University of Pavia

Analysis Of Changes In Scientific Research Through Graph Theory

Financial Data Science Project

Federico P. Bernardini
Computer Engineering - Intelligent Control System
July 2024

Abstract

This project examines the evolution of co-authorship collaborations in academic research in the scientific field by using network analysis and temporal data, focusing on two main areas: (a) Analysis of the collaborations number and volume of papers produced over the years. (b) Identification of trending topics and most popular journals over the years. The final objective is to obtain an overview of the the research evolution in the period that goes from 1980 to 2009 from a historical-technological and socio-professional perspectives.

Contents

1 Introduction

In academic research, collaboration among authors plays a crucial role in advancing knowledge and innovation: without research there is no progress. This project investigates the evolution of co-authorship networks in the scientific area from 1980 to 2009, providing a comprehensive analysis of how these collaborations have developed and transformed over time. These aspects are studied through a detailed analysis of co-authorship patterns and trends, the two main areas on which the analysis focuses are:

(a) The dynamics of the collaborations structure and volume of papers produced,

(b) The most popular journals and areas of research.

After the graphs presentation, a discussion of the overall results will be done, trying to explain also from an historical-technological point of view the seen phenomenons. Finally, some other comments on the social-work prospective will be given, together with the analysis weaknesses and some possible developments for future work.

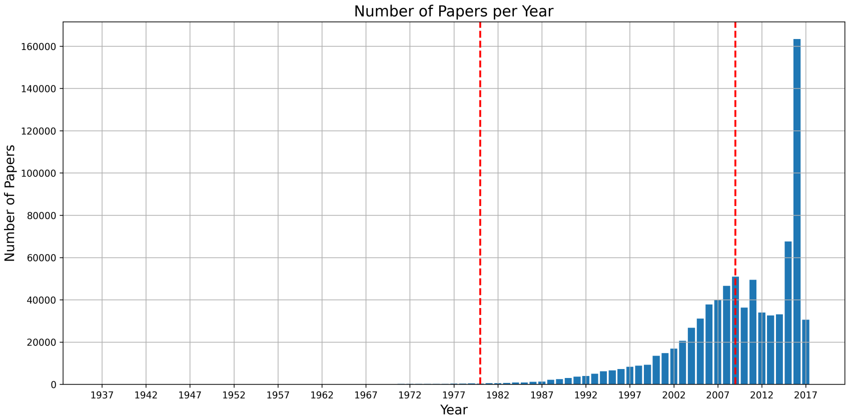


Fig. 1: Numbers of papers published each year from 1937 to 2017. (- -) indicate the extremes of the studied time interval, 1980 and 2009.

1.1 Hardware

Experiments are carried out using Matlab R2023b and Python 3.12.3 on a laptop with Apple M1 Pro, 16GB RAM and MacOS Sonoma.

2 Dataset

The dataset used to conduct the studies is taken from Kaggle, a detailed explanation on the data extraction technique is provided in the following paper [5].

In total the dataset has 10^6 different observations spanning from year 1937 to 2017 and each element is composed by the following objects:

- *id*: a unique identifier for each paper,
- *title*: the title of the research paper,
- *authors*: the list of authors involved in the paper,
- *venue*: The journal or venue where the paper was published,
- *year*: the year when the paper was published,
- *n°citation*: the number of citations received by the paper,
- *references*: a list of paper IDs that are cited by the current paper,
- *abstract*: the abstract of the paper.

Given to the limited power of the available hardware, the dataset cannot be entirely used, some data sampling processes are required.

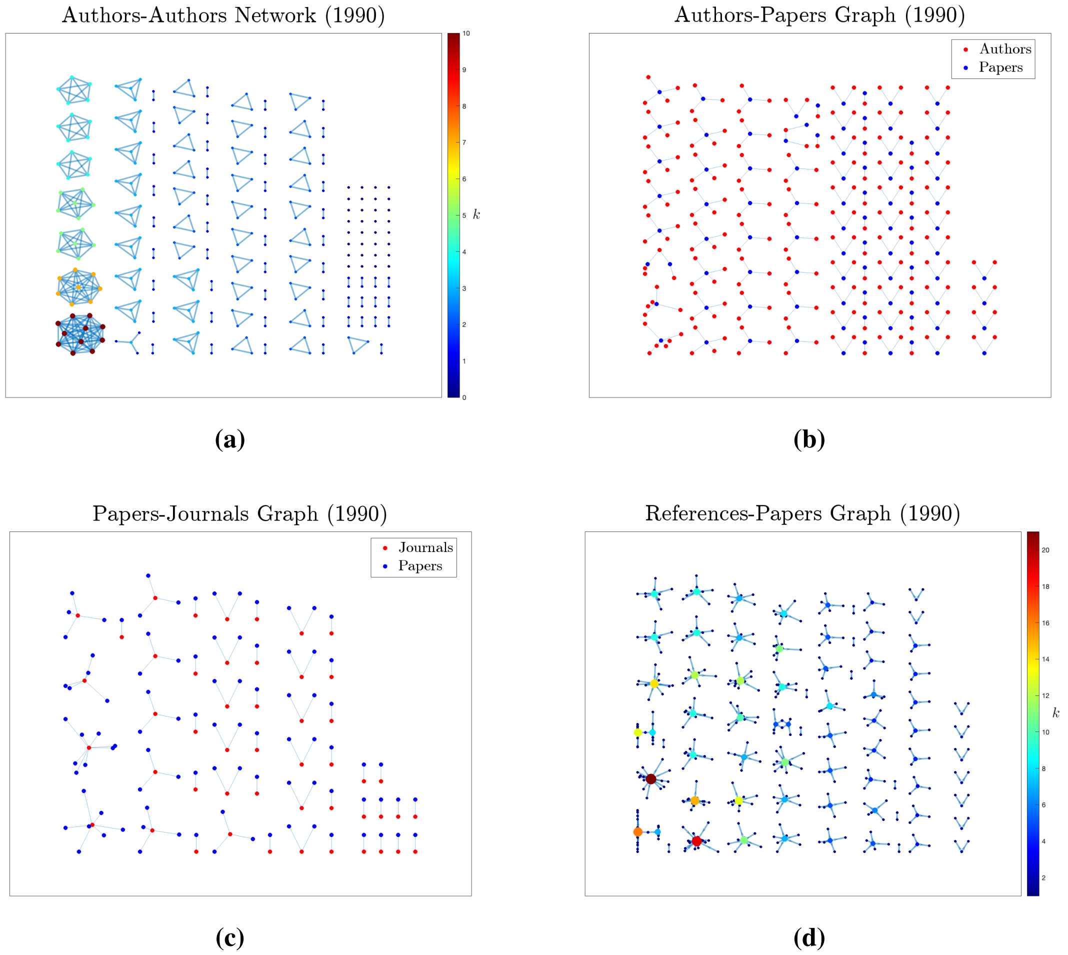
2.1 Data Handling

Given the large amount of available data, and the inability of using all of them, the first sampling procedure done was discarding all the samples containing `NaN` (Not a Number) elements, in order to avoid ambiguities.

Then a restriction of the considered timeframe was performed. The lower limit has been set in 1980 because, as illustrated in Fig.1, the majority of papers were published after this year. The upper limit instead is 2009 because, as shown in article [2] from Michael Fire and Carlos Guestrin, the number of publications has an exponential growth over the year, while in the used dataset there is an evident decreasing number of papers beyond 2009, except for 2016.

Finally, a random selection of just 25000 papers was performed from the remaining dataset. This was necessary because some of the generated graphs were too large to analyze within a reasonable computational time.

NOTE: Raw data analysis and data cleaning procedures have been performed using the Python library Pandas.


Authors-Authors Network (1990)
k
0
1
2
3
4
5
6
7
8
9
10
Authors-Papers Graph (1990)
● Authors
● Papers
Papers-Journals Graph (1990)
● Journals
● Papers
References-Papers Graph (1990)
k
2
4
6
8
10
12
14
16
18
20

(a) **(b)** **(c)** **(d)**

Fig. 2: Example of utilized network during the analysis; k is the node degree, see section 4.1

3 Graphs

A graph $\mathcal{G}$ is a structure that can be mathematically described as $\mathcal{G} = \{\mathcal{N}, \mathcal{E}\}$, where $\mathcal{N} = \{n_1, n_2, \ldots\}$ is the set of nodes and $\mathcal{E} \subseteq \{l_{mn} = (n,m) | (n,m) \in \mathcal{N}\}$ is the set of possible links between nodes. It is useful to define $N = |\mathcal{N}|$ as the number of nodes in graph $\mathcal{G}$.

NOTE: In dealing with oriented graphs it is important to keep in mind that $l_{mn} \nleftrightarrow l_{nm}$.

A special type of graph utilized in this work are the bipartite ones. Their peculiarity is that nodes belong to two different sets, $\mathcal{U} = \{u_1, u_2, \ldots\}$ and $\mathcal{V} = \{v_1, v_2, \ldots\}$. In these networks links l_{uv} are possible just between elements of different sets.

3.1 Graph Presentation

In this subsection the graphs created for this study case are introduced.

Authors-Authors Graph: Is a non-oriented unweighted graph, where nodes n represent the authors and links l_{mn} represent a collaboration between researcher m and n. In Fig.2a it is possible to see an example of these networks structure for year 1990.

References-Papers Graph: Is an oriented unweighted graph, where nodes n represent different papers and links l_{mn} represent the citation made by node m of node n. In Fig.2d it is possible to see an example of these networks structure for year 1990.

Papers-Journals Graph: Is a bipartite unoriented graph, where nodes $u \in \mathcal{U}$ are papers and nodes $v \in \mathcal{V}$ are journals. A link l_{uv} represent that paper u is published on journal v. In Fig.2c it is possible to see an example of these networks structure for year 1990.

Authors-Papers Graph: Is a bipartite unoriented graph, where nodes $u \in \mathcal{U}$ are authors and nodes $v \in \mathcal{V}$ are papers. A link l_{uv} represent that author u has contributed to the publication of paper v. In

Fig.2b it is possible to see an example of these networks structure for year 1990.

4 Graph Analysis

In this section as a first step, the metrics used for the study of the four networks will be introduced, and then some technical conclusions will be drawn.

All the metrics presented in this section are implemented on Matlab using the BC Toolbox, deeper insights of the used functions are given in in [4].

4.1 Metrics

Node Degree: The degree k_i of the i-th node represent the number of connection that this node has with the other nodes in the graph.
In the context of oriented networks, it is more appropriate to distinguish in-degree, k^{in}, and out-degree, k^{out}. These metrics represent the number of incoming and outgoing connections of a node, respectively. In this work, when we refer to the degree in such networks without specifying k^{in} or k^{out}, then we mean the sum of the two: $k = k^{in} + k^{out}$.

Average Node Degree: It is the mean computed over the whole nodes degree of a graph,

$$\langle k \rangle = \frac{1}{N} \sum_{i \in \mathcal{N}} k_i. \tag{1}$$

In case of oriented graph $k_i = k_i^{in} + k_i^{out}$.

Node Degree Distribution: It represents the probability $P(k)$ for a node of having a degree k. In this project two type of distributions will be used, the Poissonian

$$P(k_i = k) = \frac{\mu^k}{k!} e^{-\mu}, \tag{2}$$

where $\mu = \langle k \rangle$ is the average node degree of the graph, and the Power-law

$$P(k_i = k) = ck^{-\alpha}, \tag{3}$$

where c and α are appropriate tuning parameters.

Average Path Length: The shortest path length between two nodes m and n is denoted with d_{mn}. In this work, it refers to the minimum number of nodes that must be traversed to connect m to n. The average path length $\langle d \rangle$ represent the average of all the shortest path lengths in the network. It is computed as

$$\langle d \rangle = \frac{1}{|\mathcal{N}_{sp}|} \sum_{i \in \mathcal{N}_{sp}} d_i, \tag{4}$$

where $\mathcal{N}_{sp} = \{n_1 m_1, n_1 m_2, ...\}$ is the set of all the possible connected pairs of nodes in the graph.
In case of a directed graph $d_{mn} \neq d_{nm}$.

Assortativity: In this work, when we refer to the assortativity we are talking about the property of a network of having links between similar nodes in terms of node degree. A network is said to be assortative if the connections happen between nodes with a similar node degree, while we speak of a disassortative network if connections occur between nodes with different node degrees.

To quantify the entity of this phenomenon the assortativity coefficient described by Newman in [3] is used, that is

$$r = \frac{L^{-1} \sum_{(m,n) \in \mathcal{N}} k_m k_n - \left[L^{-1} \sum_{(m,n) \in \mathcal{N}} \frac{1}{2} (k_m + k_n) \right]^2}{L^{-1} \sum_{(m,n) \in \mathcal{N}} \frac{1}{2} (k_m^2 + k_n^2) - \left[L^{-1} \sum_{(m,n) \in \mathcal{N}} \frac{1}{2} (k_m + k_n) \right]^2} \tag{5}$$

where $L = |\mathcal{E}|$ is the number of links in the network.

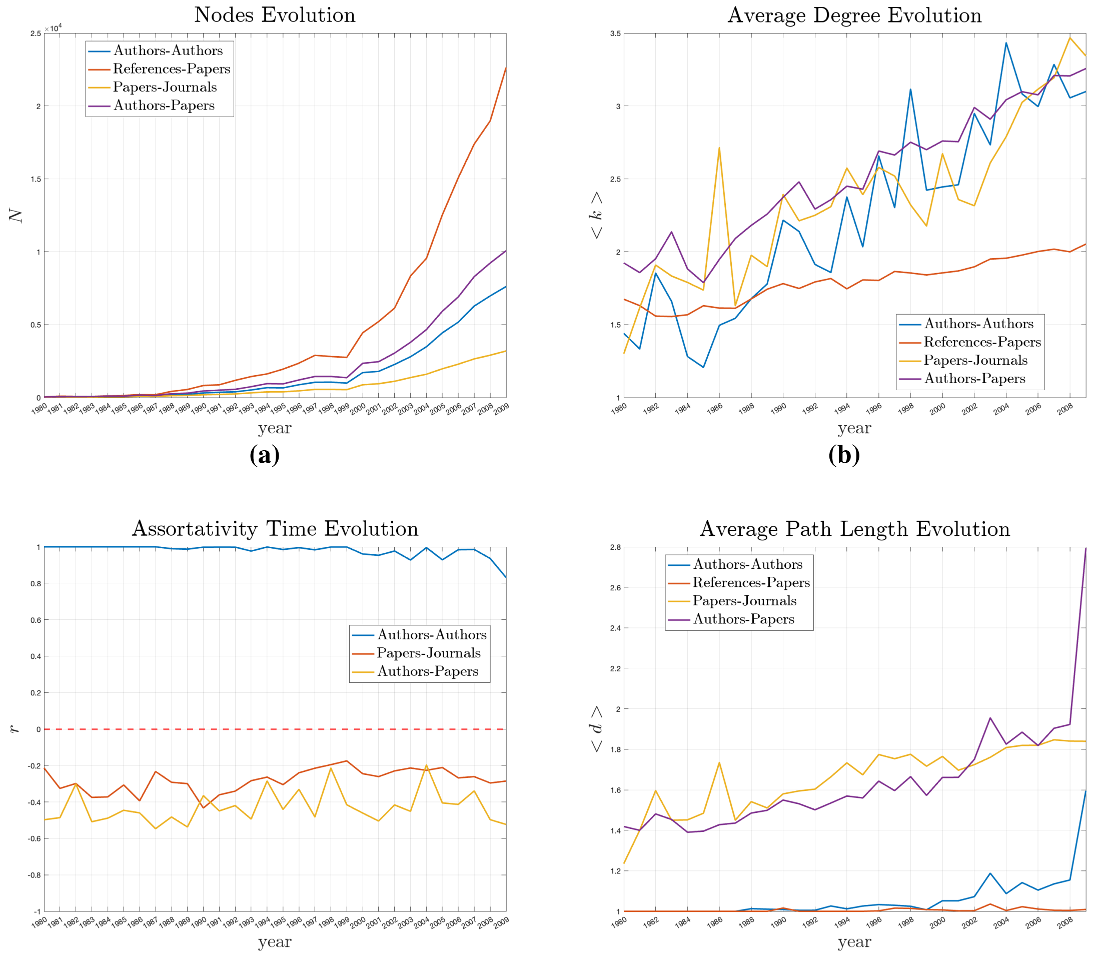
For oriented networks it is appropriate to consider a different indicator able to manage the presence of in-degree and out-degree. The used one is

$$r^{\rightarrow} = \frac{L^{-1} \sum_{(m,n) \in \mathcal{N}} k_m^{out} k_n^{in} - \left[L^{-1} \sum_{(m,n) \in \mathcal{N}} \frac{1}{2} (k_m^{out} + k_n^{in}) \right]^2}{L^{-1} \sum_{(m,n) \in \mathcal{N}} \frac{1}{2} ((k_m^{out})^2 + (k_n^{in})^2) - \left[L^{-1} \sum_{(m,n) \in \mathcal{N}} \frac{1}{2} (k_m^{out} + k_n^{in}) \right]^2} \tag{6}$$

Instead of r, the symbol $r^{\rightarrow}$ will be used in order to avoid notation ambiguity when we are referring to it. In particular $r^{\rightarrow}_{in-in}$ will be used for in-in degree, $r^{\rightarrow}_{in-out}$ for in-out degree $r^{\rightarrow}_{out-in}$ for out-in degree and $r^{\rightarrow}_{out-out}$ for out-out degree.

Based on the value taken by $r/r^{\rightarrow}$[1] it is possible to distinguish between three different cases,

[1] This notation means "r and $r^{\rightarrow}$", is not the ratio between the two assortativities.



(a)

(b)

(c)

(d)

Fig. 3: **(a)** Plot of the nodes number evolution per year; **(b)** Plot of the average degree evolution per year; **(c)** Plot of the assortativity evolution per year; **(d)** Plot of the average path length per year.

- $r/r^{\rightarrow} > 0$: that implies an assortative network,
- $r/r^{\rightarrow} = 0$: that implies a neutral network,
- $r/r^{\rightarrow} < 0$: that implies a disassortative network.

Betweenness Centrality: This indicator evaluates the importance of a node based on the number of shortest paths that pass through it, mathematically

$$B_C(i) = \sum_{n \neq i \neq m} \left(\frac{\sigma_{nm}(i)}{\sigma_{nm}} \right), \tag{7}$$

where

- $\sigma_{nm}(i)$ represents the number of shortest path between node n and m that pass through node i,
- σ_{nm} represents the number of shortest path between node n and m.

More insights of this indicator are given in the following article [1].

When performing normalization, it is important to note that each subgroup should be normalized using its actual number of elements, rather than the global one. To avoid this inconvenience, which would take a considerable amount of time, it has been decided to keep the value without normalization, as it is more truthful than the one normalized at the global level.

4.2 Technical Considerations

In this subsection all the presented graphs will be analyzed one by one from a technical view point, trying to categorize them. Just in subsection 4.3 we will try to give an explanation of the results obtained.

For better analyse the networks structure and classify them, three different years are taken as references: the first one (1980), the middle one (1995) and the last one (2009); in this way a more detailed overview is provided.

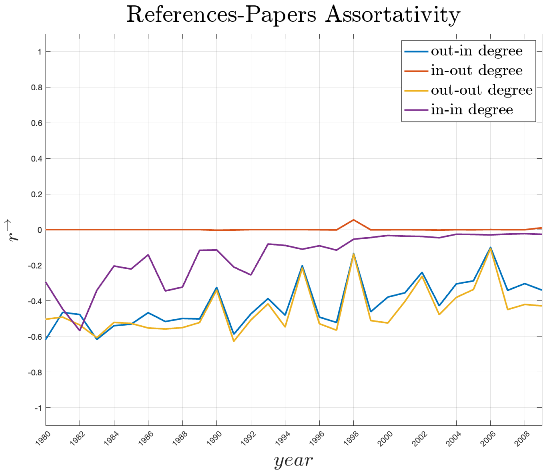


Fig. 4: Assortativity evolution per year of the References-papers network.

4.2.1 Authors-Authors Graph

As depicted in Fig.9 this type of graph follows a Poissonian distribution (2), typical of random networks, throughout the entire period considered in this study. In this case the randomness is given by the fact that authors collaborate without preferences.

This networks are strongly assortative, Fig.3c, given that all the authors that collaborate on a paper are connected between them, so the majority of the linked nodes have the same degree. From 1999 the assortativity starts to decrease a little bit, from $r \simeq 1$ to $r \simeq 0.8$, due to the birth of more complex kind of nodes connections, for example: two different research groups coordinated by a single person as connector. This phenomenon also has resonances on the average path length, Fig.3d. Since the end of the last century, it is evident that the emergence of more complex structures has led to a lengthening of the shortest paths lengths to connect nodes, as well as to new connections between nodes that were previously unconnected.

Combining results from Fig.3c and Fig.3b we see that while $\langle k \rangle$ is increasing, r remains almost constant, this allow us to conclude that no giant component is being born. This trend occurs because newer authors tend to collaborate more frequently among themselves rather than with older authors.

4.2.2 References-Papers Graph

Given the results displayed in Fig.12 we see that these kind of graphs follow a Power-law distribution (3), typical of scale-free networks. In this case some giant components are present, and each of them represents a different macro research area. The smallest ones instead represent niche research areas or areas for which there is limited data available in the dataset.

From Fig.4 four main conclusions can be drawn from assortativity $r^{\rightarrow}$:

- Considering $r^{\rightarrow}_{out-out}$ we see a disassortative behavior, in fact there is the tendency to connect to non-similar nodes. This limitation arises because many cited papers are represented in the dataset only by their ID, without their full information. Therefore, it is challenging to retrieve detailed information about these papers unless the dataset size is increased, which is currently not feasible due to the hardware constraints mentioned earlier. This phenomenon causes the connection of a nodes with an high out-degree, those whose information are fully available, to nodes with low out-degree, those whose information are not available.

- Considering $r^{\rightarrow}_{out-in}$ we see a disassortative behavior, almost equal to the out-out degree one. This phenomenon is probably due to the same problem enhanced in the previous point. It is reasonable to think that nodes belonging to same subgraph are likely to connect between themselves, but given that their information are not available this is not possible, and so both their out and in-degree is lower respect the actual ones.

- Considering $r^{\rightarrow}_{in-in}$ we see an initial disassortative behavior that has become basically neutral in recent year. Since 1999 the number of nodes for these graphs has exploded, Fig.3a, thus allowing the creation of more complex structures and more balanced connections. These complex structures appear because the number of recent papers cited increases compared to the past, and the most recent ones are more easily present in the dataset.

- The $r^{\rightarrow}_{in-out}$ shows a neutral behavior during all the analyzed period, so there seems to be no correlation between in and out degree when it comes to making connections.

Fig.3b and Fig.3d show that $\langle k \rangle$ and $\langle d \rangle$ slightly increase over 30 years. This happens because even if new connections born these are usually within new papers and, due to the dataset problem, 2^{nd} order connection are rarely possible, forcing $\langle d \rangle \simeq 1$.

4.2.3 Authors-Papers Graph

To study this kind of network it is interesting to make some separated considerations for the two sets contained: $\mathcal{U} = \text{authors}$, $\mathcal{V} = \text{papers}$. Fig.10 illustrates that the authors set follows a Power-law distribution, indicating that a small number of authors have an higher number of collaborations compared to the majority. Conversely, the papers set follows a Poissonian distribution, suggesting that most papers are authored by a similar number of contributors.

This network shows a disassortative behavior, Fig.3c. This is expected because most of the subgraphs assume star-shaped structures, where only a few authors contribute to multiple publications.

Fig.3b indicates that the number of authors contributing to papers is increasing over time. Additionally, Fig.3d reveals that the number of authors involved in multiple publications is also on the rise, in fact the increase in the average path length $\langle d \rangle$ can be attributed to authors collaborating on multiple papers, which leads to the formation of more distant connections between them.

4.2.4 Papers-Journals Graph

Even if this network is bipartite like the previous one, in this case analyzing both sets is not meaningful, in fact set $\mathcal{V} = \text{papers}$ is described by a degenerate distribution where all the nodes have $k = 1$. This happens because a paper can be published just on a single journal, forcing all the subgraphs to be star-shaped. This type of structure imposes a disassortative behavior, Fig.3c. In this configuration, the lateral nodes typically have a lower node degree k, whereas the central nodes exhibit a higher node degree k.

Set $\mathcal{U} = \text{journals}$ shows a more interesting individual behavior, in Fig.11 we see that journals nodes follow a Power-law distribution, so some bigger hubs are present representing the most influential journals in terms of publications number.

The increasing of $\langle d \rangle$, shown in Fig.3d, should not be confused with an increase in the complexity of subgraph structures, the increase is due to the number of published papers over a journal that moves the average toward 2. This value cannot be exceeded, in fact the maximum distance between two nodes in these graphs is $d = 2$.

For what concern $\langle k \rangle$ instead, its increase is due to the greater number of papers published on each journal, the only nodes whose node degree k is rising are the central ones in each star graph.

4.3 Real World Considerations

In this subsection we will attempt to explain the technical results seen above by associating them with possible causes in the real world and translating nodes and connections into actual papers, authors, and journals.

The first analysis was aimed at discovering the most popular research topics over the years. The most effective way to do this was to compile a top 5 of the journals with the most publications by each year, in Fig.5 results are shown. We see how over the years the three macro strands of research are: Computer Science, Electronics & Electrical Communications and Automation & Robotics. These results are not at all surprising; over the past 40 years, the greatest technological innovations have been in these three areas, revolutionizing people's lifestyles and the techniques by which companies provide products and services. These areas of research are also increasingly interconnected, it is not possible to talk about one of them without also considering the other two.

NOTE: It is interesting that although it is a hot topic nowdays, the study of artificial intelligence starts as early as the 1980s, and its possible practical applications have been studied since the 1990s for intelligent control systems & robotics and for computer vision.

During the technical analysis it has been seen that since the end of the 1990s the number of publications and collaborations between authors has in-

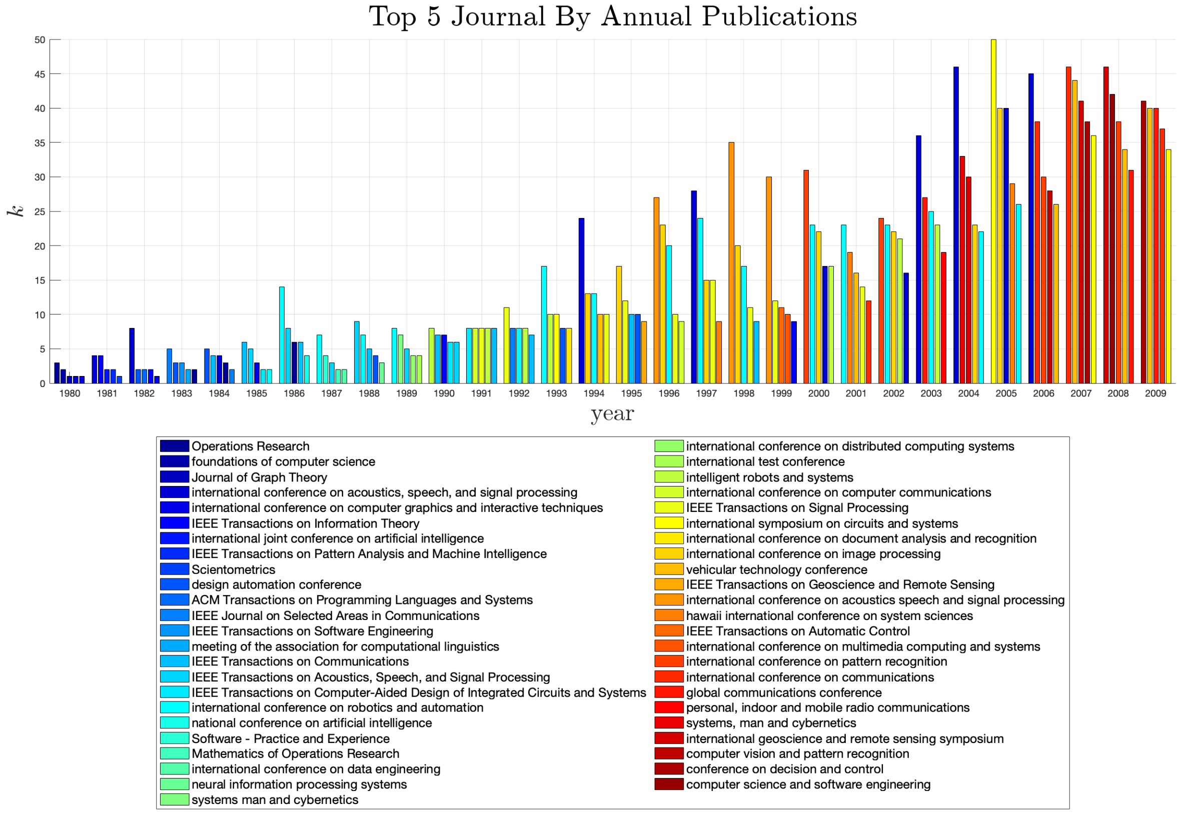


Fig. 5: Top 5 journals for number of pubblication on them by year.

creased considerably, this is certainly due to the global spread of the Internet. Thanks to it, getting in touch with researchers from all over the world and discovering new papers has become much easier.

It can be interesting to discover the top 5 cited papers by year, Fig.6, in order to find out what are the central themes for the study of the three macro areas highlighted above. Unfortunately, the variety of micro topics covered is much wider and it is no longer possible to clearly identify the main research themes, indicating that some of these papers have cross-disciplinary importance. A more detailed analysis of each individual paper should be conducted to draw stronger conclusions.

Thanks to Fig.7 it is possible to see how the most influential authors in terms of publications are almost always different with each passing year, underlining how dynamic and based on collaboration, rather than individuals, the research world is. Furthermore, Fig.8 shows the top 5 authors in terms of betweenness centrality (7) by year, and it is evident once again that after the mid-1990s the number of collaboration is exploded and the team work began to be central in achieving important results, more then ever before.

5 Conclusions

In this project, it was possible to fulfill the established goal of finding out how the scope of research in the years between 1980 and 2009 was desired through the use of graph theory. The results obtained seem to be consistent with the historical-technological events. There are some aspects, which cannot be quantified through observation of the data alone, that might be interesting to point out[2].

The increase in the number of researchers, and the resulting increase in published papers, clearly indicates how the importance of research is increasingly being recognized. This event is due to: the already explained worldwide spread of the Internet, the increasing level of specialization required for a job position and the ever-increasing competition in academia. It is not necessarily the

[2] These considerations are based solely on personal assumptions and comparisons with people who work in this field; they are not supported by any scientific evidence reported in this work.

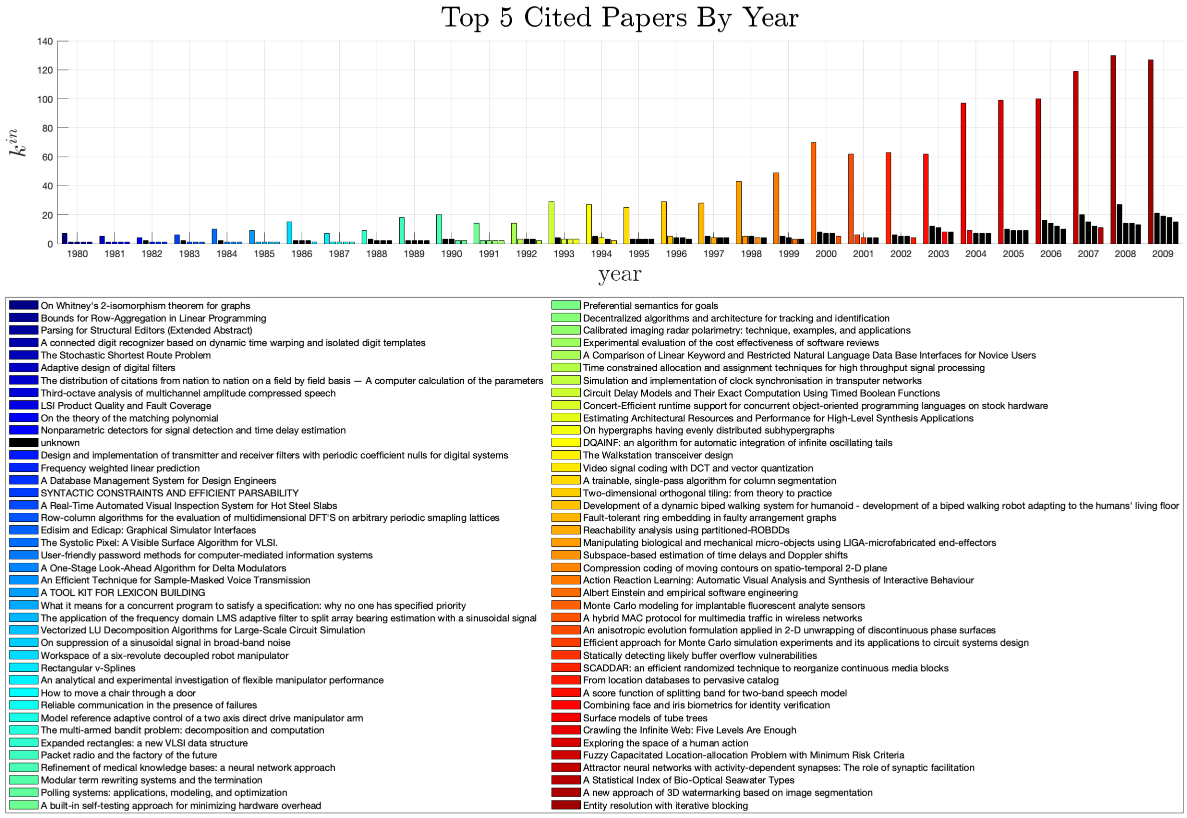


Fig. 6

case that if the number of papers is increasing so is their quality, nowadays it is of central importance to have many publications and citations in order to succeed in academia, but this may lead in some cases to an increase in quantity at the expense of quality. This consideration is not meant to be denigrating toward those who do research; every step forward, no matter how small, can be crucial for future research. However, it is important to emphasize also the negative aspects as well for the purpose of this project.

5.1 Limitations & Future Works

The main technical limitation encountered during this work is due to the use of a small amount of data compared to the total available. Although data were taken randomly, this may not have been sufficient to eliminate some bias that led to erroneous or inaccurate conclusions.

A limitation of human nature, on the other hand, is that this work was done by one person. As also evidenced by the data, collaboration is essential to be able to compare and develop more complex and effective ideas.

In case this paper is shared as a reference, suggestions for extending and improving this work are:

- Increase the dimension of the used datasets. An idea is to use multiple devices for processing different data at the same time, and then average appropriately the final results,
- Overcome the hardware deficiency relying on virtual machines for parallel computation, or more simply use the strategy proposed in the previous point,
- Work in a group of at least 2/3 people, preferably with different skills and knowledge, to divide the work and encourage discussion.

One possible extension may be to use the employed metrics to train a regression model capable of predicting the number of citations a paper will receive, so that we can try to anticipate its success in terms of circulation.

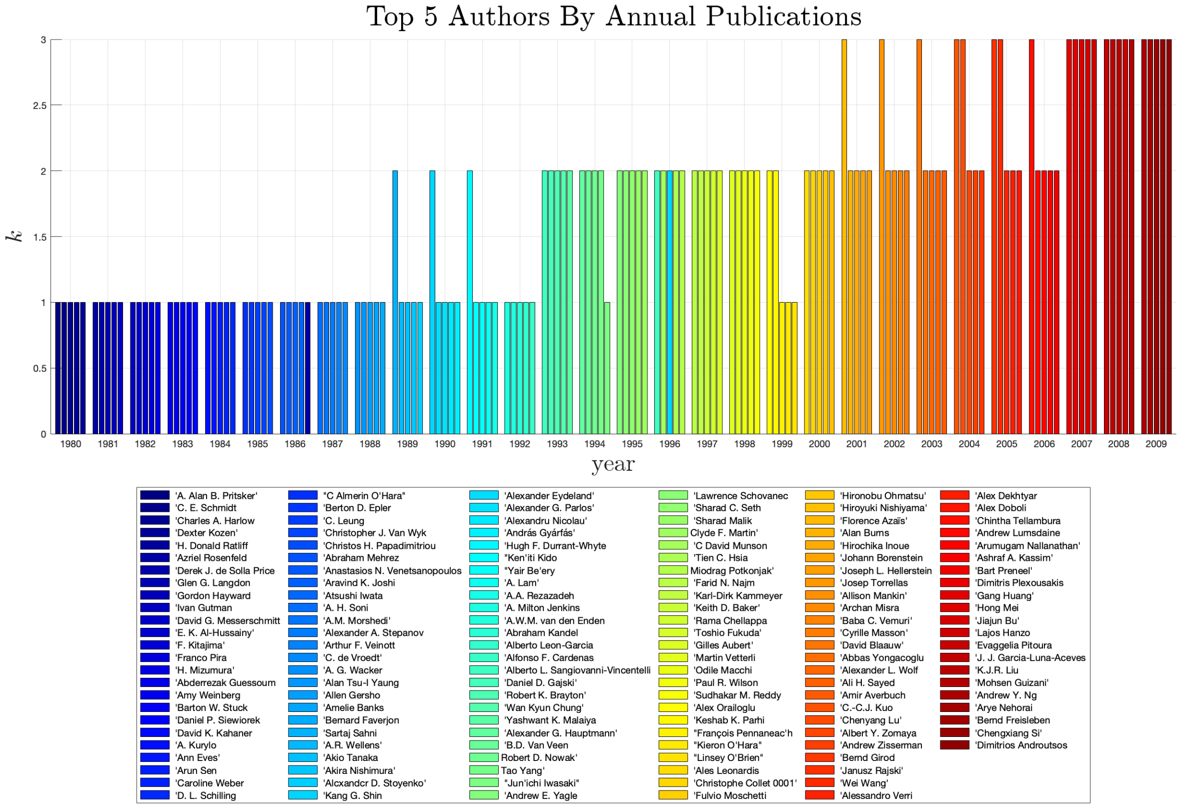

Top 5 Authors By Annual Publications

Fig. 7: Top 5 authors by year based on the number of publications.

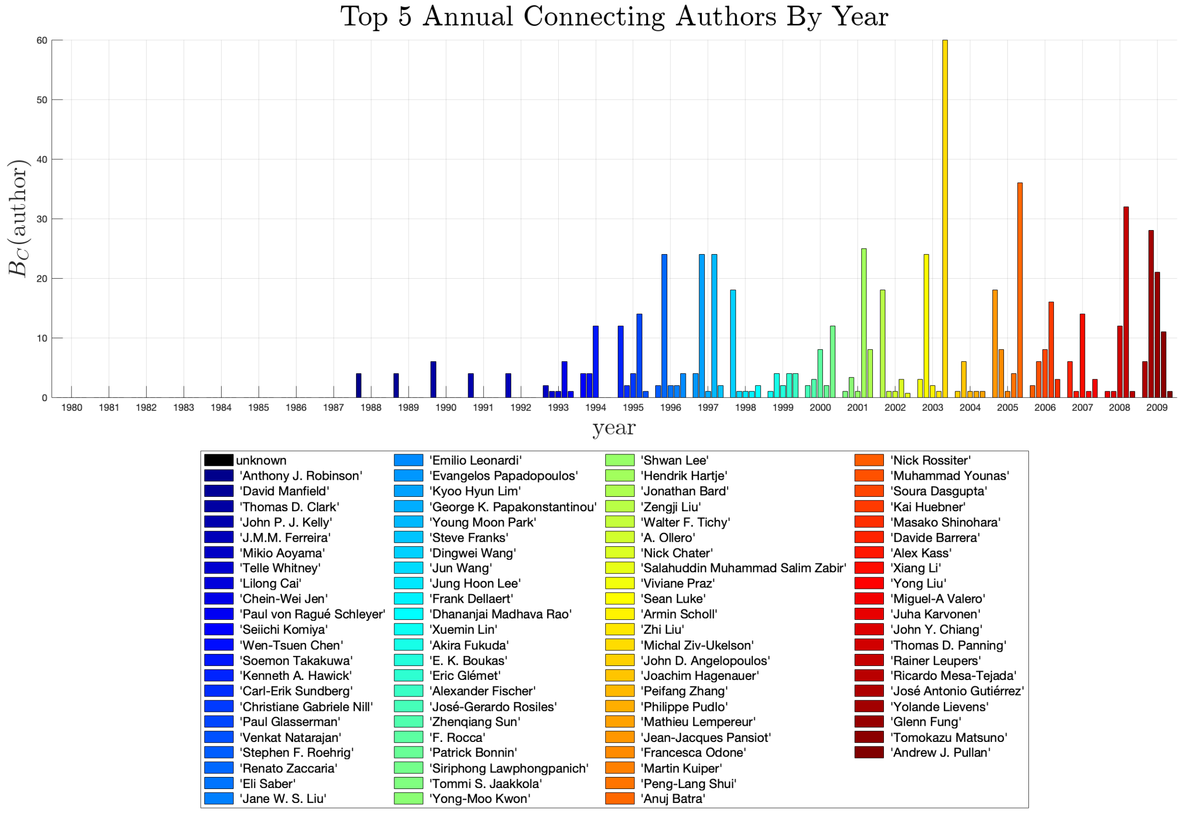

Top 5 Annual Connecting Authors By Year

Fig. 8: Top 5 authors by year based on their connectivity relevance.

A Degree Distribution Images

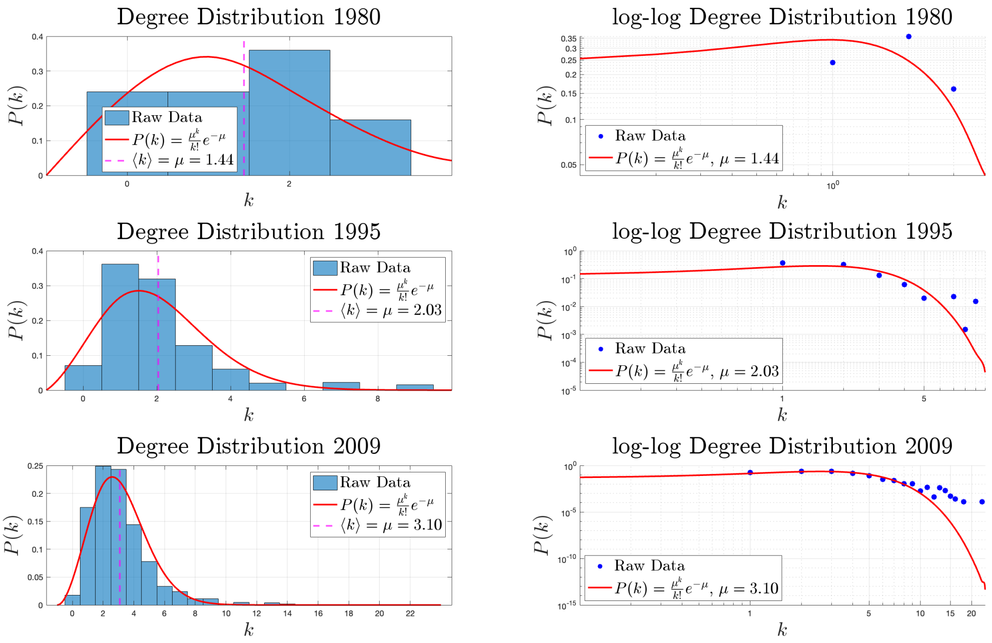


Fig. 9: Degree distribution of the Authors-Authors graph during the years 1980, 1995 and 2009.

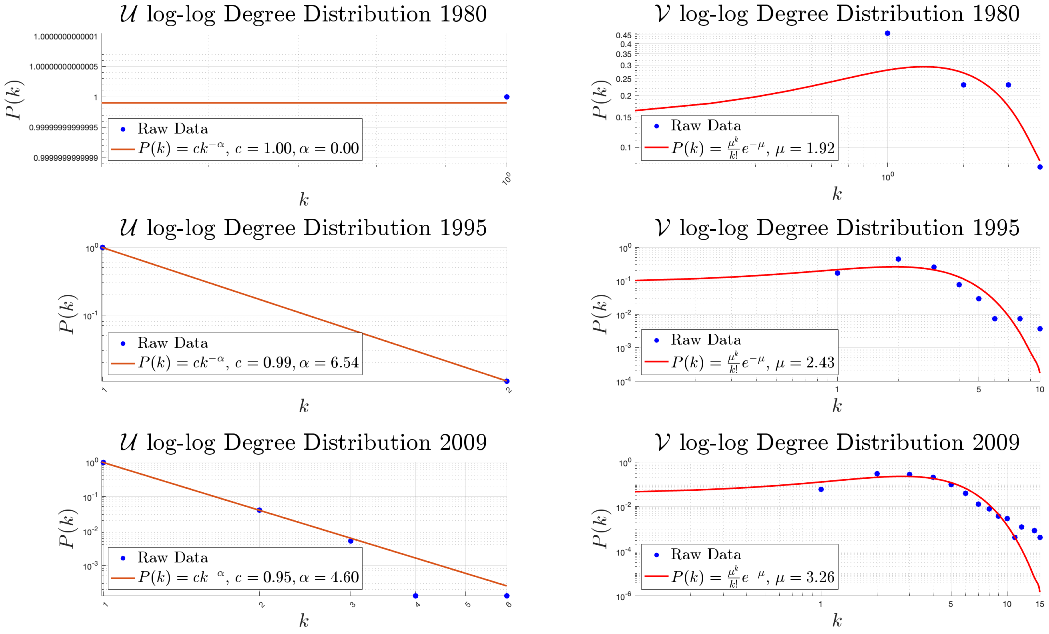


Fig. 10: Degree distribution of the Authors-Papers graph during the years 1980, 1995 and 2009. $\mathcal{U}$ indicates the sets of authors nodes, while $\mathcal{V}$ represents the sets of papers nodes.

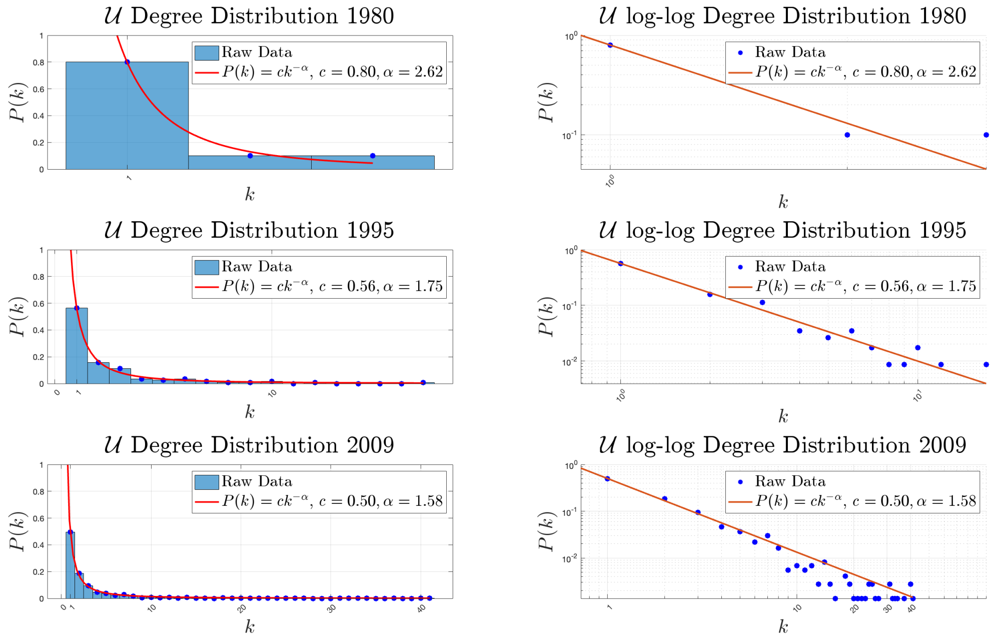


Fig. 11: Degree distribution of the Journals-Papers graph during the years 1980, 1995 and 2009. $\mathcal{U}$ indicates the sets of journals nodes.

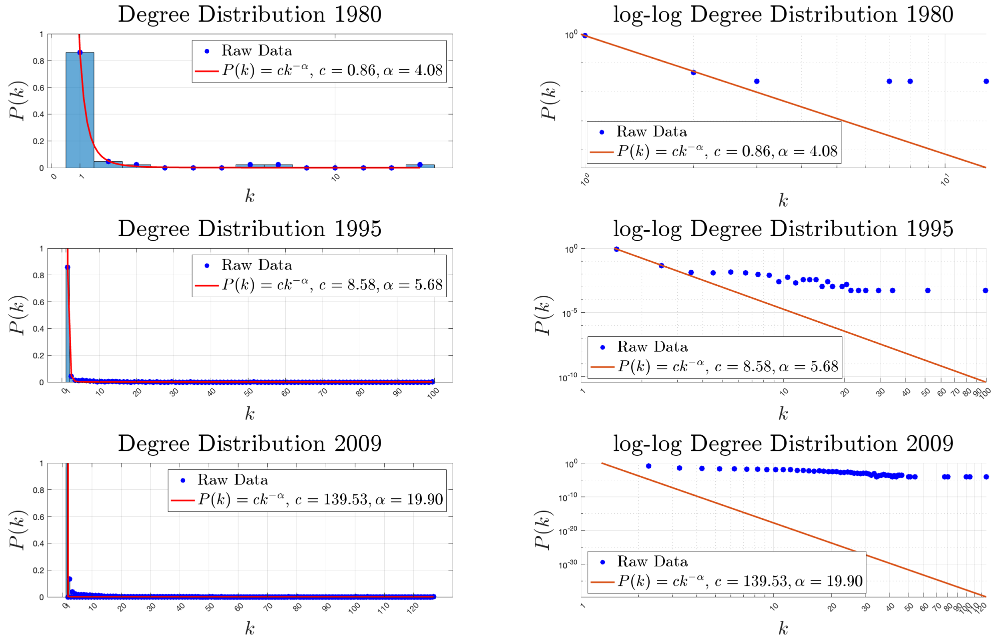


Fig. 12: Degree distribution of the References-Papers graph during the years 1980, 1995 and 2009.

References

[1] M. Barthélemy. “Betweenness centrality in large complex networks”. In: *The European Physical Journal B* 38.2 (Mar. 2004), pp. 163–168. ISSN: 1434-6036. DOI: 10.1140/epjb/e2004-00111-4. URL: https://doi.org/10.1140/epjb/e2004-00111-4.

[2] Michael Fire and Carlos Guestrin. “Over-optimization of academic publishing metrics: Observing Goodhart’s Law in action”. In: *GigaScience* 8 (June 2019). DOI: 10.1093/gigascience/giz053.

[3] M. E. J. Newman. “Assortative Mixing in Networks”. In: *Phys. Rev. Lett.* 89 (20 Oct. 2002), p. 208701. DOI: 10.1103/PhysRevLett.89.208701. URL: https://link.aps.org/doi/10.1103/PhysRevLett.89.208701.

[4] Mikail Rubinov and Olaf Sporns. “Complex network measures of brain connectivity: Uses and interpretations”. In: *NeuroImage* 52.3 (2010). Computational Models of the Brain, pp. 1059–1069. ISSN: 1053-8119. DOI: https://doi.org/10.1016/j.neuroimage.2009.10.003. URL: https://www.sciencedirect.com/science/article/pii/S105381190901074X.

[5] Jie Tang et al. “ArnetMiner: Extraction and Mining of Academic Social Networks”. In: *KDD’08*. 2008, pp. 990–998.