Scientific Collaborations in High Performance Computing

Onur Cankur

Motivation

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

Background

- HPC is a research area in computer science
  - Involves both core CS and very multidisciplinary topics.

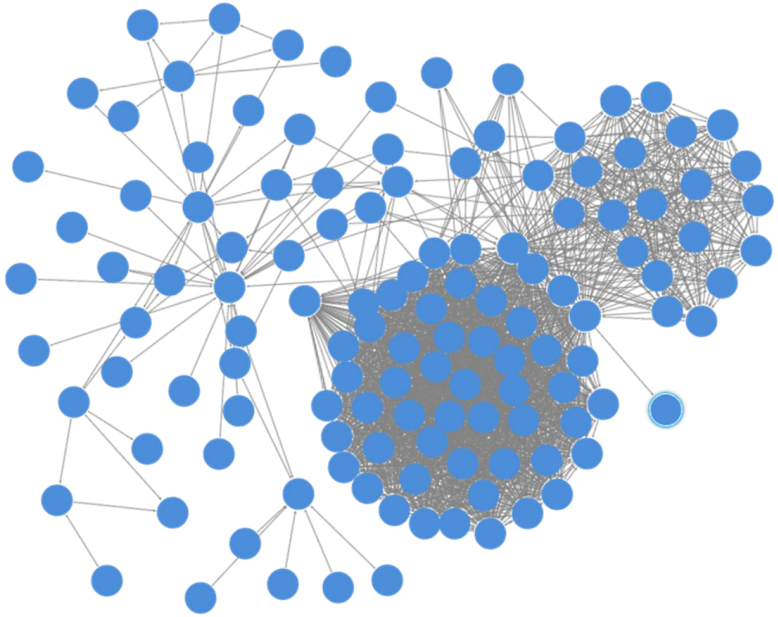
- The collaboration between researchers forms co-authorship networks.

- Nodes represent the researchers, and the links represent the collaboration between them.

Network Description

- Nodes represent the authors.
- Links represent the collaboration between authors.

- 119428 nodes.
- 468969 unweighted and undirected links.



Created using Neo4j Bloom

Research Questions

1. Do the collaborators of a scientist also collaborate with each other?

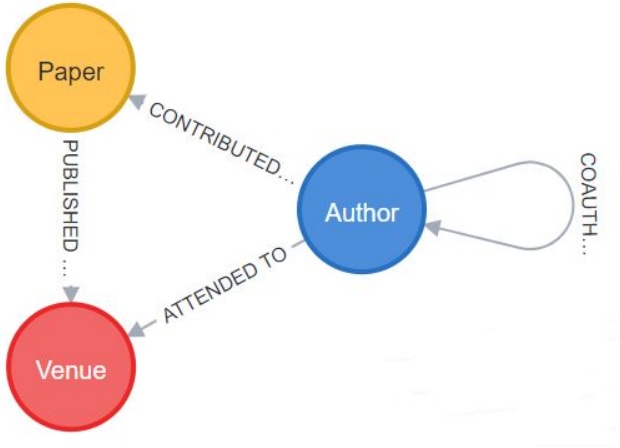
2. Do the most connected researchers connect more nodes next year?

3. Do researchers form clusters based on the venues where they published the most of their papers at?

Collecting the Data

1. Collect all papers published at 35 different venues from DBLP computer science bibliography.

2. Create objects and relationships and store them using *Neo4j graph database*.



Neo4j Database Schema

Analyzing the Network

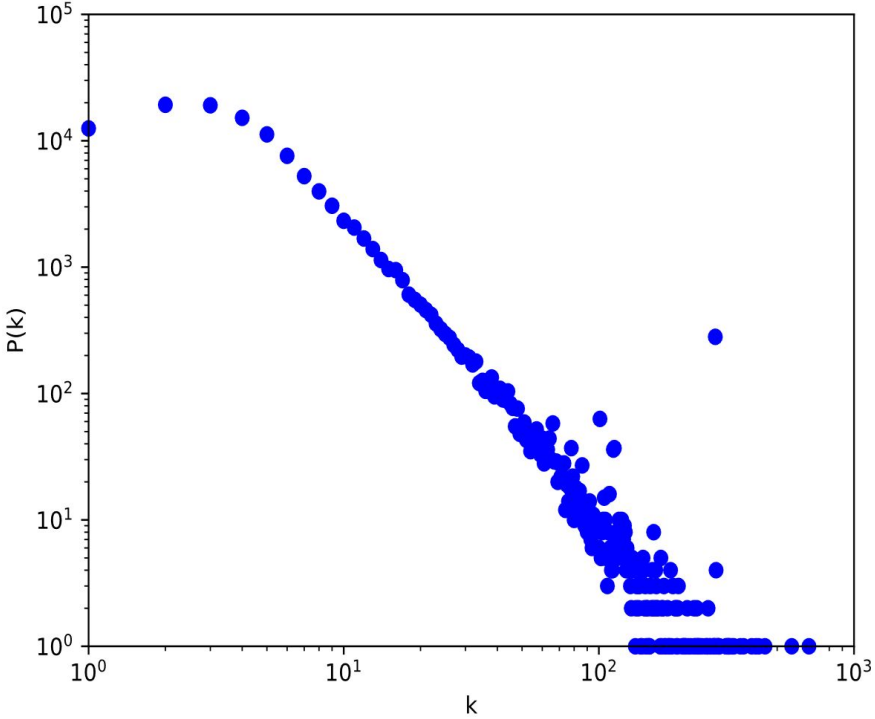
- A separate co-authorship network is created for every year from 1985 to 2022 by cumulatively computing the collaborations up to each year.

- Neo4j data is exported in *graphml* format and imported to *NetworkX*.

- All analyses performed using the *NetworkX* library.

Degree Distribution

- A few authors have high degrees while most authors have low degrees.
- Follows a power law distribution.



Network Statistics

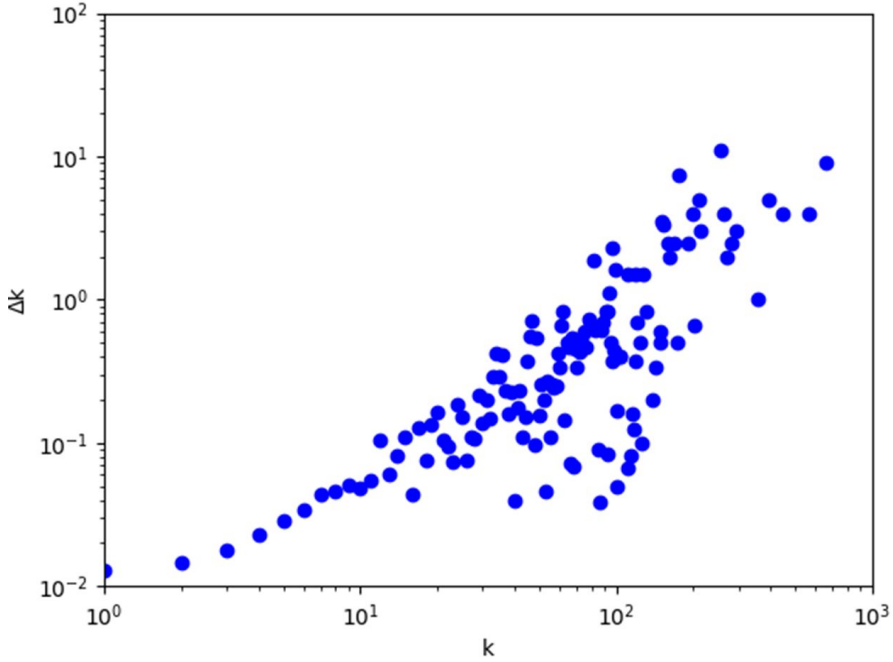
- Similar to previous studies on collaboration networks.
- Higher assortativity and clustering coefficient.
- **RQ1**: Collaborators of a scientist also collaborate with each other.

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

Attachment

- Observed the change in the number of degrees for authors that have k degrees in 2021.

- **RQ2**: The most connected researchers connect more nodes next year.



Forming clusters based on venues

- Ground truth: grouped the authors by the venues that they published most of their papers at and created a cluster for each venue.
- **RQ3**: Forming clusters based on the venues where the authors published the most of their papers at does not give the best partition.

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

Summary

- Provides the first dataset and co-authorship networks specific to collaborations between HPC researchers.
- Investigates characteristics of HPC collaboration networks.
- Allows HPC researchers to learn more about the field.
  - Find the most connected researchers, look at their works, collaborations, etc.

Fun Fact

- The second most connected researcher, Jack Dongarra, was awarded the Turing Award this year.
  - Also called Nobel Prize of computing.

References

- 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.
- DBLP Dataset. https://dblp.org/xml/release/dblp-2022-03-01.xml.gz
- Ley, Michael. "DBLP: some lessons learned." Proceedings of the VLDB Endowment 2.2 (2009): 1493-1500.
- Liu, Xiaoming, et al. "Co-authorship networks in the digital library research community." Information processing & management 41.6 (2005): 1462-1480.
- Newman, Mark EJ. "Coauthorship networks and patterns of scientific collaboration." Proceedings of the national academy of sciences 101.suppl 1 (2004): 5200-5205.
- Newman, Mark EJ. "Scientific collaboration networks. I. Network construction and fundamental results." Physical review E 64.1 (2001): 016131.