Scientific Collaborations in High Performance Computing
CMSC828O – Final Poster
Onur Cankur

Background and Motivation

Background

- The collaboration between researchers forms co-authorship networks [1, 4, 5, 6].
- A co-authorship network is a class of social networks where the nodes represent the researchers, and the links represent the collaboration between them.
- High performance computing (HPC) is a research area in computer science (CS) that involves both core CS and very multidisciplinary topics.
- Network measures such as degree distribution, assortativity, and clustering coefficient help to compare the network with previous studies.
- Greedy and Louvain clustering algorithms can create partitions with high modularity scores. The partitions created can be compared using the Normalized Mutual Information (NMI) measure.

Motivation

- Explore the co-authorship patterns in HPC.
- Create the first comprehensive dataset and co-authorship networks specific to HPC researchers.
- Allow HPC researchers to examine the current status of the field and how it has been moving forward.

Network Description and Research Questions

Network Description

- The nodes represent the authors. The links represent the co-authorship.
- Contains 119428 nodes and 468969 unweighted and undirected links.

Research Questions

1. Do the collaborators of a scientist also collaborate with each other?
2. Do the most connected researchers connect more nodes next year?
3. Do researchers form clusters based on the venues where they published most of their papers at?

Methodology – Constructing the Network

- The data is collected from DBLP, a computer science bibliography that provides a dataset and an API that allows collecting its data [3].
- 119428 authors and 92312 papers are collected from 35 different venues that are directly related to HPC.
- The data contains papers published from 1985 to 2022.

Specifically, this data is collected using the following strategy:

1. Collect all papers published in a venue using the *requests* library in Python.
2. Store paper and author information for each paper using the Neo4j graph database and make the necessary connections using the neomodel library in Python.
3. Iterate over 35 venues and follow the previous steps.

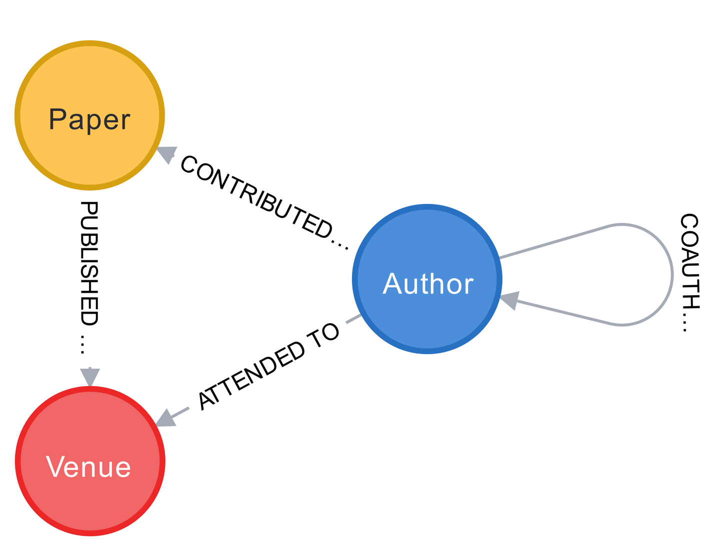

Figure 1: Neo4j Database Schema. The whole dataset contains three different nodes: authors, papers, and venues and four different links: author to author, author to paper, author to venue, and paper to venue.

Results

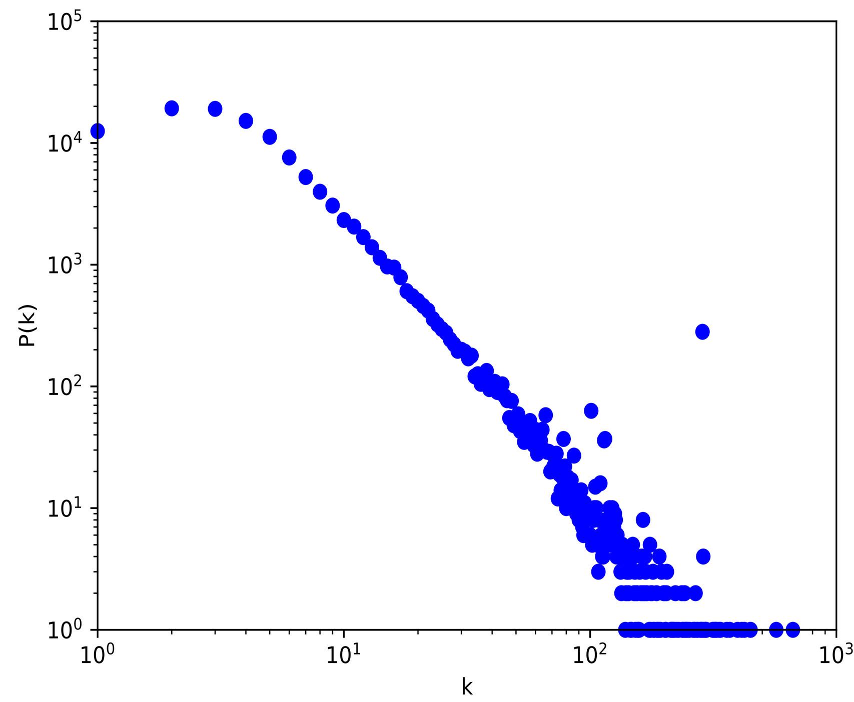

Figure 1: Degree distribution of the co-authorship data on double logarithmic axis (uniformly binned log-log plot). Gathered from papers published between 1985 and 2022.

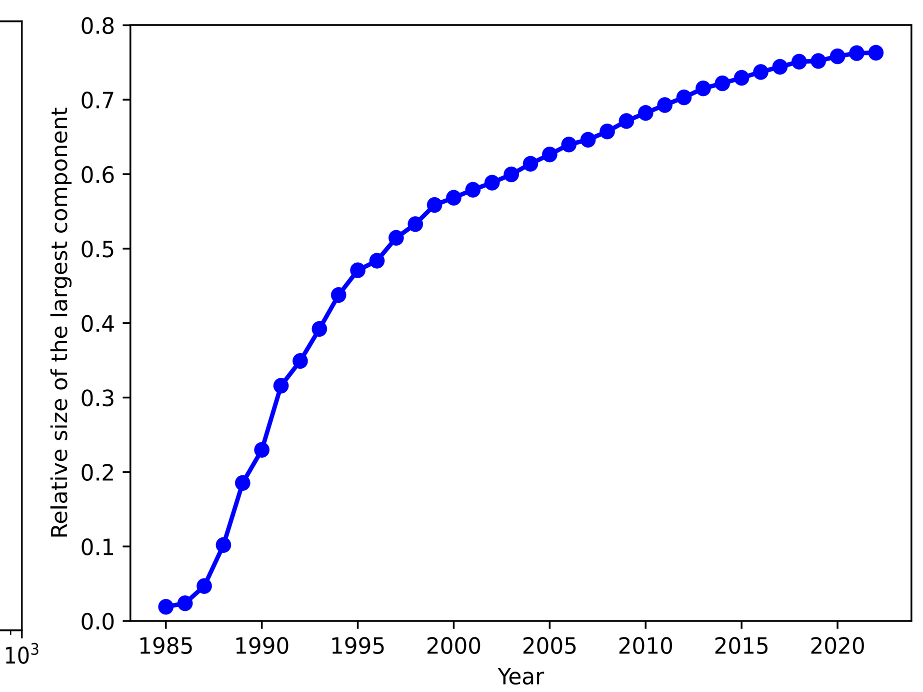

Figure 2: The change in the relative size of the largest component. The results are cumulatively computed up to the years on the x-axis.

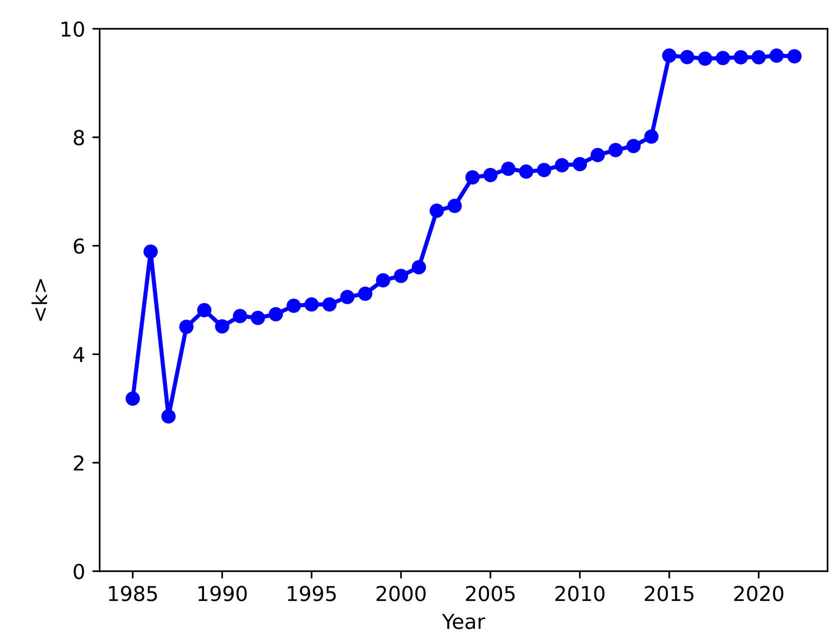

Figure 3: The change in the average degree of the largest component. The results are cumulatively computed up to the years on the x-axis.

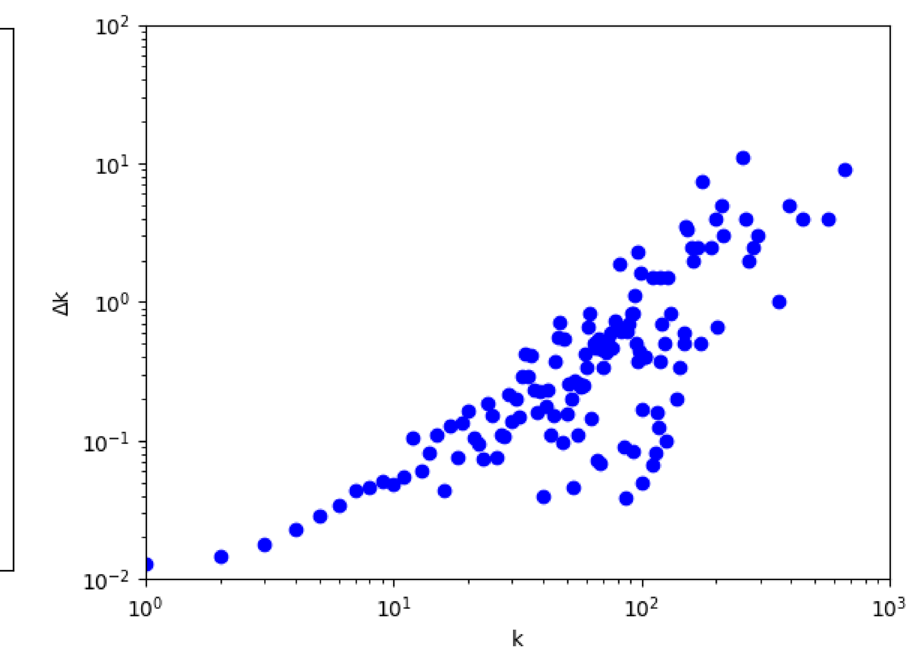

Figure 4 - The change in the number of degrees for authors that have k degrees in 2021. The x-axis represent the authors that have k degrees in 2021. The y-axis represents the change in the number of degrees from 2021 to 2022.

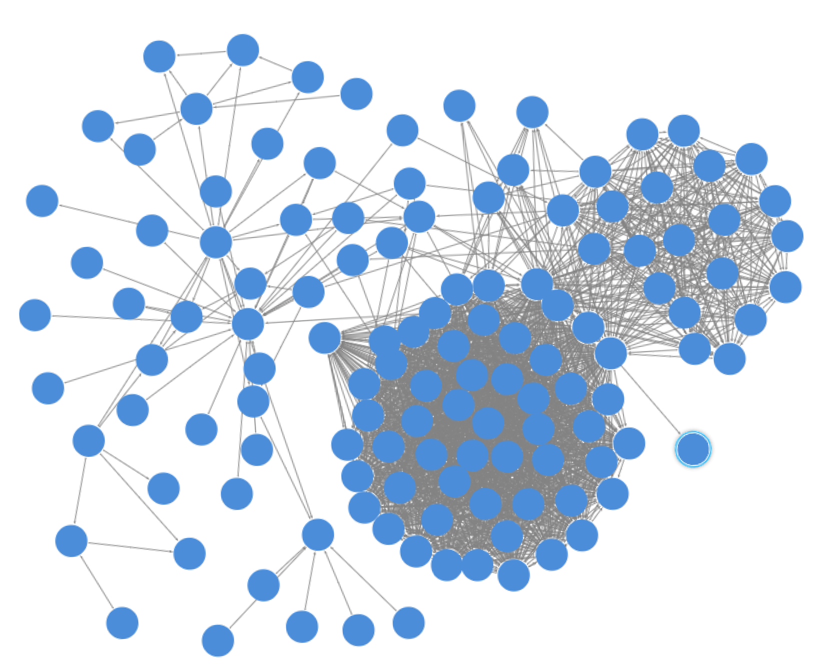

Figure 5: A subgraph from the co-authorship network. Not all links are visible. The visualization is generated using the Neo4j Bloom tool.

Table 1: Summary of the statistics for the co-authorship data. Cumulative result up to 2022.

Clustering algorithm	Number of communities	Modularity	NMI Score
Ground truth	35	0.48	-
Greedy (default)	1338	0.73	0.18
Greedy	35	0.72	0.09
Louvain	105	0.8	0.12

Table 2: Summary of the statistics for the co-authorship data. Cumulative result up to 2022.

Number of authors	119428	Number of papers	92312
Authors per paper	2.66	Papers per author	3.44
Largest component	91130	2nd largest component	39
As a percentage	(76.3%)	Collaborators per author	7.85
Largest distance	19	Average distance	6.1
Assortativity	0.72	Clustering coefficient	0.74

Methodology – Network Analysis

- A separate co-authorship network is created for every year from 1985 to 2022 by cumulatively computing the collaborations up to each year.
- Neo4j data is exported in *graphml* format, which can be read in NetworkX.
- Analyses are performed using the NetworkX library.
- Clustering coefficient tells how a node's neighbors connected with each other. Therefore, it is used for the first research question.
- The change in the number of degrees for authors that have k degrees in the previous year is observed for the second research question.
- For the last research question, the authors are grouped by the venues where they published most of their papers and created a cluster, which is called ground truth, for each venue. Then the ground truth partition is compared with the partitions created by other clustering algorithms.

Discussion and Conclusion

Discussion

- Similar to the previous studies on collaboration networks, it follows a power law distribution, hence it is a scale-free network [4, 5, 6] (Figure 1).
- The largest component significantly increases from 1985 to 2022 (Figure 2). The average degree increases linearly over time (Figure 3). The results are similar to [1].
- The HPC co-authorship network has similar characteristics to co-authorship networks studied by Newman (Table 2) [5, 6].
- RQ1: Collaborators of a scientist also collaborate with each other (Table 2).
- RQ2: The most connected nodes become more connected next year (Figure 4).
- RQ3: Forming clusters based on the venues where the authors published most of their papers does not give the best partition (Table 1).

Alternative Approaches

- More advanced community detection algorithms can be used.
- The second question can be further investigated by considering the existing node to new node and existing node to existing node connections separately.

Impact

- Provides the first dataset and co-authorship networks specific to HPC researchers.
- Makes it possible to conduct more advanced studies on collaboration networks.
- Allows HPC researchers to learn more about the field.

Future Directions

- The citation information can be added to the study either by scraping from Google Scholar or creating a proxy to ask more complex questions.
- A webpage with an interactive network visualization that includes all the information can be created for HPC researchers to explore the field.

References

[1] Barabâsi, Albert-Laszlo, et al. "Evolution of the social network of scientific collaborations." Physica A: Statistical mechanics and its applications 311.3-4 (2002): 590-614.

[2] DBLP Dataset. https://dblp.org/xml/release/dblp-2022-03-01.xml.gz

[3] Ley, Michael. "DBLP: some lessons learned." Proceedings of the VLDB Endowment 2.2 (2009): 1493-1500.

[4] Liu, Xiaoming, et al. "Co-authorship networks in the digital library research community." Information processing & management 41.6 (2005): 1462-1480.

[5] Newman, Mark EJ. "Coauthorship networks and patterns of scientific collaboration." Proceedings of the national academy of sciences 101.suppl 1 (2004): 5200-5205.

[6] Newman, Mark EJ. "Scientific collaboration networks. I. Network construction and fundamental results." Physical review E 64.1 (2001): 016131.