The Structure of Academic Collaboration Networks Within an Elite Liberal Arts College

Anthony Lorson[1] **and Nicholas Dorogy**[1,*]

[1] *Washington and Lee University, 204 W Washington St, Lexington, VA 24450*
[*] *Corresponding author: dorogyn22@mail.wlu.edu*

Compiled July 30, 2021

We explore the formation of collaboration networks among university professors. Two professors are considered to be connected if they have authored a paper together. We have constructed a network from these connections using Washington and Lee University's publications data set. This network will be studied to determine whether or not it adheres to the small world, six degrees of separation hypothesis. We will further show the nature of this network and explore its average degree of separation, clustering, degree distribution, and connectivity.

1. INTRODUCTION

A social network is defined to be a network in which individuals are represented by nodes (or vertices) and edges (or ties) are determined by some relationship between individuals. This connection could be as vague as friendship on a social networking platform, or as straightforward as immediate relation in families.

Since the first social network articles in the 1960's, the study of these networks has become a significant topic of interest. These networks can be used to create targeted advertisements, predict disease percolation, or even track information spread [1–4]. Changes in the structure of these networks can result in significant outcomes. Take for example the COVID-19 virus. The spread of the virus can be reduced to interactions among personal social networks or expanded to account for networks of international travel. Understanding these networks is crucial in hindering transmission and accurately forecasting disease spread.

One of the earliest and most significant analyses on the formation of social networks was carried out by Stanley Milgram [3]. Milgram asked individuals in Nebraska to deliver a letter to a particular Boston stockbroker. However, the individuals could only pass the letter on to someone whom they were acquainted with on a first name basis. Thus, the best method for delivering this letter was to pass the letter on to a person whom they believed to have a better chance of knowing the specified stockbroker. This transfer was likely determined by geography, vocation, or perhaps social influences. By tracking the intermediaries between the stockbroker and original letter sender, Miligram found that, on average, six people separated any two randomly chosen individuals. This conclusion is generally cited as evidence for the "small world hypothesis" in which most members of a network can be connected by a small set of intermediaries, regardless of the size of the network.

Often times, when studying social networks, complications arise in setting parameters for the network. Consider a network of acquaintanceship between citizens of a small town. For those conducting the interview and for each individual questioned, the definition of acquaintanceship may be quite different. Even if guidelines are set to restrict what an acquaintance is, there is a large margin of error in gathering data as the measure of this network is inherently subjective. Thus, for studying social networks, analysis with inscrutable definitions of edges yield more accurate results. Our study takes advantage of such inscrutable rules. There is no argument error when using co-authorship as edges because it is a clearly stated parameter. To further solidify our definition, we restricted our co-authorship to conference papers and articles published in journals. Thus, we have removed a large amount of error in our results. Furthermore, with large unorganized networks, results can be skewed by naming issues. Not only can multiple authors have the same name, but authors can also publish under different names, initials, etc. Fortunately, because of our networks limited size and prior data-organization, authors were assigned unique ID's that removed any sort of error resulting from naming conventions. Thus, our social network provides ample, accurate data to investigate.

While similar studies have been conducted with significantly larger data sets [6], in this paper, we will study a sub-network of personal acquaintanceship that is quite small, and possesses a precise definition of acquaintanceship. We expect to find a relatively large deviation from similar studies that utilized larger data sets because of the irregularities inherent in small data sets that tend to average in larger systems. Moreover, the culture developed in small, elite liberal arts universities will likely cause differences between similar results from larger networks and our own small, localized network.

2. ACADEMIC COLLABORATION NETWORKS

We study networks of professors in which two professors are considered connected if they have coauthored a paper together. This is a fairly reasonable definition of connectivity: most academics who have collaborated on a paper together know one another quite well. This is also a moderately stringent definition considering the number of academics who may be acquainted yet have never co-authored a scientific journal article before. However, we prefer to have a rigorous definition over a looser one as long as this rigor can be applied consistently throughout the network.

We have constructed collaboration graphs for professors at Washington and Lee University who have self-reported at least one publication in an academic journal. The database of professors is maintained and updated annually by Washington and Lee's University library.

This idea of having a scientific collaboration network based on publication records is not a new one. In the field of mathematics, the idea of the **Erdos number** is infamous. Paul Erdos was the most prolific mathematician of all time, having published at least 1400 papers in his lifetime. The concept of the Erdos number is a measurement of how close a mathematician, in bibliographic terms, is to Paul Erdos. An Erdos number of 1 means that the mathematician in question has published a paper with Erdos. An Erdos number of 2 means the mathematician has published a paper with another mathematician who has coauthored at least one paper with Erdos. And so on. An exhaustive list exists of all mathematicians that have Erdos numbers of 1 or 2.

There are many additional things one can study from a scientific collaboration network like the one we have created. we will be discussing the total number of authors, average degrees of separation between agents, **Giant Component**, density, and the clustering coefficient of our network.

3. METHODS

We have chosen to use multiple methods of analysis to determine our results. While some characteristics of the network were individually programmed, many other measures were simply pre-defined functions that have either been modified or directly used from existing libraries in NetworkX. Cytoscape and Gvedit were also used in the construction of graphs and to perform analyses not included in the NetworkX library.

First and foremost, the network data contained in our excel sheet was passed to our Python-based script using Openpyxl. The nodes, identified by unique IDs were formulated into dictionaries that contained information on each author. Edges were constructed by analyzing the list of works to determine whether or not more than one ID was present in the list of WL author IDs. For each ID present, an undirected edge was created between these ID numbers using the edge addition NetworkX function. it Since it is an undirected edge if two authors have already worked together, they will not be counted twice. Finally, we used a variety of different libraries and applications to create graphs. These include, NetworkX, Matplotlib, Cytoscape, and Gvedit.

In order to calculate the number of nodes, unconnected nodes, edges per author, and total edges, we wrote simple statements or used simple Python operators. However, to determine the clustering coefficient and diameter we used NetworkX functions. To find degree distribution, degrees of separation, and network density, we used Cytoscape for its rich variety of built in tools.

4. RESULTS

Here we offer a summary of the results of our network analysis. The following sections examine different aspects of the network in detail.

A. Authors

When calculating the number of authors, it is important to consider complications with counting the true number of individuals. First, authors can change their signature or identification. Authors may switch from using one initial, to middle and first initials, or perhaps first initial and surname. This could result in creating two new nodes for one author, inaccurately creating a larger network. Moreover, if an author is to change their name, they may be recounted as a new author. Once again, this erroneously inflates the size of the network. Second, authors may have the same name. Incidentally, this would shrink the size of the network.

Because individual authors were assigned a unique ID in our data, this network is not burdened with naming issues. In addition, by using IDs, any misspellings or unusual data entries will not create complications with node uniqueness because each unique ID spawns its own node.

We found that author ID 246, Associate Professor of Biology, G.B. Whitworth, possessed the highest degree of co-authorship by having collaborated with nine other Washington and Lee professors. This result is not particularly surprising considering the highly co-operative nature generally found in the Biological Sciences.

Author ID 433, Henry S. Fox Professor of English, Lesley Wheeler, has published the most papers at 222 individual works. Once again, this result is not surprising when considering the high-volume output generally found in poetry.

B. Average Degrees of Separation

Degrees of separation, or the typical distance between a pair of nodes, is an important characteristic of network connectivity. In order to determine the average separation in a network, the network must be reduced to include only sub networks of connected nodes. Then, by selecting any two random nodes in the sub network, the degree of separation is equal to the number of intermediaries that must be passed through to connect the two nodes if they are not directly related. Performing this calculation on all connected nodes in the network returns the average degree of separation.

For Washington and Lee professors with at least one coauthored paper, we found that there are about 3.7 degrees of separation. This number is quite surprising considering the fact that social networks generally conform to Milgram's six degrees of separation [3]. However, this unusually low degree of separation can be slightly misleading. Nearly 66% of this network is not connected to any other individual, which means they are not considered when determining degrees separation. Thus, there are 3.7 degrees of separation on average *only for the sub-network that requires at least one fellow Washington and Lee coauthor*. For 66% of the network there does not exist any degree of separation. It appears that those who work together are highly collaborative and form tight, intimate cliques.

We also analyzed the maximum and minimum separation between any two connected nodes. The maximum distance between nodes is often referred to as the diameter, while the minimum is commonly referred to as the radius. The diameter of our network, or the maximum distance *ever* travelled between nodes, is 33. The radius, the shortest distance *ever* travelled between nodes is 1.

C. Giant Component

In all social networks there is a possibility of a **percolation transition**[7]. That's to say, in networks with very small numbers of connections between agents, all agents belong only to small islands of collaboration or communication. In Newman's 2001 paper on scientific collaboration networks, he found that the networks he had created from a much larger set of databases had a **giant component**- a large group of individuals who are all connected to one another by paths of intermediary acquaintances.

We found that in our network, the largest group of such collaborators comprised of approximately 10% of the population. This percentage, which is a relatively small fraction of the population, can be interpreted several ways. One of the most viable theories is that because we did not limit our study to only publications in the sciences, these small islands of collaboration are reflective of the various subjects included in our database. It's clear that a professor publishing poetry will be on the humanities island whereas a professor publishing work on nanoscience will be on a completely separate island that is quite distant from that of the poet.

D. Network Density

Network density is a measure employed to show how densely a network is populated with edges. Density, a value returned between 0 and 1, gives a good indication into the amount of collaboration that occurs in social networks like our own. For networks with a high network density (a large number of edges), there is a high degree of co-authorship.

Further confirming speculation that this network is not particularly collaborative, we discovered our density to be .002. Simply, there are far fewer edges in comparison to the number of nodes. Although this measure says nothing of Washington and Lee professors who publish with faculty from other institutions, it does suggest that Washington and Lee professors do not work among themselves often.

E. Clustering

In 1998, Watts and Storogatz[8] pointed to another important quality of social networks which is often overlooked in many network models. Real world networks are **clustered**, meaning that within the network, they have communities in which a higher than average number of people know one another. In academic collaboration, this could be a university department or lab group. Storogatz and Watts defined a clustering coefficient, C to describe this phenomenon. C is defined as the average fraction of pairs of a person's collaborators who have also collaborated with one another.

Our network has a clustering coefficient of $C = 0.108$. This is a much lower number than we had predicted. Although it is still within the range of what Newman[5] asserts is typical of a real world social network, it is on the lower end of that range. We expected C to be much higher considering the level of connectivity that is assumed to be commonplace at a small liberal arts college.

5. CONCLUSIONS

We have thoroughly examined collaboration networks of professors using Washington and Lee's publication database. We found a number of irregular, interesting components in this network. It appears that professors rarely work with other faculty members from our own institution. Nearly 66% of all published faculty members have never worked with a colleague at this institution. Yet, in similar networks, only around 10 to 20 percent of the network is entirely disconnected [6]. Moreover, the network's clustering coefficient is on the extreme low end of expected values which generally ranges from 10 to 60 percent [5]. Our extremely low network density value, combined with a low giant component further solidify the conclusion that there are very few papers published by faculty members that include colleagues from this school. One possible explanation could be Washington and Lee's reputation in the social studies. These disciplines are not known for their large networks of collaboration, but instead contribute a great number of individual works. Meanwhile, for the sciences, papers with several hundred authors are not uncommon due to the inherently interconnected nature of their research. Thus, Washington and Lee may exhibit the behavior of a network made up predominately of social studies disciplines.

Continuing this research to track co-authorship with members outside of our community could yield interesting results. It is possible that the "Washington and Lee bubble," the idea that faculty and students at Washington and Lee are disconnected from the outside world, may also be influencing internal relationships within the community.

Further applications of this work could include analyzing the network by subject. We expect clear differences to arise between the structure of networks in disciplines that publish in much different fashions. Moreover, extending this research to similar sized universities would be invaluable for constructing an accurate picture of the formation of such social networks. If averaged over a large number of institutions, it is possible that our irregularly unconnected network is better analyzed as an abnormality.

6. ACKNOWLEDGEMENTS

We would like to first acknowledge and thank Dr. Irina Mazilu for her guidance throughout this project. We would also like to thank Jason Mickel who graciously allowed us to use his database of collaborators at Washington and Lee. Lastly, we would like to thank Rob Hageboeck who provided much of the support we needed in terms of creating a functioning network code in Python.

REFERENCES

1. Eytan Bakshy, Itamar Rosenn, Cameron Marlow, and Lada Adamic. The role of social networks in information diffusion. In *Proceedings of the 21st international conference on World Wide Web*, pages 519–528, 2012.
2. Alden S Klovdahl, John J Potterat, Donald E Woodhouse, John B Muth, Stephen Q Muth, and William W Darrow. Social networks and infectious disease: The colorado springs study. *Social science & medicine*, 38(1):79–88, 1994.
3. Stanley Milgram. The small world problem. *Psychology today*, 2(1):60–67, 1967.
4. Alan Murray, Joshua Abram, Rodney Hook, and Balaji Devarajan. Systems and methods for targeting online advertisements using data derived from social networks, February 17 2011. US Patent App. 12/191,412.
5. M. Newman. *Networks*. OUP Oxford, 2018.
6. Mark EJ Newman. The structure of scientific collaboration networks. *Proceedings of the national academy of sciences*, 98(2):404–409, 2001.
7. Dietrich Stauffer and Ammon Aharony. *Introduction to percolation theory*. CRC press, 2018.
8. Duncan J Watts and Steven H Strogatz. Collective dynamics of ‘small-world’networks. *nature*, 393(6684):440, 1998.

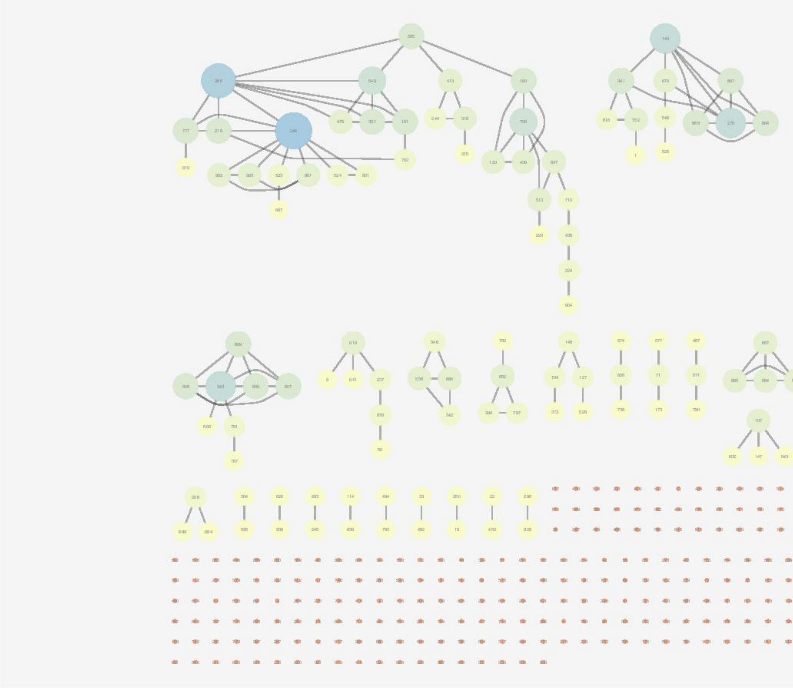

Fig. 1. Tree Graph of Our Scientific Collaboration Network