is influenced by other factors not accounted for in this model.

Citation Network (Paper-Paper Network)



Citation Network

Clustering Coefficient Over Time

Observations from the Graph:

Before 1970: The clustering coefficient is near zero, indicating a sparse citation network with little interconnection.

1970s Spike: Around the early 1970s, a sharp increase in clustering coefficient is observed, peaking above 0.02.

This suggests that research papers started citing more interconnected works, leading to a more densely connected network.

Post-1980 to 2000:

- The clustering coefficient fluctuates but remains nonzero, indicating a more interconnected research landscape.
- Multiple peaks suggest the emergence of strongly cited sub-networks (e.g., influential research areas).

Post-2000 Increase:

- A gradual rise in clustering coefficient suggests that research has become increasingly interconnected.
- More papers cite well-established sources, forming tightly connected communities.

Recent Peak (~2020):

A significant jump suggests a shift towards highly interconnected citations.

Network Metrics

Average Path Length = 11.29

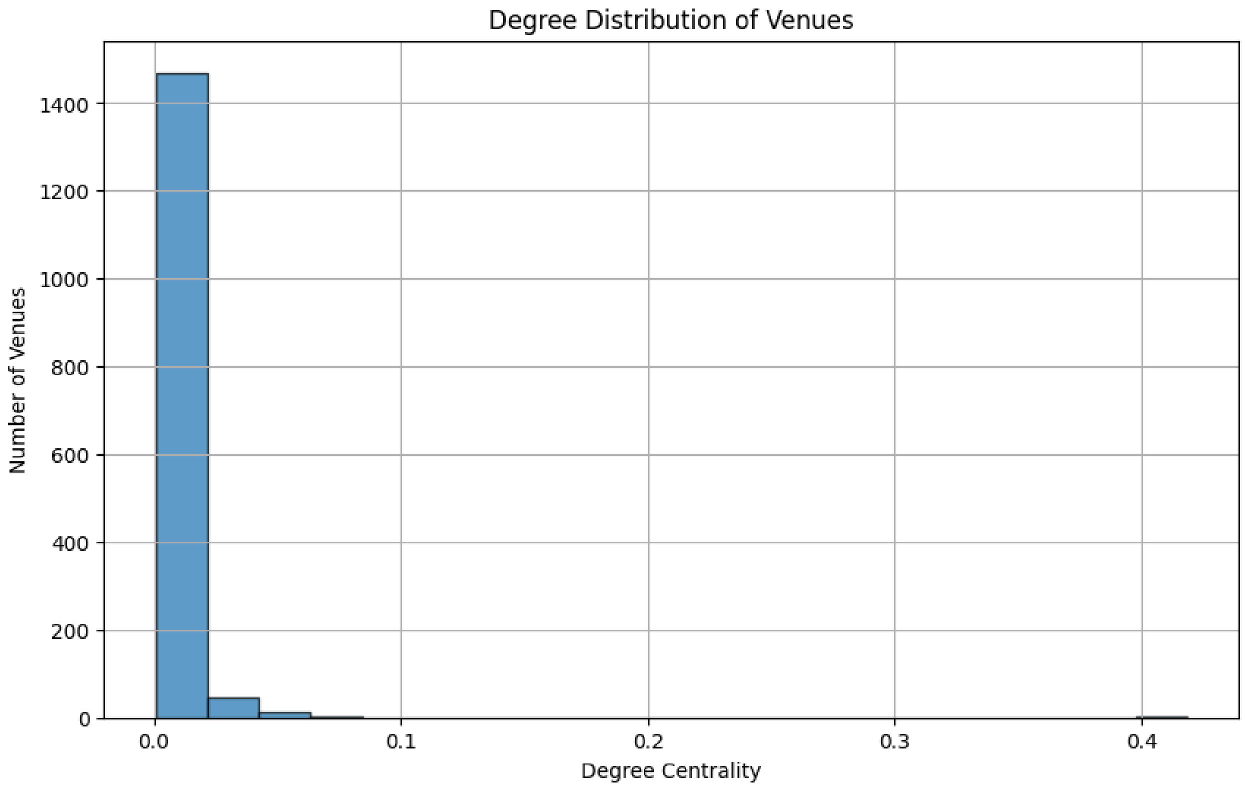
- This means that, on average, any two papers in the citation network are 11.29 citations apart.
- Suggests a moderate level of connectivity, but still requires many citation 'hops' to navigate the full network.

Diameter = 29

- The longest shortest path in the network is 29 steps.
- This indicates that while some research fields are tightly interconnected, others remain far apart, requiring multiple citations to link related works.

Co-authorship Network (Author-Author Network)

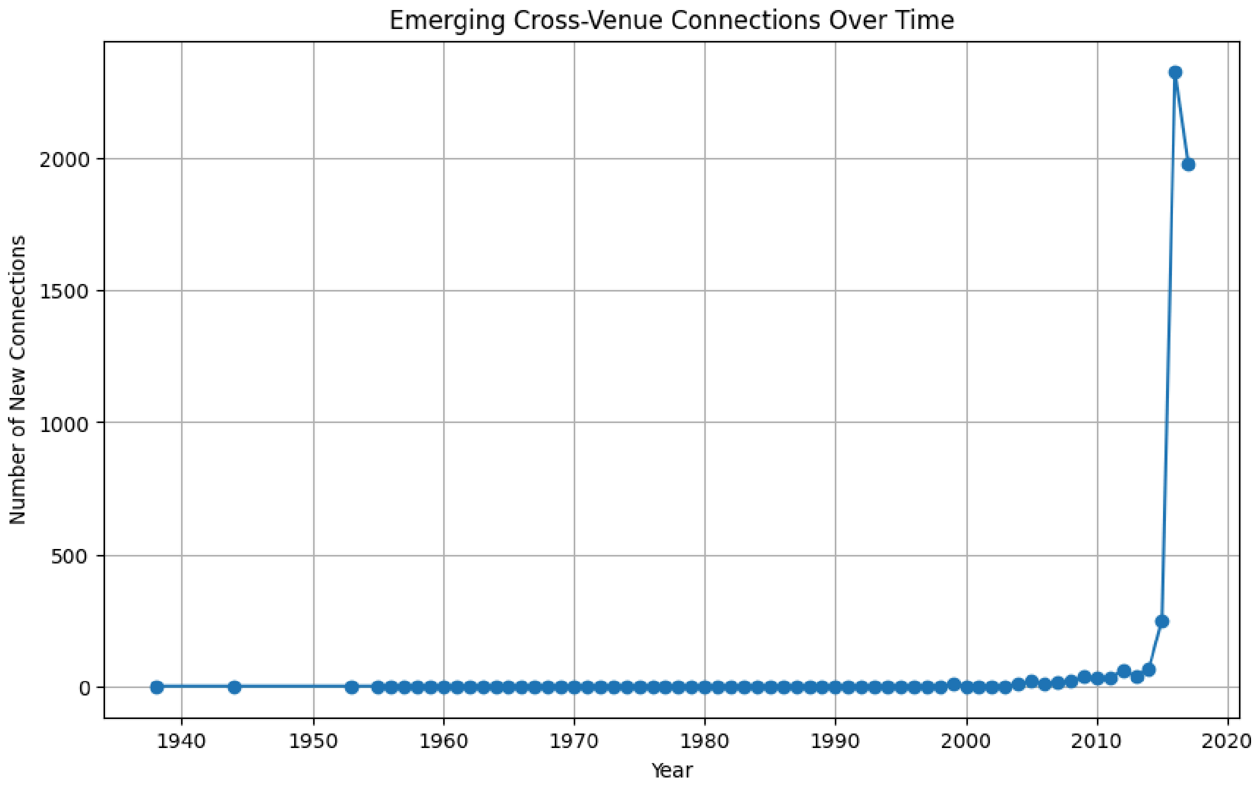
Venue Network (Conference-Journal Network)



The histogram of degree distribution shows that most venues have a very low degree centrality.The majority of venues do not frequently cite other venues, meaning they are more isolated within their own research communities.A small number of venues have a higher degree, meaning they play a significant role in connecting different venues.

Measures how often a venue lies on the shortest path between other venues. The "human factors in computing systems" venue ranks highly, suggesting it acts as a bridge between different research domains.

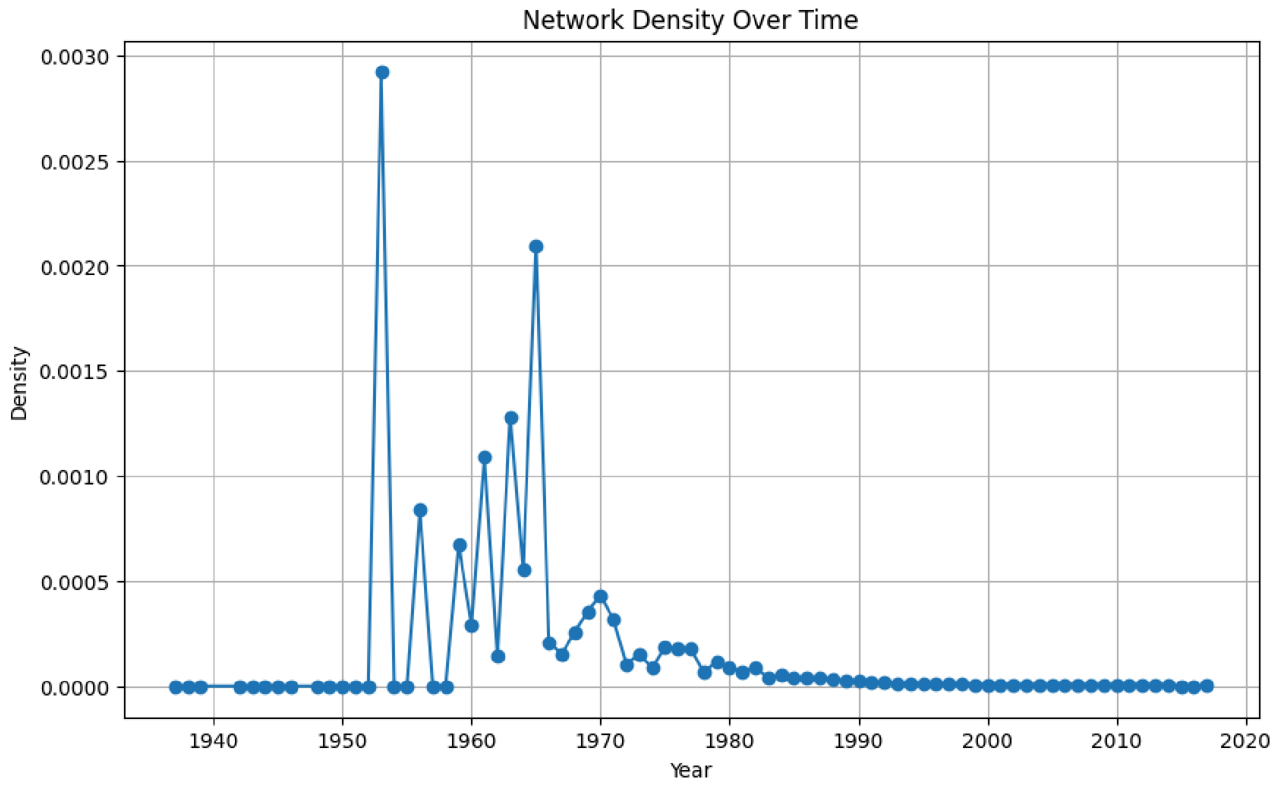
- **IEEE Transactions on Pattern Analysis and Machine Intelligence:** This venue is frequently cited, indicating its influence in the academic community.
- **Human Factors in Computing Systems (High Betweenness):** It likely connects multiple disciplines, meaning that interdisciplinary research frequently passes through this journal.
- **Neural Information Processing Systems (High Closeness):** It has a strong reachability across the network, meaning that papers from various disciplines quickly find their way to this venue.



The number of new venue connections per year has dramatically increased after 2010, with a steep rise around 2015-2020.There was relatively little interdisciplinary citation activity before the 2000s.

Temporal Evolution of the Citation Network

This section has done by a sample with fraction 0.05



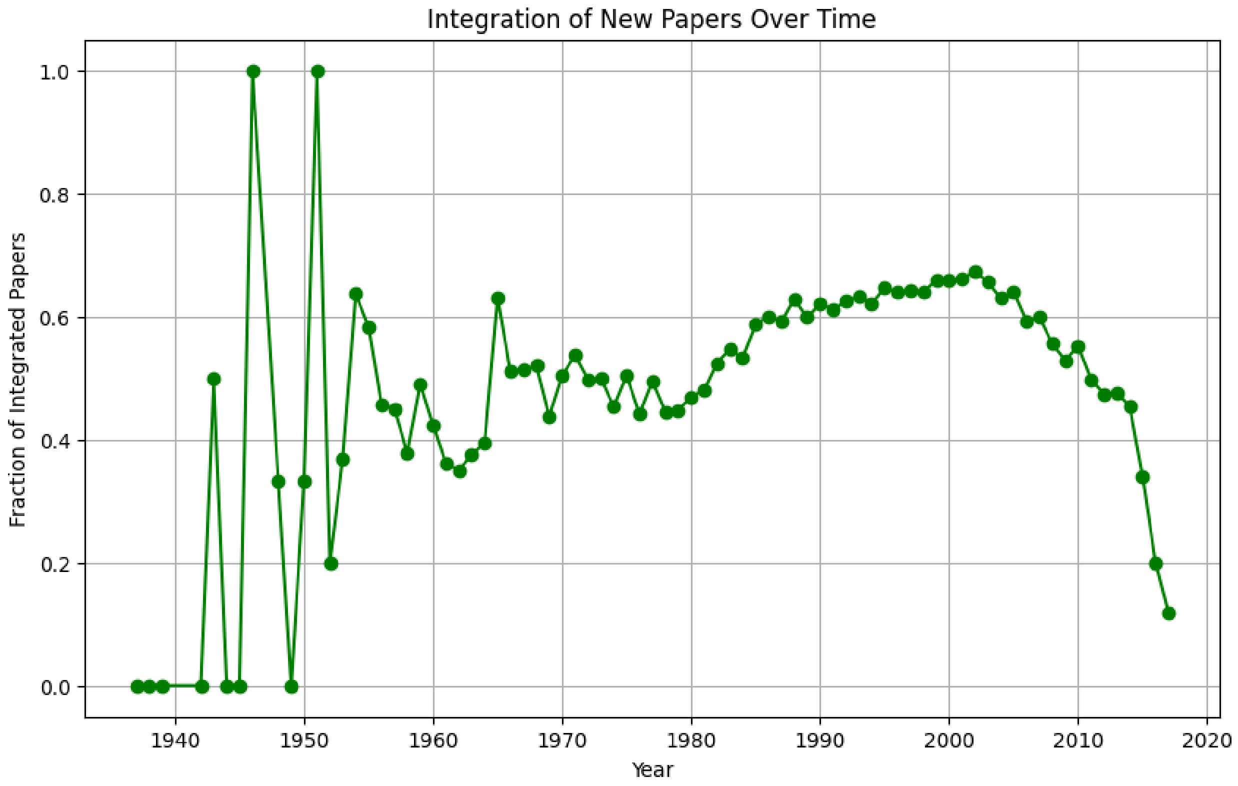
The density of the citation network fluctuates significantly across different years.There are sharp spikes in density, particularly around the 1950s and 1960s, followed by a decline in the later years.After 1980, the network density becomes much lower and relatively stable.

The spikes in density suggest bursts of citations within short timeframes. These could correspond to landmark papers or specific technological advancements that led to increased interconnections in the citation network.The decline after 1980 might indicate the expansion of academic literature, where papers are spread across a larger number of research domains, reducing the overall density.The stabilization in later years suggests a more distributed citation network, where papers are being cited across a broader field instead of forming dense clusters.

The papers with the highest in-degree values represent major breakthroughs that have significantly impacted research communities.Many of these top-cited papers correspond to computer vision and machine learning, reinforcing the dominance of AI-related research in recent citation patterns.

- The most highly cited papers include well-known research contributions, such as:
- *Distinctive Image Features from Scale-Invariant Keypoints* (5841 citations, 2004)
- *LIBSVM: A library for support vector machines* (5057 citations, 2011)
- *Histograms of Oriented Gradients for Human Detection* (3279 citations, 2005)

- *ImageNet Classification with Deep Convolutional Networks* (3242 citations, 2012)
- *Random Forests* (3235 citations, 2001)



The fraction of new papers being integrated into the citation network fluctuates over time.In the early years (1940s–1960s), integration was unstable, showing rapid increases and decreases.From 1980 onwards, the integration fraction stabilized and increased, reaching its peak between 1990 and 2010.A notable decline in integration is visible in the most recent years.