Mapping Collaborations across Fields

by Eugene Hwang, Amit Talapatra, Marissa Wiener, Wendy Zhang, and Wei Zheng

Masters of Science in Business Analytics, May 2017,
George Washington University School of Business

A Thesis submitted to

The Faculty of
The School of Business
of The George Washington University
in partial fulfillment of the requirements
for the degree of Master of Science
in Business Analytics

May 21, 2017

Thesis directed by

Dr. Shivraj Kanungo

Chair of the Department of Decision Sciences
Faculty Director of Decision Sciences
Associate Professor of Decision Sciences & of Info Systems & Tech Management

Acknowledgements

We would like to thank Dr. Anna Calcagno, Dr. Andrew Miklos, and Claire Schulkey of the National Institute of General Medical Sciences (NIGMS) for providing the topic and for mentoring our team throughout this project. We would also like to thank Dr. Shivraj Kanungo of the George Washington University School of Business for guiding our team on the network analytics techniques used in this project.

Abstract

Mapping Collaborations across Fields

This analysis explored collaboration among researchers supported by the National Institute of General Medical Sciences (NIGMS). The project was structured in the form of four study questions looking at collaboration behavior with respect to: 1. NIGMS support, 2. grant type, 3. variety in funding sources, and 4. scientific discipline. The team found that homophily was apparent in the network when grouping by each of these factors. Additional findings included the following: clustering was higher within the network of NIGMS-supported researchers compared to other NIH-supported researchers; cooperative grant types effectively foster collaboration; the 'R01' grant type, despite being the most common, result in relatively low levels of collaboration; variety in funding sources correlates with increased collaboration; disciplines with more interdisciplinary collaboration (such as chemistry and chemical engineering) have higher average degrees than other disciplines; and finally, that the discipline of 'medicine' is a particularly tightly-knit network with a relatively high amount of internal collaboration with the network studies.

Executive Summary

This analysis was produced for the National Institute of General Medical Sciences (NIGMS) in order to provide a better understanding of collaborative behavior within the NIGMS network of researchers. The project was framed around the following four study questions provided by NIGMS:

1. Do NIGMS-supported researchers collaborate more often with other NIGMS-supported researchers?
2. Does the type of grant awarded influence collaboration behavior?
3. Does variety in funding sources influence collaboration behavior?
4. Does the scientific discipline influence collaboration behavior?

Each of these questions was answered using a network analysis approach where researchers were identified as nodes in a co-authorship network and common publications were used as edges to connect researchers. For each of the study questions, nodes were labelled based on the feature being measured for its influence on collaboration behavior (i.e. NIGMS support, type of grant, variety in funding sources, or scientific discipline). A series of network metrics were calculated to evaluate homophily, clustering, and the completeness of networks with respect to these categories. The networks produced for each study question are shown in the following figure.

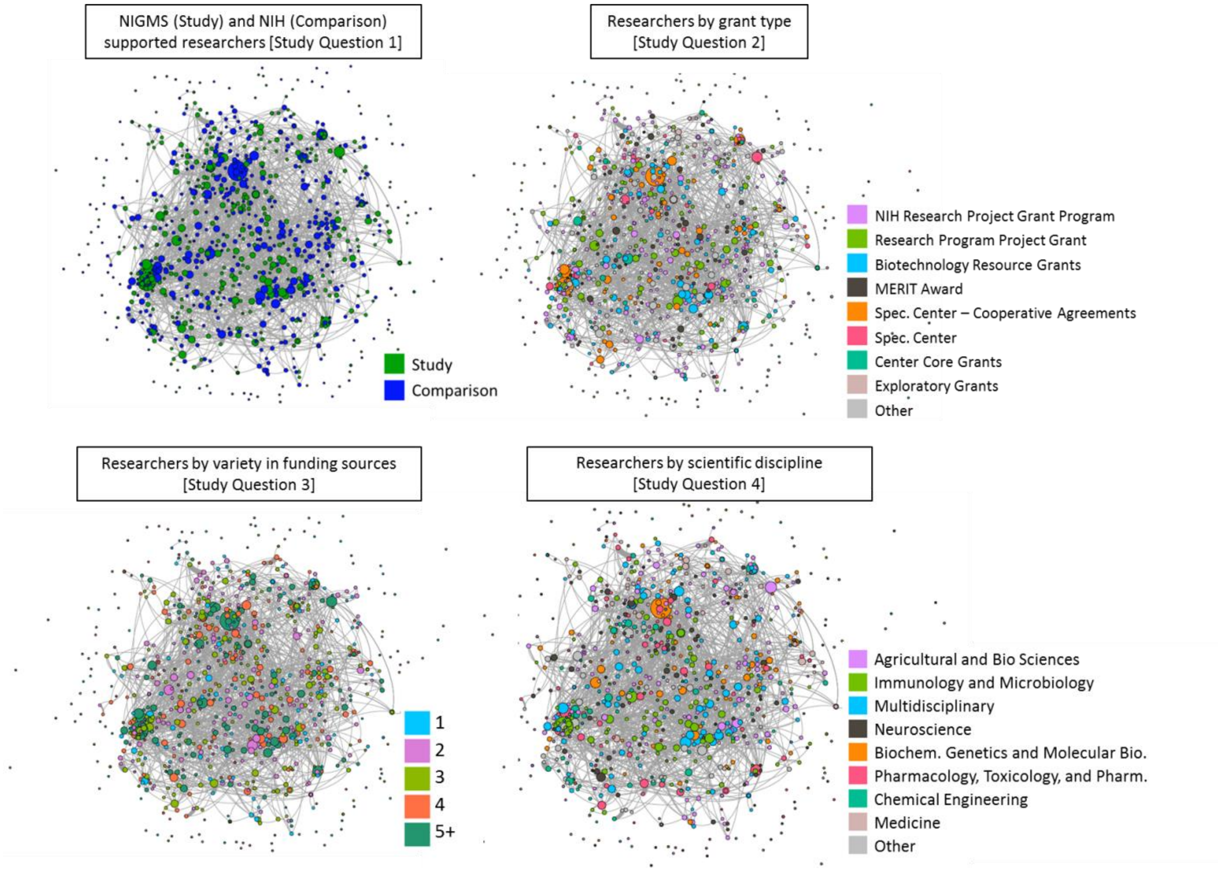

NIGMS (Study) and NIH (Comparison) supported researchers [Study Question 1]
Study
Comparison
Researchers by grant type [Study Question 2]
NIH Research Project Grant Program
Research Program Project Grant
Biotechnology Resource Grants
MERIT Award
Spec. Center – Cooperative Agreements
Spec. Center
Center Core Grants
Exploratory Grants
Other
Researchers by variety in funding sources [Study Question 3]
1
2
3
4
5+
Researchers by scientific discipline [Study Question 4]
Agricultural and Bio Sciences
Immunology and Microbiology
Multidisciplinary
Neuroscience
Biochem. Genetics and Molecular Bio.
Pharmacology, Toxicology, and Pharm.
Chemical Engineering
Medicine
Other

Figure 1. Network graphs created for each of the four study questions.

This analysis determined that each of the network attributes considered for the study questions had an impact on collaboration behavior. For study question 1, it was determined that there is homophily among the study and comparison groups, and that clustering is higher within the study group, indicating that NIGMS-supported researchers tend to collaborate more with other NIGMS-supported researchers. For study question 2, the analysis found that grant type is an indicator of homophily, and that cooperative agreements demonstrate high levels of collaboration. Additionally, this analysis found that 'R01' grants, one of the most common grant types, had low levels of collaboration. For study question 3, the analysis found that variety in funding sources is an indicator of

homophily, and that there tends to be more collaboration among researchers with greater variety in funding sources. The team also identified that the tier of the researchers with low variety in funding sources had relatively low levels of collaboration, despite having a similar number of researchers as the tiers with higher levels of variety in funding sources. Finally, for study question 4, the analysis found that scientific discipline was an indicator of homophily. The team also identified that disciplines with more interdisciplinary collaboration (such as chemistry and chemical engineering) had higher average degree counts than other disciplines. It was also found that the discipline of Medicine is a particularly tightly-knit network with a relatively high amount of internal collaborations compared to the other disciplines studied.

Table of Contents

List of Figures

Glossary of Terms

Principal Investigator (PI): Individuals who contribute to the scientific development of a project.

Activity Code: A 3-character code used to identify a specific category of extramural research activity, applied to financial assistance mechanisms

Grant: Financial assistance mechanism providing money to an eligible entity to carry out an approved project or activity.

Scientific Discipline: A particular branch of scientific knowledge.

Node: A visual representation of a researcher.

Edge: A visual representation of a publication between two researchers.

Homophily: "Birds of a feather stick together". Individuals tend to congregate with other like-minded individuals to form tightly knit clusters.

Assortativity: A measure of homophily, ranging from -1 to 1. A positive assortativity indicates that each group tends to collaborate within themselves.

Degree Centrality: A measure of the number of direct connections (degrees) one node has to other nodes. The number of connections is a key measure of importance or influence within the network.

Dyad: Consists of two connected nodes

Triad: Consists of three connected nodes.

Transitivity: A measure of social process such that friends of friends will become friends. A high transitivity measure indicates a tightly knit collaborative environment (clique), while a low transitivity measure indicates that neighboring researchers do not collaborate with each other (star).

Chapter 1: Introduction

The concept of this project was provided by the National Institute of General Medical Sciences (NIGMS). The Institute developed this project as a way to better understand collaborative behavior within its network of researchers in the United States. Understanding the relationship between collaboration and other attributes of researchers could lead NIGMS to identify opportunities for new grants that would accelerate scientific progress. NIGMS presented the project scope as four study questions:

1. Do NIGMS-supported researchers collaborate more often with other NIGMS-supported researchers?
2. Does the type of grant awarded influence collaboration behavior?
3. Does variety in funding sources influence collaboration behavior?
4. Does the scientific discipline influence collaboration behavior?

For the purposes of this project, collaboration was identified as co-authorship between researchers on journal publications. In the following network analysis, researchers are represented by nodes while common publications between researchers are represented by edges.

For the first study question, nodes in the network were assigned as either the ‘study group,’ meaning fully supported by NIGMS from 2001-2015, or the ‘comparison group,’ meaning fully supported by other NIH sources from 2001-2015. The purpose of this study question was to investigate collaborative behavior within the network based on these two groups, as shown in the following figure.

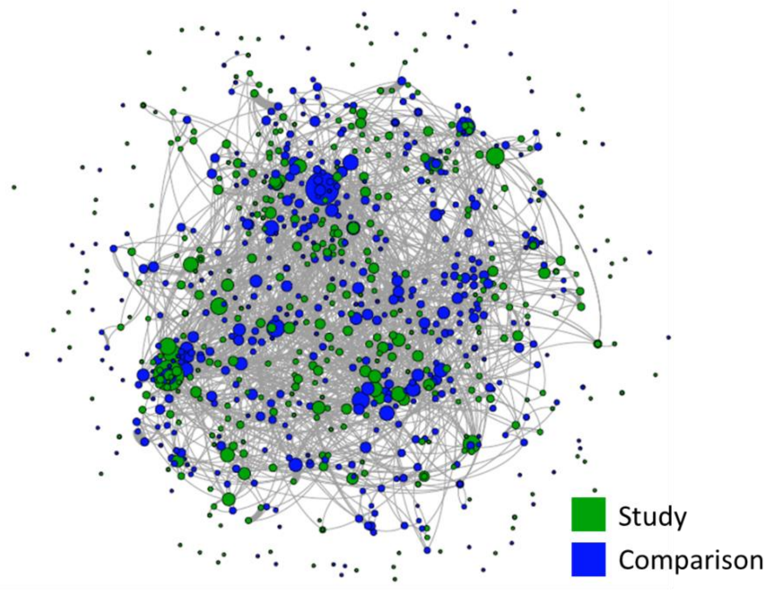


Figure 2. Degree centrality graph of the study and comparison groups.

The second study question explored collaboration based on the type of grant most influential to each researcher, which is reflected in the figure below.

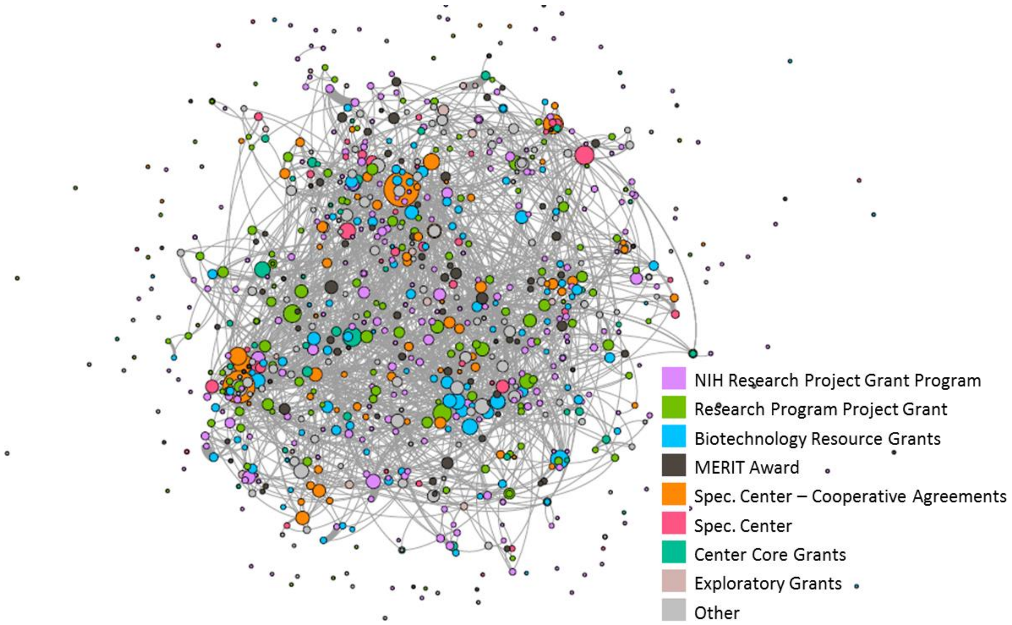


Figure 3. Degree centrality graph based on the most common type of grant associated with each researcher.

The third study question looked at the impact of variety in funding sources on collaboration behavior. In this project, variety in funding sources was defined as the number of unique funding sources attached to each individual researcher. The network graph based on this attribute is shown below.

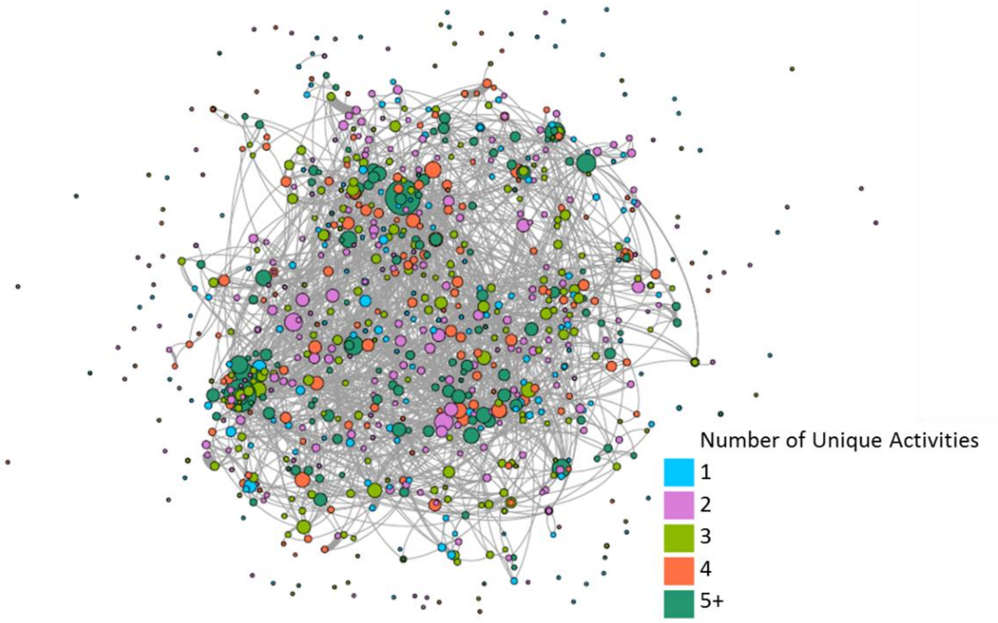


Figure 4. Degree centrality graph based on the number of unique funding sources (activities) associated with each researcher.

Finally, the fourth study question explored the impact of scientific discipline on collaboration behavior. Researchers were labelled by scientific discipline based on the discipline that most accurately represented the journals that they most often published in, as shown in the following figure.

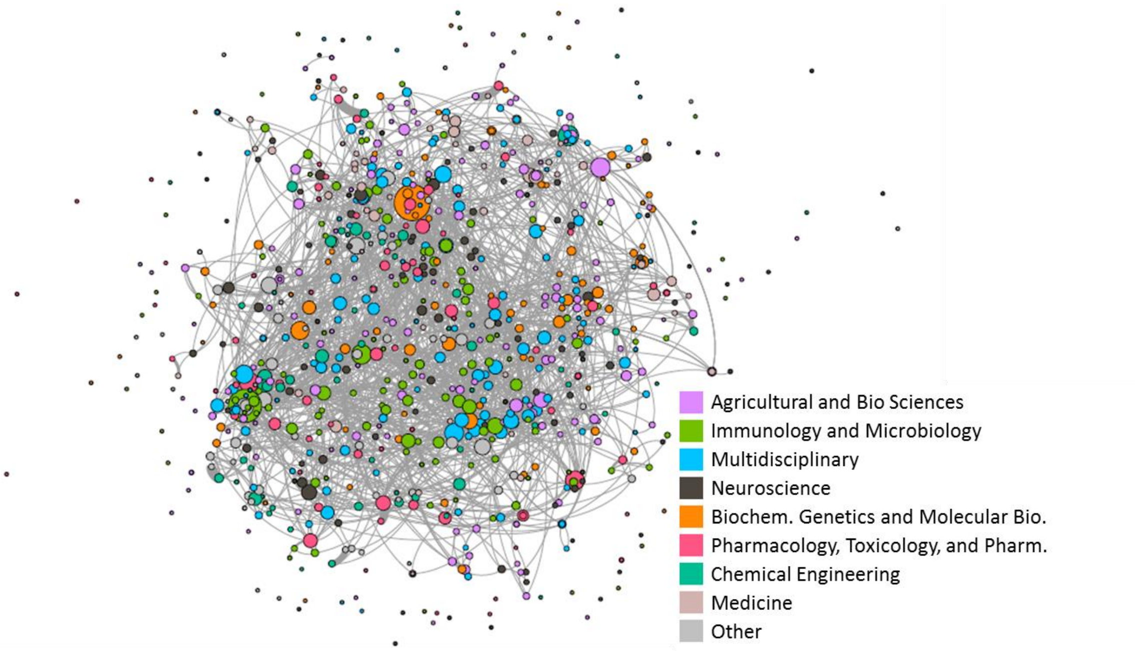

Agricultural and Bio Sciences
Immunology and Microbiology
Multidisciplinary
Neuroscience
Biochem. Genetics and Molecular Bio.
Pharmacology, Toxicology, and Pharm.
Chemical Engineering
Medicine
Other

Figure 5. Degree centrality graph based on scientific discipline.

Each of the four network graphs presented above were investigated to determine how these groupings related to collaboration behavior. Several network analysis techniques were used in this project to estimate homophily, clustering, and the completeness of graphs with respect to the groupings identified by the study questions.

Chapter 2: Background

National Institute of General Medical Sciences (NIGMS)

The National Institutes of Health (NIH), an agency of the United States Department of Health and Human Services, is primarily responsible for biomedical and health-related research. The NIH comprises of 27 Institutes and Centers that conduct research in different disciplines of biomedical science, one of which is the National Institute of General Medical Sciences (NIGMS). NIGMS supports basic research that increases understanding of biological processes and lays the foundation for advances in disease diagnosis, treatment and prevention.

NIGMS funds a wide network of researchers across the US. Since collaboration is an important aspect of scientific research, NIGMS is interested in understanding how researchers interact with one another in order to identify gaps in collaborative networks. Gaps in collaborative networks represent opportunities for scientific progress via interdisciplinary work. Additionally, analyzing the network of researchers at NIH can reveal non-grantee researchers performing work relevant to NIGMS.

Availability of Data

Several grant databases are available for the general public. First, the Federal Reporter is a searchable database of scientific awards from federal agencies. It provides empirical data for science policy and describes federal science research investments. Federal projects can be searched using fiscal year, award’s project leader, or by a text search of a project's title, terms, or abstracts. Second, the Project Research Portfolio Online Reporting Tools (RePORT) provides access to reports, data, and analyses of NIH research activities, including information on NIH expenditures and the results of NIH-supported research. Similarly, Project RePORT queries can be constructed using the project number, NIH spending category, and Principal Investigator's name, among other fields. Third, ExPORTER provides bulk download of RePORTER data files that include information on research projects funded by the NIH as well as publications, patents, and clinical studies citing support from these projects.

In addition to these public grants databases, there are two online subscription-based scientific citation indexing services. First, Web of Science, maintained by Clarivate Analytics, provides access to multiple databases that references cross-disciplinary research, allowing for in-depth exploration of specialized sub-fields within scientific discipline. Second, Scopus, owned by Elsevier, is a bibliographic database containing abstracts and citations for academic journal articles.

Exposition of Relevant Literature

Prior studies have identified many tools to assess collaboration trends and to identify leading researchers. These tools include social network analysis, co-authorship networks (Fonseca et al., 2016), paper citation networks (Onel, Zeid & Kamarthi, 2011), and scientific collaboration networks (Newman, 2000).

In this study and related research, collaborative behavior in the scientific community was the main focus since, "researchers are no longer independent players, but members of teams that bring together complementary skills and multidisciplinary approaches around common goals" (Fonseca et al., 2016). The project focused on researchers continually supported by NIGMS as well as researchers continually supported by NIH, while other studies focused on researcher who conduct research on nanotechnology (Onel, Zeid & Kamarthi, 2011) or researchers in the database community (Elmacioglu & Lee, 2005). Using several network statistics such as degree centrality, homophily, assortativity, and transitivity and interpreting their significance to help explain collaborative behavior, several studies concluded that "scholars collaborate more often than before" (Elmacioglu & Lee, 2005). The review of this literature, and other articles of similar nature, provided insight for common methods and approaches of similar investigation and a general framework for the structure of analysis.

Chapter 3: Description of Work Undertaken

Data Wrangling

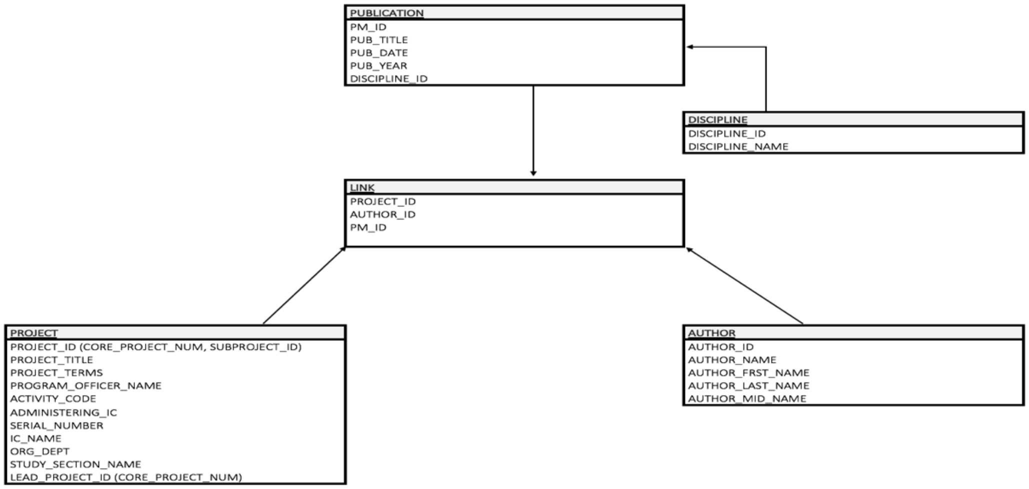
To feed data into network analysis packages, such as igraph and Gephi, data must be transformed into specific formats. A network consists of nodes and edges. Also, additional attributes must be added to those nodes and edges to allow for more extensive analyses. Unlike other data science algorithms which require data in a tidy, tabular format, network study requires one node file and one edge file. In the node file, a unique id column should be created for each node, and other attribute columns can be added to describe the nodes. Alternatively, the edge file should contain at least two columns that contain node ids from the node file. Each row of these two columns represents an edge connecting two nodes. A special weight column may be added to quantify significance of each edge. Like a node file, an edge file can contain multiple attribute columns to enhance the network study. In this collaboration network, authors are designated as nodes, while publications on which they co-authored are edges. To proceed with the network analysis, an author file and a publication file must first be created. Additionally, attributes for both nodes and edges needed to be extracted from multiple sources.

Data preparation was time consuming and challenging, as expected. Just like in any data science projects, much of time and effort was spent on gathering and cleaning data. There were multiple factors complicating this process. First, there was not a central location to pull data. The project required raw data from NIH ExPORTER and Elsevier Scopus. Downloading relevant data through their web interfaces was labor intensive. Moreover, the annual files downloaded from NIH ExPORTER needed to be stacked.

However, new columns were gradually added over years. As a result, columns were not aligned correctly. Furthermore, not only did files need to be vertically concatenated, but also they needed to be horizontally joined. There was no such direct link between NIH researchers and publications. The only way to establish connections was to merge NIH grant data with publication by multiple joins. However, the csv format of downloaded NIH ExPORTER files was not suitable for joining operation. More importantly, researcher information was mired in the grant dataset. Therefore, a fair amount of effort was required to extract individual researchers and their attributes.

Creation of database

To tackle the challenges listed above, choosing the right set of tools and software was critical. The most appropriate way to handle the data provided was a database. Despite the time investment upfront, downstream analysis tasks would greatly benefit from its query capabilities. SQLite was chosen to be the database of this project. Unlike Postgres and MYSQL, the installation is painless. It is very fast because of its C implementation and because it is lightweight. A star schema was designed to provide a logical structure for the project data. In the schema, one project fact table was linked to multiple dimension tables by id columns. Those dimension tables provided an effective way to slice and dice data.



Notes:
If there are no projects associated to the publication or author, then the PROJECT_ID would be -999.

Figure 6. Star Schema of Database

For data retrieval, Python was employed to automate downloading processes for its powerful libraries, database, and web APIs. A python program was designed to process a list of NIH Exporter web requests to download zipped csv files for NIH annual grant, publication, and linkage datasets. Those files were then read into Pandas data frames for further processing. The citation database Scopus provided a web API. PMIDs, which uniquely identified each publication, were submitted with the Scopus API key to its server to retrieve discipline information. The returned results were temporarily saved in a python dictionary object before being saved to csv files.

Next, to vertically stack NIH files, three SQLite tables were used to efficiently align columns in each csv files. To make sure that all columns were present before each file was attached to the tables, the files were appended in chronologically reverse order. The 2015 file was written to the database first to create the tables so that all columns could be created. Indices were generated on key columns to increase query efficiency.

After this step, the relevant columns for those datasets could be retrieved from the database for further cleaning and transformation.

Following the last step, the author column in the publication dataset required parsing to retrieve all unique authors to create the node file for this collaboration network. Additional attribute columns were added to the file to represent qualitative information related to those authors. For the edge file, the program looped over each publication to create a row containing a pair of co-authors. Similarly, additional publication attributes, such scientific discipline, were added.

Study Group and Comparison Group

Before creating the node file and the edge files with targeted authors, two interested groups of authors had to be identified to answer study questions. The first group was Study Group, which would contain researchers who were continually supported by NIGMS between 2001 and 2015. The other group was Comparison Group. This group would consist of researchers who were continually supported by other NIH agencies in the same period. But they also needed to be supported by NIGMS in one of those fifteen years. Additionally, for both group, the researchers had to receive more than $50, 000 in a year to be classified as supported for the year.

To identify researchers in both groups, researcher information had to be extracted from the grant datasets. A grant could be associated with multiple researchers who could be uniquely identified by their PI IDs. As a result, in both researcher name and PI ID columns, researchers and their PI IDs for a grant were recorded as comma separated strings. Therefore, to extract the name and PI ID of each researcher and his or her related grant attributes, a Python program was created. For each row of the grant dataset, the program split the researcher name and PI ID strings by comma, turning them into two arrays. The lengths of these two arrays were checked to make sure that one PI ID corresponded to one researcher. Then items in those two arrays were processed to remove white spaces and unnecessary characters. Finally, the names and PI IDs of researchers, along with the grant attributes, were recorded into expanding arrays. At last, the arrays were organized into a Pandas data frame and written into a researcher table in the database.

As mentioned above, a researcher had to receive over $50, 000 to be considered supported in a year. So, the program calculated yearly sums of the total costs of grants for each author. With the information on hand, group memberships of researchers were determined based on the business rules aforementioned.

Now researchers were ready to be joined with the publication dataset using grant ids. Although a researcher was associated with a grant and the grant which, in turn, was associated with a publication, it didn't mean the researcher was a co-author of the publication. Like the researcher column in the grant datasets, the names of co-authors for a publication were concatenated into a string in the Author column of the publication dataset. To filter out those non-author researchers, the names of researchers were checked if they were a substring of the author string in the Author column. After relevant publication subset were obtained, the author column was parsed to retrieve all co-authors in both groups. Authors who were not in both groups were categorized into a group called "others." The "others" group was not included in the analysis because of the scope of the project and computation capacity required to analyze a large network. Also, they were excluded since these authors did not meet the requirements mentioned previously and were often lab assistants, technicians, and graduate students. However, the structure and design of the database allows for the other researchers to be included in future studies.

igraph Functions

A prudent programming approach was developed to analyze the network. Although the four study questions required the network to be segmented by different node attributes, they were similar in a way that they all aimed to study the sub graphs of the network. The goal of the analysis R program was to make functions as general as possible so that code reusability could be maximized.

Among network statistics, within and outside group collaboration percentages, transitivity, complete triad count, strength, and assortativity were identified as most helpful in answering those four questions. Furthermore, those statistics were calculated over time to identify temporal trends. The annual statistics were not very helpful in answering the study questions, however, because of data quality issues in years prior to 2010. Block-modeling and QAP were also considered. Unfortunately, the answers to the study questions required determined segmentations. For this reason, block-modeling was deemed to be irrelevant to this project. If “other” researchers were included in the study, this method could be used for data reduction. QAP could potentially be used to develop a predicative model that would be appropriate and applicable to scope beyond this project. Additionally, degree distributions were calculated as a data quality measure to ensure that degree distributions of all sub graph had approximately power law shapes. After functions were created for each statistical measure, they were grouped in a wrapper function called “analyze_graph”, which took a graph object in for analysis then return all above statistics. Then, a function was created to produce an igraph object from the node file and the edge file. Node and edge attributes were assigned to this object to enable network analyses. Using those attributes, the network was dynamically segmented into sub graphs, as required by each study question. Then those sub graph were passed into the analysis function to produce network statistics for further studies.

Similarly, multiple plotting functions were created to produce charts, such as heat maps and distribution plots, to assist analyses. But the generalization of plotting functions posted greater challenges than statistics functions. Every chart type for each study question may require different colors and axis limits. For this reason, a plotting function for each question was created to enable customizations of plots.

Gephi Functions

Gephi is the leading visualization and exploration software for all types of network from social networks to biological networks. Since Gephi is an open-source software, there are continual enhancements to the tool based on user feedback, resulting in a more robust software as well as an increasing user community.

There were several benefits to using Gephi for this scientific collaboration network analysis. Gephi was designed with users and their needs in mind, and it doesn’t take long to navigate through its various functionalities. Gephi also made it easy to switch to and from different layouts, such as the Fruchterman-Reingold layout to the OpenOrd layout. This was useful in quickly detecting differences in layouts, and creating exceptional visualizations of the overall network. Gephi is a powerful instrument for exploratory data analysis, enabling users to understand more about a given network. There is also functionality to compute various network statistics such as degree centrality and clustering coefficient, which made it easy to understand and visualize the statistics and the meaning behind them.

However, there were certain aspects of Gephi that made analysis difficult. First, different versions of Gephi load data very differently. In an earlier version of Gephi, the weight of parallel edges is counted separately, assigning each a weight of one, whereas in later versions of Gephi, weights of parallel edges are summed. Second, there are differences in naming conventions in Gephi and igraph, which made it difficult to compare results. For instance, modularity in Gephi as compared to assortativity in igraph, and clustering coefficient in Gephi as compared to transitivity in igraph. It was clearer to understand and trace the metrics produced from igraph, and the package was more flexible to use. Gephi did not have the ability to attain specific statistics by activity codes or discipline, which made detailed analysis difficult. For these reasons, the calculations used in this study were calculated using igraph.

Feature Engineering

It was necessary to adjust some of the features in the data to answer each of the study questions. Much of the joining of datasets, along with the identification of researchers collaborating was generated using PI name. The initial data file downloaded from NIH ExPORTER had each author’s full name as a single string, with last, first, and middle name separated by comma. To ensure consistent formatting and avoid potential duplicative records, full names were split into three separate fields, so that the combination of all three columns could represent a unique author.

Next, grant type needed to be assigned to the node/author level. Assigning this attribute to the node level allowed for certain network statistics to be computed, which were essential to analyze the nature of collaboration behavior and answer the study questions. Since grant type was at the project level, researchers and publications could be associated with multiple grant types. After gathering the unique grant types for each researcher and publication, the grant type with the most funding was used to determine the classification for each. There are inherent differences in the distributions of connections for the different grant types by definition, and these differences are more often associated with the funding amount associated with each type of grant. For this reason, along with the fact that funding amount was used to assign grant type classification to researchers and publications, it was useful to calculate the normalized weights of network edges to account for these interactive factors. Ultimately, this weighting would be used to produce the strength metric to calculate the weighted average degree. To calculate weighting of edges, it is essential to first look at the distribution of grant types over all publications, using the assignment method discussed above. The distribution of publications by grant type was calculated for publications in the entire network population, including those brought in from researchers in the “other” category, resulting in about 50,000 publications. From here, the reciprocal of the distribution was calculated and normalized between the values of zero and one to obtain a factor which could be applied to each publication/edge. Now, one collaboration would count as that value other than 1 and each grant type is weighted taking into account the natural distribution of connections.

Study question 3 required interpretation to define what variety of funding sources meant, as it was not explicitly provided in the data. To answer this question, funding varieties were interpreted as the number of unique grant types for each researcher. Taking into consideration that researchers with more projects would have a greater variety in funding sources, adjustments were made to ensure that all researchers could be compared despite the number of projects they had. A rate of number of unique grant types over total number of projects was created to get a grant type per project value. The results were then grouped into quantiles so that groupings could be discrete. The transformed data was then used to discover patterns of collaboration associated each "bucket" of variety in funding group.

For question 4, discipline data was not included in the initial data file. This data was available at the publication level, so it needed to be gathered and assigned to the authors at the node level. Similarly to grant type, there is a tendency of certain fields to inherently have more publications than others, so it was necessary to normalize publication weights. As discussed previously, discipline information for publications was obtained from the Scopus API. Multiple disciplines were assigned to publications, and in turn, multiple disciplines were associated with authors in the network. First, a distribution of publications by discipline and their reciprocals were calculated and normalized between the values of zero and one. To calculate the discipline weights associated with each edge, these values were applied as a factor to each of the publications associated with that particular discipline. To calculate the weight for an edge with multiple disciplines, the average weight was assigned. Finally, to assign disciplines to the node level, PI information was attached to publications before averaging discipline weight, and then publication information was removed. Grouping by PI, total weighted collaborations were summarized for each of the disciplines. Researchers were then assigned the discipline associated with the maximum value of weighted collaborations.

Chapter 4: Analysis and Results

Study Question 1

The first study question posed was whether NIGMS supported researchers tend to collaborate more often with other NIGMS supported researchers. As mentioned previously, data was separated into comparison and study groups, which by definition are the two groups of interest for this study question. The study group contains researchers continually supported by NIGMS from 2001-2015 and the comparison group contains researchers continually supported by other NIH organizations from 2001-2015. When grouping by study and comparison group, the assortativity was positive at 0.21, indicating that there is homophily by study or comparison group. When breaking down the distributions of collaborations on an overall basis, 37% of collaboration was across group. Therefore, the majority of collaborations were within either the comparison group or study groups. The exhibit below shows the percentage of within group and outside group collaboration for the study and comparison group separately. The dark blue bar is the percentage of within group collaboration and the light blue bar represents the outside group collaboration. Each of the bars sums to 100%, so that the proportion of within group collaboration can be compared across groups, controlling for distributional differences.

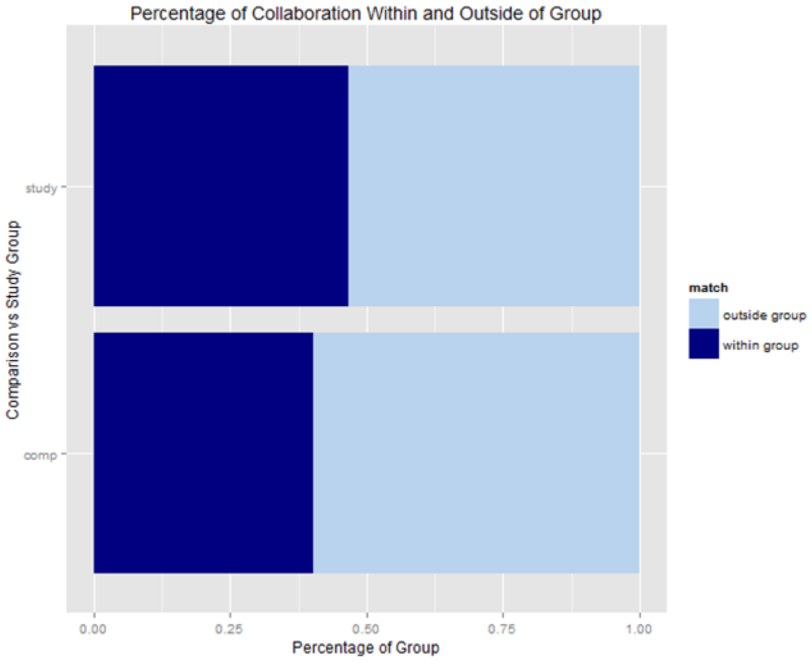


Figure 7. Percentage of within versus out of group collaboration by study and comparison group.

When comparing the percentage of within group collaboration for the study and comparison groups, it is found that the study group has a higher percentage compared to the comparison group. In order to gain insight on how tightly knit the communities are for each of the groups, it is helpful to calculate the number of complete triads. The study group has more complete triads in comparison to the comparison group, indicating that the network is more clustered. Therefore, yes, NIGMS supported researchers collaborate more with other NIGMS supported researchers than the comparison group, and their network is more clustered.

Study Question 2

The second study asked whether the type of grant awarded influences collaboration behavior. In terms of data, "type of grant" corresponds to the Activity Code associated with a certain project. Since grant information was provided at a project level, there were multiple activity codes for each researcher and for each publication. As mentioned previously, researchers and publications were assigned Activity Codes based on the grant type associated with the grant type with the highest amount of funding associated with the grant. Assortativity by grant type is slightly positive at 0.10, indicating a tendency for researchers awarded similar grant types are more likely to collaborate. The figure below shows the percentage of within group collaboration versus outside group collaboration for each of the most common grant types.

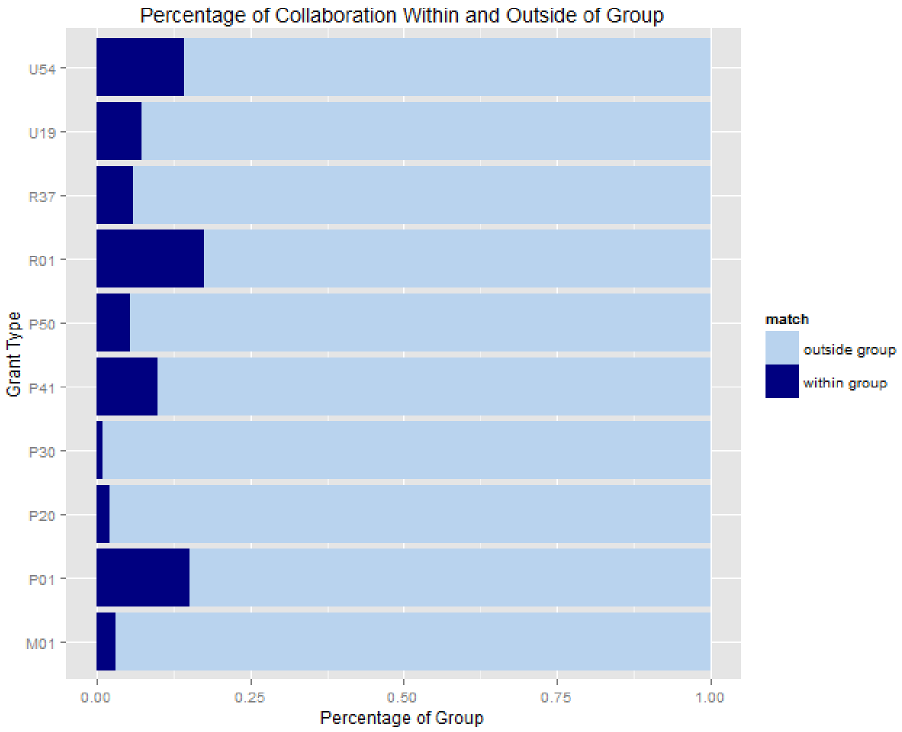


Figure 8. Percentage of within versus out of group collaboration by grant type.

As compared to the other grant types, researchers classified as the following tend to collaborate more within their respective grant type than with other grant types:

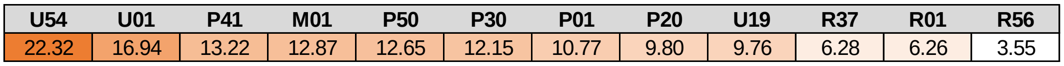
- R01 - NIH Research Project Grant Program

- P01 - Research Program Projects
- U54 - Spec. Center - Cooperative Agreements
- P41 - Biotechnology Resource Grants

Alternatively, researchers classified as the following tend to collaborate more outside of their own primary grant type:

- P30 - Center Core Grants
- P20 - Exploratory Grants
- M01 - General Clinical Research Centers Program
- P50 - Spec. Center
- R37 - MERIT Award
- U19 - Research Program--Cooperative Agreements

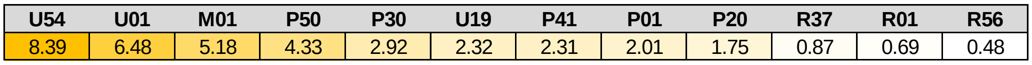
In the figure below, grant types are listed in the order of highest to lowest average degree. Researchers with grant type U54 and U01 tend to have the highest degree of connection, and these are both cooperative agreement types, which tend to be more expensive and expansive grants.



U54	U01	P41	M01	P50	P30	P01	P20	U19	R37	R01	R56
22.32	16.94	13.22	12.87	12.65	12.15	10.77	9.80	9.76	6.28	6.26	3.55

Figure 9. Ranking of grant types from highest to lowest average degree per researcher.

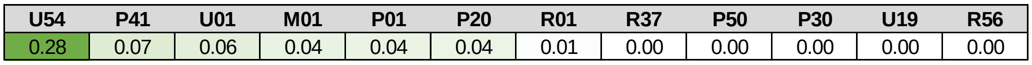
Because grant type was assigned to researcher by highest funding amount, this can cause a bias since grants with more funding are likely to have more collaboration. Grants with more funding are usually larger projects, and some even require a certain level of collaboration to be awarded. Therefore, it is necessary to also look at the average degree when weighting edges to control for these differences. The weighting method gives less weight to publications classified as grant types which have more collaboration by nature, and higher weight to publications which naturally show up less frequently. Grant types ranked in order of descending strength are shown below. The order of rank did not change significantly when controlling for distributional differences.



U54	U01	M01	P50	P30	U19	P41	P01	P20	R37	R01	R56
8.39	6.48	5.18	4.33	2.92	2.32	2.31	2.01	1.75	0.87	0.69	0.48

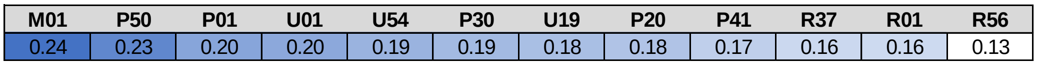
Figure 10. Ranking of grant types from highest to lowest weighted degree.

Triad count and transitivity both measure the tightness of the community. In order to account for the differences in node distribution, the triad count is set relative to the total number of nodes for that particular category. U54 and U01 decreased in rank for transitivity, but still are in the upper half of the ranking.



U54	P41	U01	M01	P01	P20	R01	R37	P50	P30	U19	R56
0.28	0.07	0.06	0.04	0.04	0.04	0.01	0.00	0.00	0.00	0.00	0.00

Figure 11. Ranking of grant types from highest to lowest node to triad count ratio.



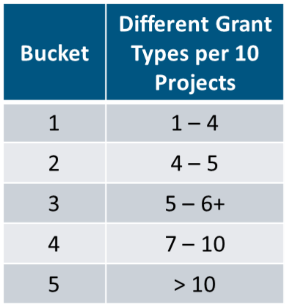
M01	P50	P01	U01	U54	P30	U19	P20	P41	R37	R01	R56
0.24	0.23	0.20	0.20	0.19	0.19	0.18	0.18	0.17	0.16	0.16	0.13

Figure 12. Ranking of grant types from highest to lowest transitivity.

Overall, grant types influence collaboration behavior and are an indicator of homophily. Cooperative agreements tend to foster more collaboration. This is shown from U54 and U01 grant types performing well for basically all of the network measures. U54 (Specialized Center) performs better than U01 (Research Project) in terms of collaboration. U54 has more collaboration per researcher, while U01 shows more collaboration within the same groups of researchers, which is shown from the transitivity ranking and within and outside of group bar chart. R56 (high-priority, short-term project award) has extremely low collaboration. Finally, R01 grants, despite being the most common, has low levels of collaboration.

Study Question 3

As a direct extension of the second study question, study question 3 asked whether a variety of funding sources influence collaboration behavior. As mentioned previously, for the purposes of this project, variety in funding sources is defined as the number of unique grant types for each researcher. Since there is a greater chance that researchers will have a greater variety in funding sources if they have more projects, the number of unique grant types had to be adjusted so that researcher could be compared regardless of the number of projects they were involved with. To do so, the number of unique grant types was divided by the total number of projects, which provide a value representative of number of unique grant types per project. Since this was a continuous number, values were grouped into quantiles. The following figure shows the number of unique grant types per ten projects for each of the quantile buckets.



Bucket	Different Grant Types per 10 Projects
1	1 – 4
2	4 – 5
3	5 – 6+
4	7 – 10
5	> 10

Figure 13. Legend of the number of distinct grant types per project for each quantile bucket.

When grouping by quantile, there is a slightly positive assortativity of 0.03. This indicates that there is some homophily between nodes with similar levels of variety in funding. The figure below is the bar chart showing the within and outside group collaboration by quantile.

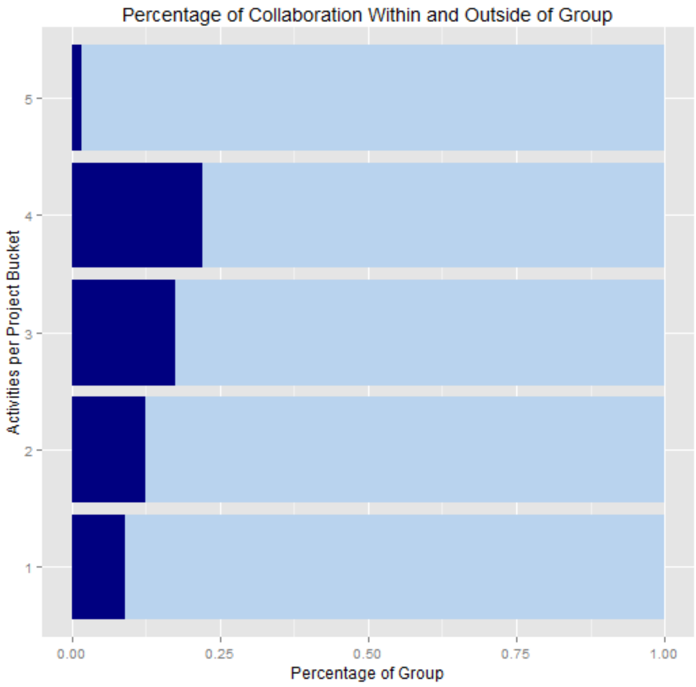
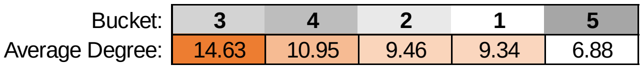


Figure 14. Percentage of within versus out of group collaboration by grant type per project quantile.

The results for quantiles 1-4 show that likelihood of within group collaboration increases as the variety of funding sources increase. The fifth quantile has a smaller amount of publications because of the treatment of ties when creating the buckets, and realistically, there are not a lot of researchers having greater than10 different grant types per 10 publications, so this grouping naturally captures the outlying researchers. Because the data within this bucket is thin, conclusions should not be made with confidence from the behavior of the fifth bucket. For the majority of data, researchers with a similar level of variety in funding tend to collaborate more at an increasing rate.

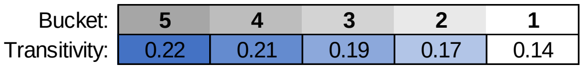
To further analyze collaboration behavior, similar network statistics to the previous study questions were calculated. The following table ranks the quantiles from highest to lowest average degree. It is shown that researchers with a moderate amount of variety in funding sources have the most collaboration, on average.



Bucket:	3	4	2	1	5
Average Degree:	14.63	10.95	9.46	9.34	6.88

Figure 15. Ranking of quantile from highest to lowest average degree per researcher.

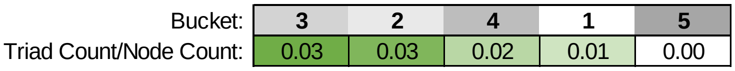
To better understand clustering of the quantiles and how tightly knit communities within the groups are, transitivity and triad count to node count ratio values were calculated. A clear pattern emerged from the transitivity ranking by quantile. As variety in funding sources increases, the transitivity also increases, indicating that researchers who have more variety in funding tend to have more tightly knit communities.



Bucket:	5	4	3	2	1
Transitivity:	0.22	0.21	0.19	0.17	0.14

Figure 16. Ranking of quantile from highest to lowest transitivity.

The triad count to node count is another metric for indicating connectedness among the researchers' collaborative partners. Researchers with a moderate amount of variety in funding sources have the highest triad to node ratio, as seen in the figure below.



Bucket:	3	2	4	1	5
Triad Count/Node Count:	0.03	0.03	0.02	0.01	0.00

Figure 17. Ranking of quantile from highest to lowest triad to node count ratio.

In assessing the statistics overall, Group 1, which represents researchers with 1-4 different grant types per 10 projects tend to have low transitivity and triad count to node count ratios. They are low ranking for degree, but when looking at actual values, have a relatively average number of connections. Therefore, researchers with low variety in funding have less tightly knit communities although average connections are comparable to the remainder of quantiles. Overall, variety in funding sources does influence collaboration behavior.

Study Question 4

The last study question asked whether scientific discipline influences collaboration behavior. Discipline information was attached to publication information and then assigned to researchers from the methodology discussed previously. When grouping by scientific discipline, there was a positive assortativity measure of 0.16, indicating that there is homophily by scientific discipline. The figure below shows the within versus outside group bar chart by scientific discipline is particularly useful to answer this study question, as the outside group collaboration represents interdisciplinary collaboration.

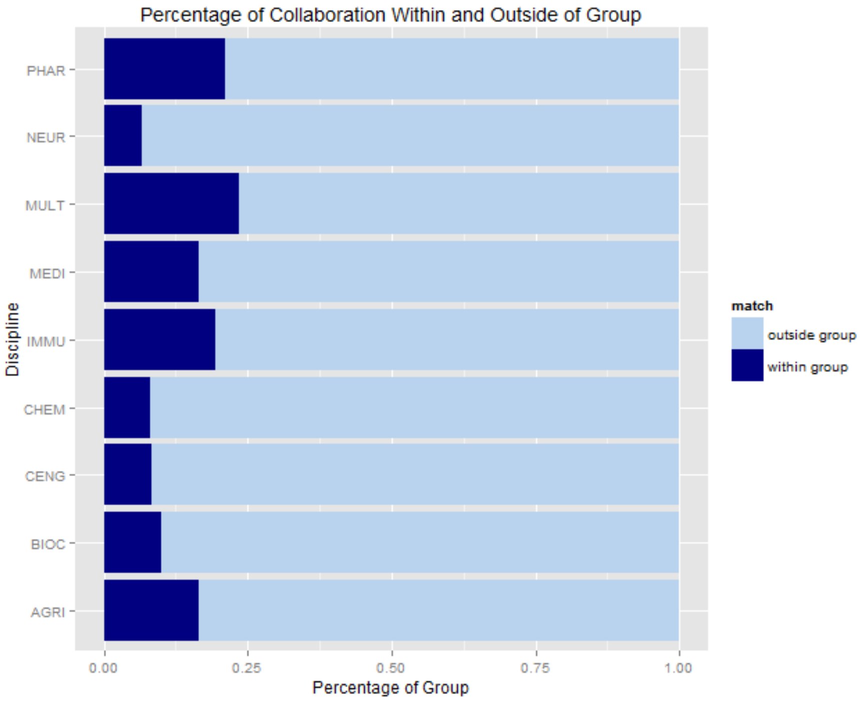


Figure 18. Percentage of within versus out of group collaboration by discipline.

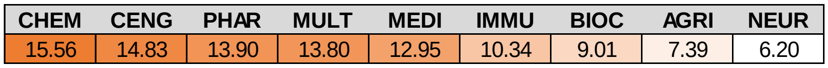
As compared to the other disciplines, researchers classified as the following tend to collaborate more within their respective discipline than with other disciplines:

- Pharmacology, Toxicology, and Pharmaceutics
- Medicine
- Immunology and Microbiology
- Agricultural and Bio Sciences

Researchers classified as the following tend to collaborate more outside of their own discipline:

- Neuroscience
- Chemistry
- Chemical Engineering and Biochemistry
- Genetics and Molecular Biology

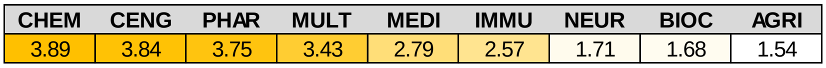
To further analyze collaboration behavior, similar network statistics to the previous study questions were calculated. As seen in the figure below, disciplines with more interdisciplinary collaboration such as Chemical Engineering and Chemistry have a higher average degree than the other disciplines.



CHEM	CENG	PHAR	MULT	MEDI	IMMU	BIOC	AGRI	NEUR
15.56	14.83	13.90	13.80	12.95	10.34	9.01	7.39	6.20

Figure 19. Ranking of discipline from highest to lowest average degree per researcher.

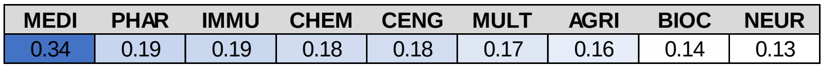
Similarly to the grant types, certain disciplines inherently tend to publish more than others. The weighted degree measure accounts for those differences. Publications are assigned weights based on each discipline’s frequency of publications, so that differences in distribution can be controlled for. Even when accounting for these differences, the order of the weighted degree stayed mostly consistent. This is partially due to the fact that distribution was taken into consideration when assigning discipline to node, as discussed previously.



CHEM	CENG	PHAR	MULT	MEDI	IMMU	NEUR	BIOC	AGRI
3.89	3.84	3.75	3.43	2.79	2.57	1.71	1.68	1.54

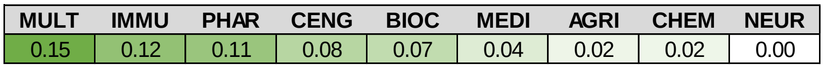
Figure 20. Ranking of discipline from highest to lowest weighted average degree per researcher.

For transitivity and triad to node count ratio, disciplines with more within group collaboration tended to have more tightly knit communities. Medicine has a very high transitivity measure as compared to the other disciplines, which means that it has a very close-knit network and a relatively high amount of internal collaboration. In terms of triad count to node count ratio, Medicine drops down in ranking and Immunology and Pharmacology increase in rank.



MEDI	PHAR	IMMU	CHEM	CENG	MULT	AGRI	BIOC	NEUR
0.34	0.19	0.19	0.18	0.18	0.17	0.16	0.14	0.13

Figure 21. Ranking of discipline from highest to lowest transitivity.



MULT	IMMU	PHAR	CENG	BIOC	MEDI	AGRI	CHEM	NEUR
0.15	0.12	0.11	0.08	0.07	0.04	0.02	0.02	0.00

Figure 22. Ranking of discipline from highest to lowest triad to node count ratio.

Overall, scientific discipline does influence collaboration behavior and is an indicator of homophily. Disciplines with more interdisciplinary collaboration (such as Chemistry and Chemical Engineering) have higher average degrees than other disciplines.

Chapter 5: Conclusions

The purpose of this paper was to answer four study questions focused on collaboration behavior, specifically focusing on aspects associated with the researchers and their publications, such as grant type, variety in funding and discipline. With extensive data cleaning, wrangling, and employing various tools such as python and igraph in R, valuable insights were discovered and observations formed, in the hopes of providing actual and practical implications that could foster collaborations.

Patterns of collaboration behavior were observed within the study and comparison group. The study group exhibited more collaborations within themselves than the comparison group. NIGMS supported researchers are more likely to collaborate with other researchers within the same group. Also, grant type, or the activity code attached to each researcher, was deemed to be relevant in its effect on collaboration among researchers. Interestingly, the most common grant type didn’t have the most influence. With the extension of grant type analysis, funding variety provided further proof that the effect was indeed apparent. As the variety in funding increased, the likelihood of researchers collaborating within the same group increased accordingly. Discipline played a role in collaboration as well. In particular, interdisciplinary fields tended to have more collaboration than the rest. The exact effect of the discussed factors are hard to measure quantitatively. Also, these are a few of many factors that could have profound impact on overall collaboration behavior. Nonetheless, the observations and insights derived from this paper could be utilized and considered for future studies. These factors, considered and applied, individually or collectively, can serve as qualitative measures to assist with efforts that aim to foster and inspire collaborations among researchers.

References

Cherven, K. (2015). Mastering Gephi network visualization: produce advanced network graphs in Gephi and gain valuable insights into your network datasets. Birmingham, U.K.: Packt Pub.

De Paula Fonseca, B., Sampaio, R. B., Vinicius de Araújo Fonseca, M., & Zicker, F. (2016). Co-authorship network analysis in health research: method and potential use. Health Research Policy and Systems, 14(34). doi:10.1186/s12961-016-0104-5

Elmacioglu, E., & Lee, D. (2005). On Six Degrees of Separation in DBLP-DB and More. SIGMOD Record, 34(2), 33-40.

Newman, M. E. (2001). The structure of scientific collaboration networks. PNAS, 98(2), 404-409.

Onel, S., Zeid, A., & Kamarthi, S. (2011). The structure and analysis of nanotechnology co-author and citation networks. Scientometrics, 89, 119-138. doi:DOI 10.1007/s11192-011-0434-6

The Open Graph Viz Platform. (n.d.). Retrieved May 04, 2017, from https://gephi.org/

National Institutes of Health (NIH). (n.d.). Retrieved May 04, 2017, from https://www.nih.gov/

(n.d.). Retrieved May 04, 2017, from https://projectreporter.nih.gov/reporter.cfm

(n.d.). Retrieved May 04, 2017, from https://exporter.nih.gov/

(n.d.). Retrieved May 04, 2017, from https://federalreporter.nih.gov/

(n.d.). Retrieved May 04, 2017, from https://www.scopus.com/home.uri

(n.d.). Retrieved May 04, 2017, from http://apps.webofknowledge.com