Are there meaningful differences in collaboration networks among scientists from different domains? Or Does network science affect networking among scientists?

University of Padova
Learning from Networks Project
Hamid Mohammadi
January 2025

Author's Note: The chapter for conclusion is based on the author's limited experience and shallow understanding of networks. Surly, a more enlightened individual would make a different judgment based on the numbers reported here. Most of the time spent on this work (both literature review and programming) has led us to the fact that using a normalization factor for (at least) these kinds of networks is not helpful and adds more complexity and bias. Unfortunately, procedures with these kinds of outputs cannot be presented in detail (summarized in Chapter IV).

***Abstract*—This work aims to analyze the different properties of four co-authorship networks across four scientific fields. We divided this project into two phases. Phase I consists of computing each network's properties from node-level to graph-level. In Phase II, we investigated the principles of the theory of Homophily and small-world network properties. Then, based on our outcome in this project, we have a section about why trying to normalize the degree of nodes is not a good idea.**

***Index Terms*—Clustering coefficient, Assortativity, Homophily, Small-world properties.**

I. Introduction

In human society, every individual belongs to specific networks. Comprehending how our networks are structured provides valuable information on our understanding of our position and plays a key role in our future decisions. Collaboration networks convey information about how researchers across different scientific domains interact and form connections.

The primary focus of this work is to examine the differences in network structure across the four datasets and investigate whether these differences align with the principles of Homophily and small-world network properties. Homophily refers to the tendency of similar nodes to connect, which can be measured through degree assortativity. Small-world properties, on the other hand, are characterized by a network's high clustering coefficient and short average path length. By comparing these networks to degree-preserving random graphs, we aim to identify whether the real-world networks exhibit small-world characteristics. This report is structured into two main phases of analysis. Phase I is computing key network properties, such as average degree centrality, clustering coefficients, and assortativity measures. Phase II delves deeper into analyzing the principles of Homophily and small-world properties based on the results from Phase I.

Through this analysis, our aim is to uncover patterns that reveal whether collaboration tendencies differ across scientific fields and whether the study of network science itself influences how researchers form connections. Additionally, special attention is given to the CSphd network, which requires unique considerations due to its directed and disconnected nature. This network presents a distinct case for analyzing assortativity in directed graphs. In this work, we analyze four scientific collaboration networks, each representing co-authorship connections among researchers where nodes represent researchers and edges are co-authorship relationships. Table I.

TABLE I: Network Data Statistics

Dataset	Number of Edges	Number of Nodes
AstroPh	198.1K	18.8K
MathSciNet	820.6K	332.7K
netscience	914	379
CSphd(directed)	1.7K	1.9K

Finally, we assume that collaboration between researchers is represented by co-authorship. This is essential for the discussion and methodological purposes.

II. Phase I
Network Properties

The capstone of deriving meaningful insights through advanced technical computations is computing the network's structural properties like Global Clustering Coefficient, Average shortest Path Length, Average Degree, and Assortativity (Table II).

However, the average path length cannot be defined for the CSPhD network because of its disconnectivity, and because it is a directed network, computing Assortativity is also different, and all values are reported in (Table III). First, we tried to solve the

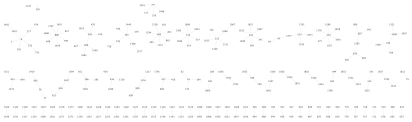

Fig. 1: CSphd Network.

issue of calculating the average shortest path length by considering it for each connected component and weighing them by component size. Then, we concluded that this action was not approved because we were introducing some properties that were not defined. Because this network hosts a high level of fragmentation (Fig. 1) and lacks paths between many nodes, it does not seem meaningful to compute the weighted average shortest path length. Indeed, we have found programming ways to compute it by weighing and not considering non-values, but we decided not to report them because it would produce unreliable and misleading results.

III. Phase II
Analyzing Network Features: Homophily and Small-World Properties

In this stage, we analyzed the Homophily and Small-world properties of the four mentioned networks. For this study, as mentioned in previous sections, for our analysis, we consider four networks, one of which is directional and disconnected (CSPhD).

A. Homophily

Generally, there are two types of homophiles: Attribute Homophily and structural Homophily. Because our datasets do not have attributes, we compensate on structural Homophily. Structural homophily refers to the tendency of nodes to connect based on similar structural properties, such as their degrees. Degree assortativity is a measure of this specific type of homophily. Therfore, in this study, we are considering degree assortativity.

Structural Homophily and Degree Assortativity: Degree Assortativity refers to the tendency of nodes in a network to connect with other nodes that have a similar degree. In simple terms, it measures whether nodes with many connections (high degree) are more likely to link to other highly connected nodes or whether they prefer to connect with less connected nodes.

- Positive Assortativity (AstroPh: $r = 0.201$):
  - Homophily Interpretation: Nodes with similar degrees (e.g., high-degree with high-degree or low-degree with low-degree) are more likely to connect.
  - Real-World Example: Social networks where influential people (hubs) tend to interact with other influential people, and less-connected individuals tend to interact with similarly less-connected individuals.
- Weak Positive Assortativity (MathSciNet: $r = 0.103$):
  - Homophily Interpretation: There is a mild tendency for nodes of similar degree to connect, but it's weak.
  - Real-World Example: A mixed network where connections are influenced by both degree similarity and random factors, leading to weaker assortative behavior.
- Negative Assortativity (netscience: $r = -0.082$):
  - Homophily Interpretation: High-degree nodes (hubs) are more likely to connect with low-degree nodes (spokes), and vice versa.
  - Real-World Example: Collaboration or hierarchical networks, where central hubs connect to many peripheral nodes. For example, in a co-authorship network, a renowned researcher (hub) collaborates with many junior researchers (low-degree nodes).

As for the CSPhD Network, because it is a directed graph, we should have chosen a different approach for calculating assortativity. However, surprisingly, in this network, all assortativity values are negative, pointing to a disassortative mixing pattern. In disassortative networks, central hubs (e.g., highly cited researchers or prolific authors) connect to fewer central nodes[1]. This might reflect that the CSPhD network behaves like an academic hierarchy, where professors/advisors (hubs) are connected to students (low-degree nodes).

B. Small-World Properties

In this section, we are reporting the analysis of the small-world properties (σ) of our networks. In Phase I, we calculated the clustering coefficient and average shortest path length for our networks, as shown in Table III. To calculate small-world coefficients (1), we need to generate a random graph and compute its clustering coefficient and average shortest path length. We did this by generating degree-preserving random graphs through edge-swapping to ensure a more realistic baseline for comparison. While maintaining the original degree distribution, this approach makes the comparison more appropriate for real-world networks. Additionally, For the "MathSciNet" network, computing the exact Average Shortest Path Length (APL) is computationally expensive, so we decided to use Random Node Sampling as an approximation method to compute its properties. This simple method Randomly samples a subset of nodes and takes the average of computed properties over sampled pairs.

Small-world coefficient (σ) is calculated based on the clustering coefficient and average path length with this formula:

TABLE II: Networks Properties

Dataset	Global Clustering Coefficient	Average Shortest Path Length	Average Degree	Assortativity
AstroPh	0.3178	4.1940	21	0.201317
MathSciNet	0.1367	7.3129	4	0.103010
netscience	0.4306	6.0419	4	-0.081678

TABLE III: Assortativity for CSphd Directed Graph

Dataset	In-In	In-Out	Out-In	Out-Out
CSphd	-0.0250	-0.0846	-0.0748	-0.0397

$$\sigma = \frac{\frac{C}{C_{\text{random}}}}{\frac{L}{L_{\text{random}}}} \tag{1}$$

Where:

- C = Clustering coefficient of the real network.
- C_{random} = Clustering coefficient of the corresponding degree-preserving random graph.
- L = Average shortest path length of the real network.
- L_{random} = Average shortest path length of the corresponding degree-preserving random graph.

As shown in Table IV, the "AstroPh" network exhibited a large small-world coefficient (σ)=24.9098, which means it has a highly clustered structure compared to its randomized counterpart while still maintaining relatively short path lengths. This suggests that collaborations in the AstroPh network tend to form tight-knit research groups while still being efficiently connected across the network.

"Netscience," on the other hand, has (σ)=10.8031, which is relatively large but not as extreme as "AstroPh". This reveals the fact that while the Netscience network also exhibits small-world properties, its structure is more balanced between clustering and efficiency, meaning that researchers are connected through both tightly clustered subgroups and moderately short paths. We tried to generate several random graphs for "MathSciNet," but unfortunately, all of them exhibited very low global clustering coefficients, suggesting an issue with the edge-swapping process. Because of this, (σ) is unreasonably high, which is not physically meaningful. We believe this is due to the fact that the randomized networks lost their structure completely.

C. Small-world Properties for CSphd network

In this section we discuss the biggest unresolved issue of this work. For generating a degree-preserving random graph through edge swap we undersood that it is different in directed networks. In fact, these random graph models are less effective in directed networks because swaps must preserve both in-degree and out-degree, significantly reducing the number of valid swaps[2]. We tried to overcome this problem by implementing some Programming techniques. First, we limited the number of swap attempts to prevent the function from endlessly retrying unfeasible swaps. Second, we allowed partial swaps, to ensure that the function would still produce a usable randomized network even if the full number of swaps could not be achieved. To guarantee a fair comparison between the original and randomized networks, we reduced the total number of swaps to ensure that only a subset of edges were swapped while maintaining the overall degree distribution.

Despite the fact that we created a random graph, In our analysis of small-world properties for directed networks, we encountered difficulties in computing the average shortest path length due to the lack of strong connectivity in the network. The concept of small-world networks is fundamentally based on the ability of nodes to reach one another efficiently, which is not always valid in directed graphs that are weakly connected or fragmented into multiple components. [2] highlights that for small-world properties to be meaningful, a network must either be strongly connected or contain a giant component that enables effective communication between nodes. However, in our case, the presence of disconnected subgraphs or small strongly connected components made the calculation of APL undefined, preventing us from obtaining a valid small-world coefficient. As a result, small-world properties cannot be directly applied unless the network is either modified (e.g., using its largest weakly connected component) or analyzed in an undirected form.

IV. Why the Normalization Factor is a Bad Idea for Our Analysis?

As we mentioned in the mid-term report, we intended to add a section to introduce a normalization factor. This idea took most of our time for this project (more than 50 percent of the working time). Trying to handle errors in implementation has reached us to the point that we are adding bias to the analysis instead of decreasing it. We took this idea from [3], but since our networks have different characteristics, we adapted the method to fit our analysis better. The formula we used is equation(2), and in this chapter, we are going to discuss why it turned out to be not helpful.

TABLE IV: Comparison with edge swap Model

Dataset	Global Clustering Coefficient	Average Path Length	edge swap Clustering Coefficient	edge swap Average Path Length	Small-World Coefficient (σ)
AstroPh	0.3178	4.1940	0.0102	3.3531	24.9098
MathSciNet	0.4115(appr)	7.3165(appr)	0.0001(appr)	6.0039(appr)	3376.7578
netscience	0.4306	6.0419	0.0247	3.7441	10.8031

$$\hat{k}i = \frac{k_i}{\langle k_n \rangle n \in N_i} \tag{2}$$

Where:

- k_i is the degree (number of co-authors) of node i,
- N_i represents the set of neighbors (co-authors) of node i,
- $\langle k_n \rangle$ is the average degree of the neighbors of node i.

The normalization factor we tried to introduce turned out to be a fundamentally flawed approach, distorted key network properties, added bias, presented numerical instability, and compromised the reliability of our results. While normalization was intended to provide a fairer comparison of node properties by adjusting degree values based on their neighbors, it instead created structural inconsistencies that affected multiple fundamental network metrics, including assortativity, average path length, and global clustering coefficient.

One of the major issues with this approach was the division by zero problem, which arose for nodes with no neighbors. However, this issue was largely irrelevant to our study because three of our networks were fully connected, meaning that every node had neighbors and well-defined degree values. This highlights another fundamental flaw in the normalization process: it introduced unnecessary complexity to a problem that did not exist in our datasets. The normalization factor was attempting to fix an issue, which is nodes with undefined degrees that were not present in our connected networks.

Moreover, since many of the metrics we analyzed, such as assortativity and clustering coefficient, are already normalized and well-defined in their standard forms, applying an additional normalization step was redundant and unnecessary. The core problem was that degree normalization modified the numerical values of node degrees without changing the underlying network topology. However, the metrics we studied are inherently topological, relying on the actual connectivity between nodes rather than on transformed degree values. Furthermore, since degree normalization introduced artificial modifications, it compromised comparability across networks by introducing biases that were not present in the original structures.

Ultimately, the normalization process did not improve the analysis; instead, it introduced errors, made standard calculations unreliable, and obscured the actual structure of the networks.

V. Implementation Details

We implemented the analysis using Python's NetworkX library. We performed the calculation on a M1 Apple Macbook Air with 8 GB of RAM.

All the codes are available here and you can find the datasets in this link.

VI. Conclusion

By calculating homophily and assortativity, we conclude that our networks have different structures, which means the collaboration tendencies among scientists are different.

Analysis of the Netscience network, in which nodes are network science researchers, reveals a hierarchical and mentor-driven collaboration pattern. In these kinds of networks, well-known researchers act as cores, and new researchers have the chance to collaborate with them. Also, this structure means that new researchers must connect with central scholars to become visible and gain access to research knowledge. We conclude that these structures facilitate knowledge propagation and will result in more innovation in this research field. The downside of this type of network could be if elite researchers do not work with each other more often, the chance of creating high-impact results will be reduced. On the other hand, if a new researcher does not establish a direct link with a prominent researcher, it may take longer to integrate into the network and become independent. Suppose we acknowledge this kind of reasoning and expand it to all interdisciplinary fields. In that case, recognized researchers encourage new researchers to produce more knowledge and catch up with more classical fields of research.

By considering the "AstroPh" network, we conclude that Astro Physicists have the smallest world based on the definition of small-world properties. They tend to form tightly-knit groups while maintaining global efficiency through highly connected hubs. Because this network showed a weak positive degree of homophily and a high degree of small world coefficient, we conclude that this network preserves an efficient knowledge-sharing structure by sharing knowledge with peers inside tight hubs while preserving connection with eminent peer researchers. Compared to network science, new researchers in astro physics can

become prominent by building small collaborations instead of depending only on a central mentor.
Since the Author of this work is more familiar with the Ph.D. students' networks, the findings alligns with our personal observation. Collaborations occur primarily between advisors and students in a one-way directional manner. Because of discontinuity, it is impossible to calculate a small world coefficient, but assortivity reveals senior advisors collaborate with new students. This totally confirms the personal observations that in Ph.D students' networks, knowledge disseminates in a top-down model. The discontinuity of the networks also shows the fact that a proportion of students do not easily build collaborations across their research group. Because of the unreliability of the "MathSciNet" network results, we cannot analyze its properties in detail. It is evident that researchers in mathematics form structures more like astro physicists. Based on this fact, these two classical natural sciences form almost the same networks, and knowledge propagation would be almost the same. Therefore, new researchers in this field may not just rely on one mentor to gain knowledge; this can be done by connecting to their peers.
Finally, the work continued for less than two months, and we do not see it as a complete defeat. The most important part for us is that excitement about using a novel method does not always lead to surprising results. By understanding the question, we can choose our methods for analysis more wisely. Finally, if we want to answer the question mentioned in the title, we can say partially yes, but compared to other fields, no. We are confident that it is because classical research fields are older, so their networks (also individuals) are closer.

REFERENCES

1 Newman, M. E. J., "Assortative mixing in networks."
2 ——, "The structure and function of complex networks."
3 Q. Ke, A. B., "normalized impact measure reveals successful periods of scientific discovery across disciplines."