PubMed AuthorMap: A Study of Research Collaboration in the Life Science Fields

An Interdisciplinary Major Qualifying Project Report
Submitted to the Faculty of the
WORCESTER POLYTECHNIC INSTITUTE
in partial fulfillment of the requirement for the
Degrees of Bachelor of Science in
Bioinformatics and Computational Biology and Biology/Biotechnology

By:

Amanda Moulaison

&

Haylea Northcott

Date

APPROVED:

Professor Dmitry Korkin Computer Science WPI Project Advisor	Professor Elizabeth Ryder Biology/Biotechnology WPI Project Advisor	Doctor Amir Mitchell Department of Systems Biology University of Massachusetts Medical School

Abstract

This project incorporates the PubMed API to create a tool that builds networks representing research collaborations across life science fields. In each network, a node represents the last author on a publication, while an edge represents publications that share one or several authors. Our tool shows that most networks demonstrate characteristics of scale-free networks but cannot be statistically proven to be scale free. We performed case studies on three networks to describe the relationship between bottleneck authors and their collaborators.

Acknowledgments

A special thank you to our three advisors of this project:

- Professor Dmitry Korkin for his time and dedication to us and the project over the past year. He taught us invaluable lessons and guided through the ups and downs of a bioinformatics research project.
- Professor Elizabeth Ryder for her constant accessibility and willingness to help by providing us with feedback from a biologist's perspective.
- Professor Amir Mitchell for always challenging us to exceed our own expectations.

1 Introduction

In 2013, it was approximated that biologists around the world produced 15 petabytes of data each year (Wired, 2013). This is widely attributed to the cost of DNA sequencing decreasing dramatically in 2009 which allowed it to become accessible to more labs than ever before. While growth in data is exciting because it means the science community is expanding its depth and breadth of knowledge, it also creates a cause for concern. How will we navigate and search such a large quantity of information? How can we still make this information useful and accessible to other students and researchers? How will researchers who are interested in similar fields share their knowledge? The answers to all of these questions lie in data science, and specifically data mining tools.

One area of particular concern regarding data management is within research publications. A single paper can take hours for a single person to read manually. There are some text mining tools on the market to automate the analysis of papers, but this is a particularly difficult task in biology due to the number of synonyms each term has, and the different ways to grammatically represent each one. For example, BRCA1 is just one gene that is related to breast cancer, but it alone has 17 synonymous terms, including BRCC1, breast cancer type 1 susceptibility protein, and BROVCA1 (PubChem 2015). This classification task is made even more difficult by the sheer quantity of publications available. At the end of 2018, it was estimated that the PubMed database, one of the most popular life science databases, contained over 29 million publications (PubMed 2018).

In this project, our goal is to build a tool using the PubMed API to visually represent the research that is being done in each life science field. Through this tool, users will be able to see where researchers and labs have collaborated, which will allow for a better understanding of the spread of knowledge through a field. From a practical standpoint, this will ease the process of researching a specific topic. For example, if a user identifies a single paper or author whose work was particularly interesting to them, they can plug the name of the paper or author into our tool. Using the network that our tool produces, the user can seek out the nearest nodes and edges around their input, which will represent papers and authors that were most similar to their input. This will ease the research process by weeding out less relevant publications, allowing researchers to focus their time and energy on the information that is most important and relevant to them.

2 Background

One of the most important aspects of successful research is collaboration. Collaboration can occur between single researchers or entire labs. Through collaboration, research topics and publications rapidly become intertwined through shared collaborators. This can be demonstrated by a network. In this section, we discuss networks and their structure, our database PubMed, and tools that have been developed in the past that we sought to improve on in our work.

2.1 Types of Networks

A network is defined as a set of nodes connected by edges or arcs. Networks are primarily used in math and computer science to model real-world problems or situations. Networks can be categorized by their characteristics such as connectivity and clustering, and there are many different types of networks. For example, one common application of graphs is to model social networks, where nodes represent people and edges represent a relationship. In our research, nodes will represent last authors on research papers and edges will represent supporting authors that connect last authors via a shared publication. The following sections will discuss a few specific types of graphs which are relevant to our research.

2.1.1 Scale-Free Networks

Scale-free networks are a specific type of network that are characterized by the presence of hubs in the network. Hubs are defined as a few nodes which are highly connected to other nodes in the network. The presence of hubs shifts the distribution of node degree - the average number of connections that each node has - of the overall graph, causing them to follow a power law distribution (Barabási,2016). This distribution is shown in Figure 1. This is the main defining feature of scale free networks. Real-world examples that tend to follow a scale free network include the transmission of diseases, the internet, and social media interactions.

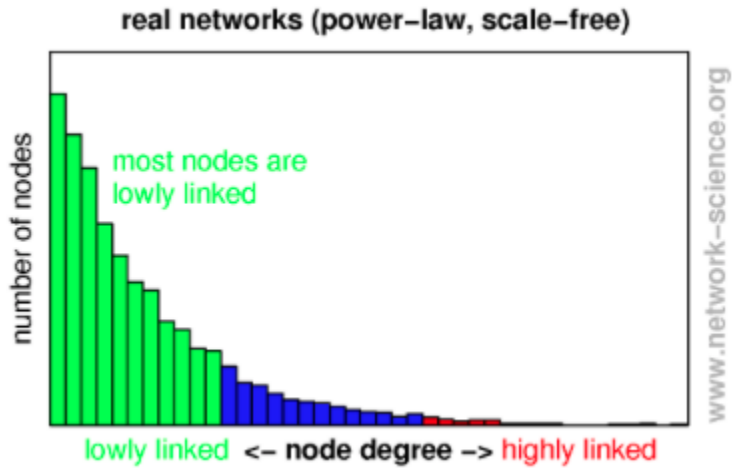

Figure 1: Node Degree Distribution in a Scale-Free Graph

2.1.2 Random Networks

Random networks are much more loosely defined than scale-free. Traditionally, a random graph is defined as a network consisting of N nodes where each node is connected by probability p (Barabási,2016). These networks do not consistently follow any patterns, like scale free or clique graphs. Random networks node degree tend to follow a Poisson distribution, seen below in Figure 2. There are no common examples of random networks in the real world and they are traditionally used only for theoretical work (Barabási,2016).

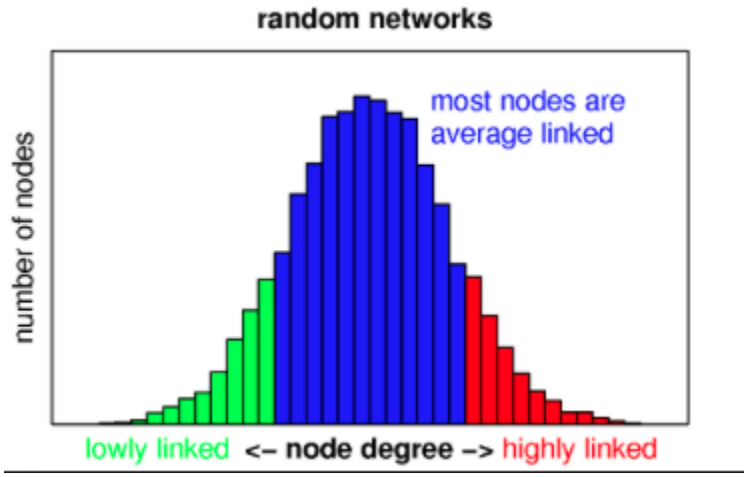

Figure 2: Node Degree Distribution in a Random Graph

2.2 PubMed, a Life Sciences Database

PubMed is a key information resource in the biological sciences in terms of diversity, breadth, and manual curation (Douglas et al., 2005). PubMed is a free search engine which is comprised of 28 million citations retrieved from the MEDLINE database. MEDLINE hosts all the references and abstracts of life science topics and PubMed allows for users to retrieve the full text of these abstracts and interact between the references on the publications. Many tools come with PubMed including citation matchers, clinical queries, and topic-specific queries. Many developers want to add to their list of tools to enhance the search engine, thus PubMeds API known as E-Utilities, is available online to developers to allow access to their database.

2.2.1 PubMed API Toolkit

In 2008, NCBI released an API search toolkit for PubMed. This toolkit was built with the intention of giving researchers the tools they need to be able to build their own PubMed search applications efficiently. The kit consists of 9 tools, and our project will implement 5: E-Search, E-Fetch, E-Summary, ECQuery, and E-Spell. The following sections will discuss these tools in detail.

2.2.2 E-Search

E-Search is the tool which provides the basic search function. This tool allows the user to input a search query and returns UIDs of all articles in the database which contain matching terms. This is the basic tool that our program will be based on as it will perform the primary searching function. It's only required parameters are the database of interest to search (default value is PubMed) and the search term (Sayers, 2010). There are a number of additional optional parameters, which allow the user to set the number of results to return, the format to return them in, and the method with which to sort the results. One additional optional parameter is History. By setting this parameter to 'Y', the user can store the search in the NCBI History server, which will allow the user to access it later in a subsequent EUtilities call. This is integral to chaining several EUtilities calls. An example output from the E-search utility is seen in Figure 3 below, in the format of an XML file. The multiple PubMed IDs are listed within the <IdList>.

```
<?xml                                    version="1.0"                                    ?>
<!DOCTYPE    eSearchResult    PUBLIC    "-//NLM//DTD    eSearchResult,    11    May    2002//EN"
"https://www.ncbi.nlm.nih.gov/entrez/query/DTD/eSearch_020511.dtd">
<eSearchResult>
<Count>255147</Count>              #    total    number    of    records    matching    query
<RetMax>20</RetMax>#    number    of    UIDs    returned    in    this    XML;    default=20
<RetStart>0</RetStart>#      index      of      first      record      returned;      default=0
<QueryKey>1</QueryKey>#       QueryKey,       only       present       if       &usehistory=y
<WebEnv>0l93yIkBjmM60UBXuvBvPfBIq8-9nIsldXuMP0hhuMH-
8GjCz7F_Dz1XL6z@397033B29A81FB01_0038SID</WebEnv>
                 #         WebEnv;         only         present         if         &usehistory=y
      <IdList>
<Id>229486465</Id>                              #       list       of       UIDs       returned
<Id>229486321</Id>
<Id>229485738</Id>
<Id>229470359</Id>
<Id>229463047</Id>
<Id>229463037</Id>
<Id>229463022</Id>
<Id>229463019</Id>
<Id>229463007</Id>
<Id>229463002</Id>
<Id>229463000</Id>
<Id>229462974</Id>
<Id>229462961</Id>
<Id>229462956</Id>
<Id>229462921</Id>
<Id>229462905</Id>
<Id>229462899</Id>
<Id>229462873</Id>
<Id>229462863</Id>
<Id>229462862</Id>
</IdList>
```

Figure 3: An Example of E-Search Output

2.2.3 E-Fetch

The E-Fetch tool returns a list of formatted records for the given inputted list of UIDs. This tool can take a list of UIDs directly as fetch parameters, but it also can be chained with requests to other EUtilities tools using the NCBI History server. For example, a user can perform an E-Search for "stem cells" as shown in the example above, which will return a list of matching UIDs. If this query is saved to the History server, the user can then make a subsequent call to E-Fetch through the History server which will allow them to get to the list of the full formatted records for each of the UIDs returned by the E-Search call (Sayers, 2010). One additional optional parameter for this tool is complexity. Within the call to E-Fetch, the user can instruct the tool how much detail to return in the full formatted record. The parameter is a scale from 0-4, where 0 returns maximum information (e.g., the entire abstract) and 4 returns minimum information (the sentence the query appears in).

2.2.4 E-Summary

PubMeds API provides a tool that can return document summaries on every publication with a list input of unique identifiers (UIDs), this tool is called E-summary. The tool utilizes DocSums as their summary tool. The designed function of the tool requires four parameters including database, UIDs, query_key, and WebEnv. The database is defaulted to PubMed but can be changed to explore genes or proteins. The UID list is required and must be comma-delimited, there is no set maximum of number of UIDs for the Esummary tool. Query_key is an output by a tool used earlier to find the object to summarize being one of the three: Esearch, EPost or Elink. WebEnv is used in conjunction with WebEnv and specifies the web environment that contains the UID list to be provided as input (Sayers, 2010). This function is especially helpful within a network graph to allow users to have a summary of the publication before further exploring the whole text. The output is an XML file with the contents seen below in Figure 4.

Figure 4: Sample XML file output from the ESummary tool

As seen in the above figure, the summary included for each publication is extremely detailed and includes the pubMed ID, date published, last author, author list, UID, title, source, in addition to many other fields. A total of thirty different fields help summarize each publication along with the list of the UIDs.

2.2.5 EGQuery

The main function of EGQuery is to provide the number of records retrieved in all Entrez databases by a single text query. EGQuery pairs well with the Esearch tool. It is a beneficial tool due to the size of some searches (e.g., cancer) may exceed capacity limits and are unable to analyze or visualize every returned record. Being able to show the proportion of records visualized on the screen provides a datapoint for the user (Sayers, 2010).

2.2.6 ESpell

ESpell is a tool to be implemented to enhance the user experience on a application. ESpell provides spelling suggestions for terms within a single text query. The required parameters for this PubMed tools is a database to search along with a term to query. All special characters must be URL encoded and any spaces have to be replaced with '+' signs. Espell increases the ease to search for any term with the ability to correct the spelling if either by accident or unknown (Sayers, 2010).

2.3 Existing PubMed Search Tools

Many developers have already taken advantage of the opportunity to help enhance PubMed's well-used search engine and have created tools with a variety of functions. Some tools create models or return summaries or even perform statistics on the results. Most tools use text mining and NLP as a base for their application but each focuses on extraction of different information. For example some tools use MEDLINE abstracts while others use MeSH terms, authors, key terms, genes, or proteins, to assist in enhancing the literature research process or make new discoveries. Every tool breaks down the current limitations of PubMed and advances its usability along with the creating possibility to extract new information.

2.3.1 Chilibot

Chilibot is a natural language processing based text-mining internet application that extracts and defines relationships networks from PubMed abstracts based on biological concepts, genes, proteins, or drugs (Chen, H. & Sharp, B.M., 2004). Chilibot is different from most text mining programs that already exist because it focuses not only on similarities in the text, but also characterizes each interaction. For example, the designer of this application included directionality in the graph as well as implemented shapes to represent the presence of inhibition versus simulation in a relationship. This is accomplished by taking the title and abstract from MEDLINE and parsing the abstract into units of one sentence to obtain higher performance levels.

A relationship map created by Chilibot is shown below in Figure 5. The PubMed database in this example was queried looking to discover relationships amongst a set of genes regulated by cocaine. In their graph they have icons on the network lines, which represent the relationship. Arrowheads indicate directionality, while the different colored circles show the interactive relationships between the two whether it is neutral (gray), stimulatory (green), inhibitory (red), or both (yellow) (Chen, H. & Sharp, B.M., 2004).

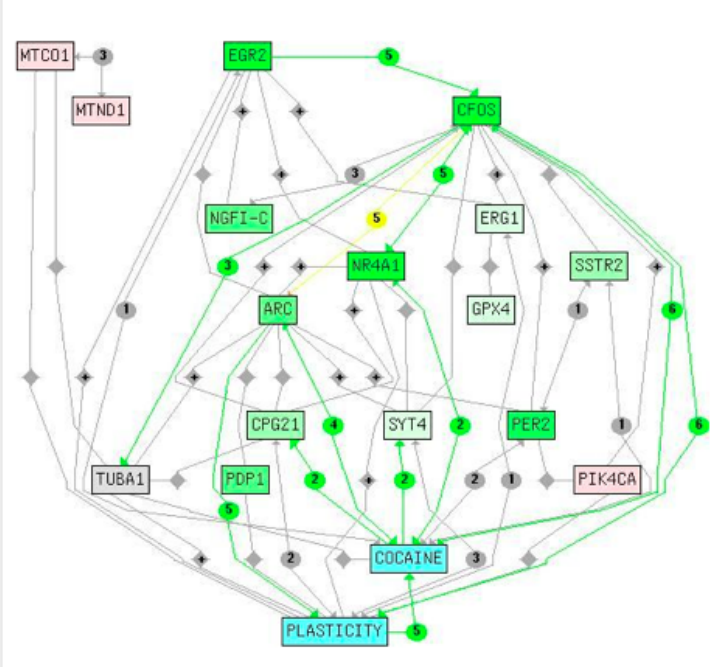

Figure 5: An example of the relationship network from the program Chilibot (Chen, H. & Sharp, B.M., 2004).

These networks have the power to help discover new trends and hypotheses about the queried biological concept, gene, or drug. Performing analyses on this network and discovering the hubs and topology class can help discover potential experimental targets. Chilibot only performs linguistic analysis on maximum of 30 abstracts per query so the possibility of the network falsely representing the queried terms due to small sample size is a possibility (Chen, H. & Sharp, B.M., 2004). This also allows for important articles to be absent from the selected data set. In Figure 5 above, the term "cocaine" is queried and Chilibot selects and dissects 30 abstracts to create the above network. However, there are 40,807 abstracts about "cocaine" in PubMed, meaning that this search result only represents 0.07% of the available research. This leads readers to question the accuracy and representation the graph provides of the queried term.

2.3.2 MeSH Map

Within the MEDLINE abstracts on PubMed, there exists controlled vocabulary that assists in indexing subjects on documents called medical subject heading (MeSH). A lot of text mining softwares rely on MeSH terms to create and strengthen links between two different entities. MeSH terms help represent similarities between two abstracts that may be on two completely different topics. MeSHmap is a prototype application that supports searches via PubMed and generates maps by "user driven exploration of MeSH terms" (Srinivasan, 2001). This application can also mine the metadata of the MEDLINE abstracts to provide a high level summary of the users' search or describe the relationship between two topics, for example a pair of drugs or procedures. MeSHmap is written in the language Java and has three major interaction phases: search, exploration, and display.

This application goes beyond just text retrieval with its three user friendly designed phases. The search phase consists of using the application user interface to query for a disease or term. The user may choose to string together a variety of terms using high level operators such as "AND" and "OR" which allows for the discovery of association between a variety of different topics that may trigger new research (Srinivasan, 2001). All results are then downloaded and analyzed using the MeSH terms and subheadings provided by the MEDLINE abstract. After analysis, a summary is formed in addition to a simple non-interactive generated map which is visualized on screen in the application. In addition, two lists are generated: one that contains a list of MeSH terms with frequency of occurrence and another list of the subheadings from all resultant abstracts. MeSHmap will also list all the titles of the papers which were analyzed in a separate window by user request with use of a fetch function.

Srinivasan remarked, "Given the explosion of information in health care it is very difficult for health care professionals, researchers and educators to keep abreast of literature in their domain." (Srinivasan, 2001). This program in addition to Chilibot gives the life sciences field critical tools to be able to filter through and obtain desired literature with ease. With these tools, we open a door of access to all and the ability to discover new knowledge through graph theory. Unfortunately, many programs including MeSHmap are not open to the public.

2.3.3 PubNet

The most successful PubMed text mining web based tool, based on usage, is called PubNet. PubNet can extract and visualize a variety of relationships between publications using aspects such as gene names, protein data bank (PDB) IDs, MeSH terms, vocabulary terms, and authors (Douglas et al., 2005). PubNet has a user interface in which a term of interest is queried to bring up the search results from PubMed and is displayed in a interactive graphical visualization. The XML output from PubMed is parsed to uncover many interesting relationships between the publications that are returned from the query.

PubNet does everything Chilibot and MeSHmap can do with some additions. PubNet is connected to a software known as TopNet. TopNet takes the graph displayed from PubNet as input and calculates the average degree, clustering coefficient, characteristic path length, and diameter for the network. PubNet is also capable of creating much more complex networks than the other two applications because it can accept multiple queries and can select node parameters for each network (Douglas et al., 2005). The application is user friendly and interactive, in just one window, each node in the graph is hyperlinked to a detailed textual report which organizes all outgoing edges and neighboring nodes with respective edges in a list.

This application was a new advancement in high throughput techniques and has made it possible to conduct biomedical research on a larger scale with ease. PubNet has different search types based on exploring a variety of relationships between publications within the database. One of the search types which closely relates to our project is their authorship function. Querying a specific author or organization would return a network where each node is an author and the edges represent co-authorship. A diverse array of graph structures is evident which highlights differences

in size, frequency in publication, and degree cooperation across the consortia. This diverse array is seen in Figure 6.

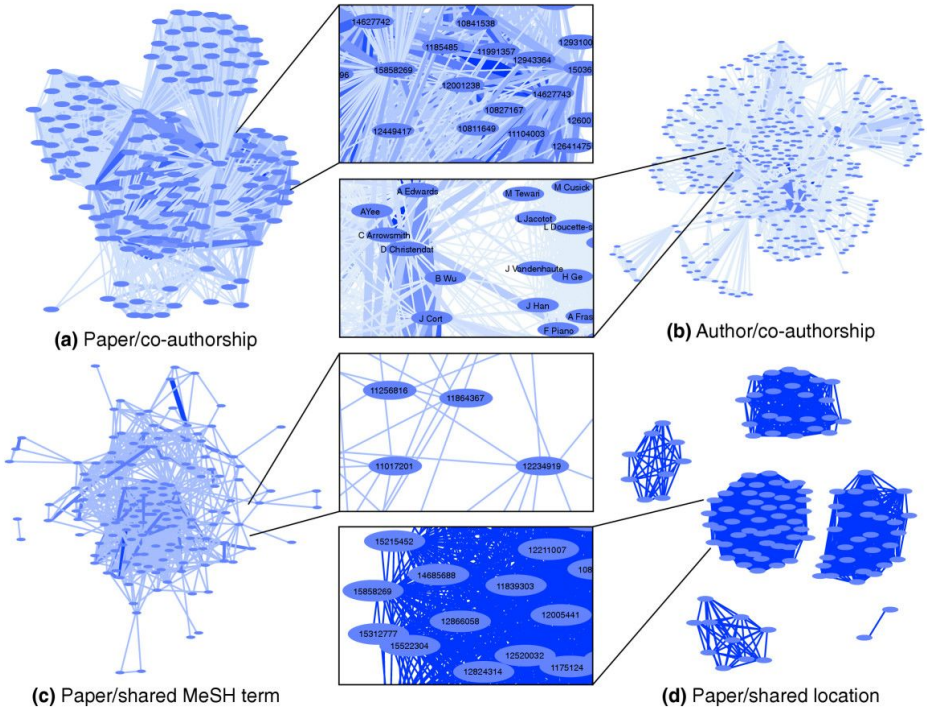

Figure 6: Examples of PubNet display network based on authorship query function (Douglas et al., 2005). (a) Network with edges representing co authorship on one paper, (b) Network with edges representing collaboration between two authors, (c) Network with edges representing shared MeSH term between papers, (d) Network with edges representing a shared location between two papers, major clustering can be seen here by country or continent.

Figure 6 illustrates the different types of relationships that can be extracted from a single query on PubMed. An example of the authorship search type being used is seen above in panel B of Figure 6. The graph represents the authorship for the NESG consortium, which shows a confederated but coordinated approach. Each author name is stored within a node of the network returned from the search query. Each edge within the network represents co-authorship on at least one paper together. A different pattern is seen in this network versus the others in the figure. Researchers believe this unique pattern is because investigators from each laboratory tend to form central anchor points, from which other members of the laboratory use to branch out (Douglas et al., 2005). The NESG consortium consists of two protein sample production centers, six sites for 3D structure detection via nuclear magnetic resonance or X-ray crystallography, and several other groups working on technology development and annotation. This network allows a user to understand the connection between authors in this consortium.

Although PubNet has broken through a lot of barriers within biomedical research, there still is a lot of improvement to come as technology continues to advance. PubNet can only handle

15 combinations of node and edge parameters although the number of different queries is unrestricted. Also, in the case of researchers with common names (e.g., “Smith”), the application has no way to differentiate between the authors. This opens the possibility of these researchers to be grouped into the same node as one author, thus providing an inaccurate network.

2.4 Our Project

AuthorMap will be designed to incorporate features from each of these preexisting tools while also having a novel approach. This application will be used to discover the connections between authors based on a topic, which is different from PubNet in which users can only search for a consortium. A user would input a string to AuthorMap to query PubMed such as “tuberculosis” or “breast cancer”. The user would also set a size to the network, based on the amount of publications they would like the search to return. For example, if the set size is 500, then 500 papers will be analyzed to create the network. In each network, every node represents a last author, who is usually the group leader or PI and each edge shows co-authorship between last authors. The tool Chilibot can only perform analysis on 30 publications per network exported, whereas AuthorMap’s maximum search size has not been found yet. The tool MeshMap is not open to the public although its in-depth analysis of the MeSH terms could offer a great advantage which AuthorMap will display for all returned papers but also be available to all users. The application AuthorMap has an objective to be open-source, user-friendly, interactive, have a new technological capacity, and answer new questions.

3 Methodology

This section describes our approach of implementing the various PubMed utilities, NetworkX, Cytoscape, and Powerlaw package in Python to obtain and analyze authorship networks. Shown in Figure 7 below is our experimental design and the application of the various tools and the sequential order in which they are implemented.

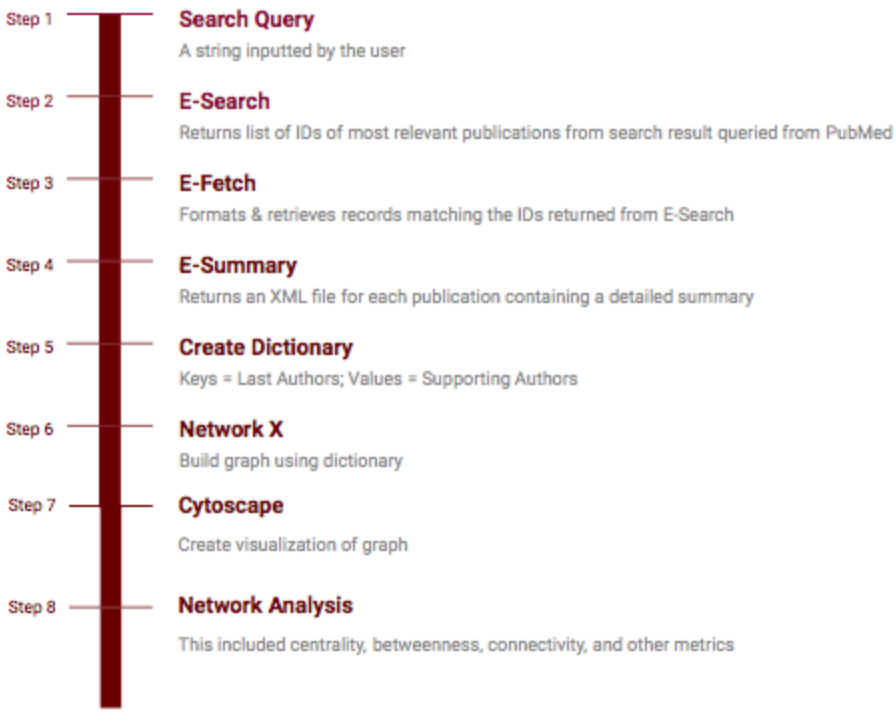

Figure 7: Experimental Design of Program in Python to Obtain and Analyze Authorship Networks

3.1 Tools, Environment

For this project, we worked primarily in Python 3.6, specifically using the Pycharm IDE. Significant libraries that we implemented included Entrez, Matplotlib, iPython, NetworkX, Py2Cytoscape, and PowerLaw. Py2Cytoscape allowed for a connection between Python and the Cytoscape application for visualization of our networks. Entrez served as the connection between our application and the PubMed API. Matplotlib, iPython, NetworkX, and PowerLaw mainly served in the creation and analysis of our networks.

3.2 Design of AuthorMap

In this section of the methodology, the AuthorMap tool is discussed in detail about of how it was created in Python 3.6 from the approach, to design, and usage of packages that made the tool fully functional.

3.2.1 Systematic Approach

To ensure our tool was systematic, first we needed to ascertain that we were selecting search terms in a non-biased method. To do this, we implemented the PubMed Mesh Terms (found at: https://www.ncbi.nlm.nih.gov/mesh?term=Biological%20Science%20Disciplines) in a systematic method. We created a list of terms at the primary level of the hierarchy to perform our preliminary search. In this way, our program automatically iterated through the list, searched each term on PubMed, created a Cytoscape graph of each network result, and performed analysis on each resultant network for each term. We also followed this exact practice with a Narrow List that was the secondary level of the hierarchy. Lastly, we delve into several case studies based on analysis of the Narrow List.

3.2.2 Search Engine

To build our search engine, we used the E-Utilities tools available through the PubMed API called Entrez. With this API, we were able to query searches to the PubMed database without having to perform web scraping, which allowed our data management to be more time and space efficient.

To gather the data, we first implemented the E-Utilities Search tool. This is the initial function that passes the user's search query to the database, and returns PubMed IDs of papers that match the search term. This function includes several optional parameters built-in. A few that we chose to use were 'retmax', which sets the maximum number of results to return, and 'retmode', which sets the return file type, for which we used xml.

After acquiring the search results, the next step was to gain more information about every individual PubMed ID returned. Because the search function only returns IDs of papers matching the search query, we wanted to next learn more about the returned papers, such as their title, date published, and authors. The E-Fetch function fetches the paper IDs that were returned by the ESearch function. Then, the E-Utilities Summary function takes in a list of formatted files from the E-Fetch function and returns key information about them in an xml file, including title and authors. The E-Summary utility allowed for retrieval of this key information and was implemented into our program.

The final step in handling our results was to parse the returned XML files from the E-Summary tool to deduce only the information that we needed. Our main interests were the authors who contributed to each paper and the paper titles, because this data is integral in building our network. After parsing, we stored the authors, last authors and supporting, in a dictionary to make the information readily available to convert into a network.

3.2.3 Cytoscape

Cytoscape is one of the most popular network visualization libraries used for bioinformatics and data science. For this project, we used it for the visualization of our author networks. After we assembled our graph in NetworkX, we passed it directly to the Cytoscape application through a local host, where it was visualized. Below, in Figure 8, is a sample Cytoscape

output of a search for “anatomy”. Each node is representative of a last author and is labeled with the author’s name and each edge signifies collaboration with another last author in the field.

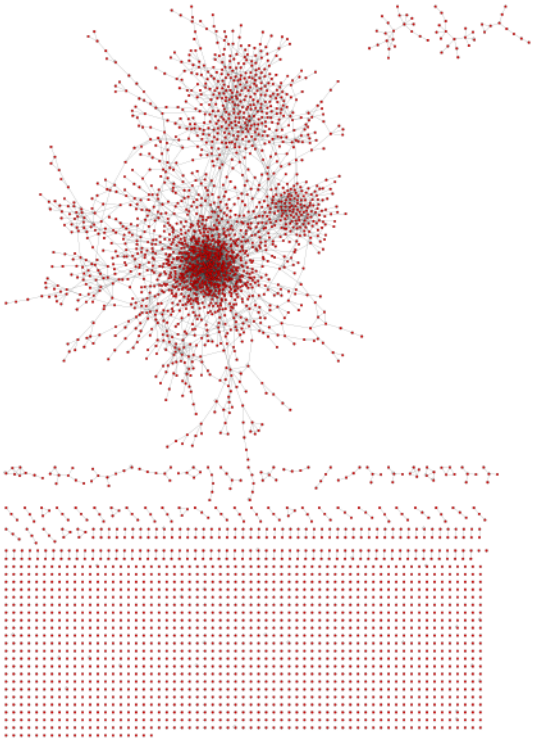

Figure 8: Example Cytoscape Output for query term “chemotherapy” with 500,000 search results. At the bottom of the network all binary nodes are seen.

3.2.4 NetworkX

The dictionary containing the last authors as the keys, and all supporting last authors as the values, were then made into a undirected simple network using NetworkX. In each network, a node represents a last author within the queried field. Supporting authors are represented as edges which connect last authors to one another. Another way to think about the network construction is that each edge represents a shared publication between 2 individuals.

NetworkX also has a vast amount of built in functions for network analysis in addition to constructing networks. The four important analysis metrics that will be focused on for the remainder of the paper are: number of communities, degree centrality, clustering coefficient, and betweenness centrality. In addition to these metrics, we programmed our tool to also perform a variety of other analysis metrics: graph density, eigenvector centrality, number of cliques and largest clique size, node centrality, closeness centrality, checking if the graph is connected and the number of connected components.

To detect the number of the communities, the community API was implemented (Aynaud, 2010). This built in analysis tool from networkX computes the partition of the graph nodes detecting hubs or communities within a large network using Louvain heuristics. This is a greedy optimization method and works in two different steps. First, the methods looks for small communities by optimizing modularity, the degree to which a system's components can be

separated and recombined, locally. Second, it builds a new network with nodes that belong to the smaller community, these smaller communities are then counted and output an integer (Blondel, 2008). The integer represents the amount of central hubs within a network and this can be used to determine the graph type.

The degree centrality built in method was heavily used in this network analysis to measure the importance of nodes based on their number of connections (Hagberg, 2008). This function takes in a NetworkX graph and returns a dictionary that contains every node as a key and the degree centrality as its value. The value represented the fraction of nodes it is connected to. Due to the dictionaries being large in size, we abstracted the data to summarize every node by getting the average degree centrality for every queried term (Hagberg, 2008). To ensure an accurate degree centrality value and knowing a large sum of nodes had a 0 value for degree centrality, we pruned the data to exclude all values of 0 before taking the average. Another value extracted from the pruned dictionary was the median. Both the average and median allowed for better visualization and comparison on a larger scale, comparing network to network instead of node to node within one network.

The same abstract approach was used when dealing with the clustering coefficient metric. This as well, takes in a networkx graph and return a dictionary of each node with its clustering coefficient (Hagberg, 2008). The clustering coefficient is the measure of the extent to which one author has collaborated with another. This metric is based on triplets of connected nodes, and the fraction of possible triangles through that node that exist (Fairchild 2012).

Betweenness centrality, also known as bottleneck measure, was used within the case studies to find the most influential authors. In large complex networks, not all nodes are equivalent and this metric helps determine the nodes that are most influential in determining the flow of the network's information. Betweenness centrality is defined as the fraction of shortest paths going through any given node (Barthélemy, 2004). This measure is not so much based on connectivity, as seen below in Figure 9; node v is not very connected, but the effect of its removal would be detrimental due to its importance in connecting two parts of the neworks.

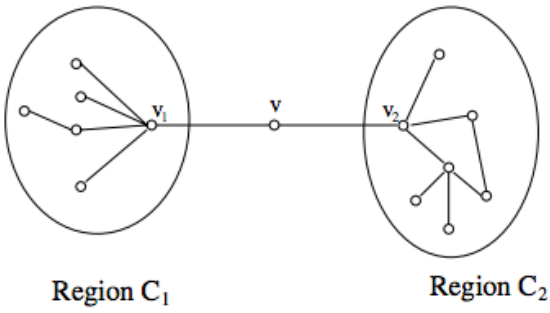

Figure 9: An example of betweenness centrality(Barthélemy, 2004). Node V represents the bottleneck of the example network, node v would have a high betweenness centrality because the amount of shortest paths that pass through the node to connect region c1 to region c2.

In networkX, betweenness centrality works identically to that described for degree centrality and the clustering coefficient. The function takes in the networkx graph and returns a dictionary, with every key being a last author's name and the value of their betweenness centrality (Hagberg, 2008).

3.2.5 Powerlaw

The powerlaw package that was imported in Python 3.6 has the ability to determine if our graphs distribution fits the power law distribution or another common distribution type. All of this is supported through statistical analysis. Previous bioinformatics papers including, "Statistical Analyses Support Power Law Distributions Found in Neuronal Avalanches" from the Plenz lab at the National Institute of Mental Health use this python package to identify the power law scaling within neuronal avalanches (Klaus, 2011).

The powerlaw package will support or refute a hypothesis claiming the graph type that each PubMed authorship map is displaying. Below in Figure 10, the definition of the power law distribution is shown in addition to several other common statistical distributions. This is the continuous distribution that the package is comparing with the formed graph data from auth. The powerlaw package functionality requires to pick two of the distributions shown in Figure 10 to compare with your graphs distribution. The two chosen for this project was power law along with log-normal (Clauset, 2009).

	name	distribution $p(x) = Cf(x)$ $f(x)$	C
continuous	power law	$x^{-\alpha}$	$(\alpha-1)x_{\min}^{\alpha-1}$
	power law with cutoff	$x^{-\alpha}\mathrm{e}^{-\lambda x}$	$\frac{\lambda^{1-\alpha}}{\Gamma(1-\alpha,\lambda x_{\min})}$
	exponential	$\mathrm{e}^{-\lambda x}$	$\lambda \mathrm{e}^{\lambda x_{\min}}$
	stretched exponential	$x^{\beta-1}\mathrm{e}^{-\lambda x^{\beta}}$	$\beta\lambda \mathrm{e}^{\lambda x_{\min}^{\beta}}$
	log-normal	$\frac{1}{x}\exp\left[-\frac{(\ln x-\mu)^2}{2\sigma^2}\right]$	$\sqrt{\frac{2}{\pi\sigma^2}}\left[\mathrm{erfc}\left(\frac{\ln x_{\min}-\mu}{\sqrt{2}\sigma}\right)\right]^{-1}$

Figure 10: Definitions of Common Statistical Distribution used in the Powerlaw Python Package (Clauset, 2009).

The power law model distribution used in the package follows the well known model Barabasi and Albert, which represents the graph type known as a scale free graph. The log-normal distribution in this package follows not specified model.

The Powerlaw package takes in your graph's distribution that is calculated using another Python package called collections. The two values that are outputted are an R value and a P value. R is the log-likelihood ratio, where a positive value for R denotes that the data is better fit by the first distribution. A negative value denotes that the data is better fitted by the second distribution specified. The size of the R value, positive or negative, determines the strength of comparison (Shaheen, 2017). The p value signifies the significance of the fitted model. For a model fitting to

be statistically significant this value must be less than or equal to 0.05. An example is shown below in Table 1.

Table 1: Example of Powerlaw Package Data Output Analysis

R (log-likelihood ratio)	p-value	First Distribution	Second Distribution
-1.398	0.0378	Log-Normal	Power Law
0.339	0.734	Log-Normal	Power Law

The first row of Table 1 has a negative R value meaning the graphs distribution is closest fit to the second distribution, power law, and it has a p-value of 0.0378 which is less than 0.05 meaning the fitting is statistically significant; we can reject the null hypothesis and conclude that the data are a good fit to the power law distribution. The second row in the table has a positive R value of 0.339 signifying that this graph is fitted closer to log-normal than power law but the p-value is much higher than 0.05 meaning that the fitting is not statistically significant. A scientist would conclude that the graph is probably better fit to another distribution.

3.3 Hypothesis Design & Testing

Using the powerlaw package, we were then able to test our hypotheses about the type of network acquired from our systematic tool. Through building our networks and analyzing them we were searching for answers to two research questions. Each research question is applied to every network. The first research question is: *Is the graph generated by our program a scale free graph that follows a power-law distribution?* As discussed in section 2.1.1, scale-free graphs follows a power-law node degree distribution, which is tested with the Python package described previously. Graphs that follow a scale-free distribution contain more hubs, or "communities", implying that in a case such as our network of authors, more collaboration is occurring. From this research question, our two testable hypotheses were developed. These are seen below.

H_0: The graph is not a scale-free graph
H_a: The graph is a scale-free graph

With the outputted R and p values we can accept one of these hypotheses and reject the other. If the p value is statistically significant, less than a value of 0.05, then the H_0 is rejected. Thus, we tried to support any field that does not have a significant p value with the NetworkX analysis metrics that were chosen to support scale-free networks. Looking for fields that demonstrate strong scale-free network characteristics (high degree centrality median & average, high clustering coefficient, and high number of communities) within the analysis metrics would allow us to help claim if a network is truly scale-free.

The other research question that was posed was about possibility of being a random network. This question asks: *Is the graph generated by our program a random graph that follows*

a log-normal distribution? This node degree distribution type of log-normal was the second distribution tested using the Power Law Python package. If the graph is random instead of scale-free, this implies that there is less collaboration occurring between authors in the network. The two hypotheses extracted from this research questions are presented below.

$\mathbf{H_0}$: The graph is not a random graph
$\mathbf{H_a}$: The graph is a random graph

We followed the same analysis tactic explained for the first research question to accept and reject one of the hypotheses above using both the R and p values as well as the NetworkX metrics. We are looking for a positive R value with a significant p value to be able to accept our H_a. Unless we are able to identify patterns within the computed metrics that demonstrate random networks characteristics. This would consist of few communities, low degree centrality, and low clustering coefficient. A combination of all of our results will be used to answer both research questions and accept a hypothesis that supports the network type exhibited for each field in both the Broad and Narrow Lists.

3.4 Usage of Tool

For the purpose of this project, the tool was used mainly as a method to carry out hypothesis testing using two lists: a Broad List and a Narrow List. Data was collected from each run of the program and some network statistics including clustering coefficient, degree distribution, and number of communities, were collected in order to compare the resultant networks.

3.4.1 Broad List

The "Broad List" of search terms contains the highest-level terms from PubMed's Biological Science Disciplines Mesh Terms (NCBI). For the purpose of this research, the list included ten terms: 'anatomy', 'biochemistry', 'biology', 'biophysics', 'biotechnology', 'chronobiology', 'neurosciences', 'pharmacology', 'physiology', and 'toxicology'. This list allowed us to explore author interactions within broad fields of approximately 1-5 million publications each.

We passed this list as input to the AuthorMap generator and through the systematic tool several quantitative metrics were produced based on the network generated from the PubMed queried results. To gain access to the PubMed database through our tool, PubMed's Entrez API was implemented. The systematic tools implements three of the Entrez functions: E-Search, E-Fetch and E-Summary. After all of these are executed in the program, an XML file is exported for every paper containing a variety of information including the desired last authors name. Next in the program, we built a dictionary to create a link between the last authors in the field (the keys) and their respective supporting authors (the values). For example, if the last author on one paper was a supporting author on another, this would connect their nodes. By pruning through every

paper's supporting authors looking for a common last author, we were able to build the network representing their relationships.

The first step to analyzing each network was discovering what type of network it was. We looked at each network in Cytoscape, and realized they didn't follow normal scale-free or random network patterns as expected. This was interesting and called for more research. Using the package Powerlaw in Python, we then tested several hypotheses to discover what types of networks we were producing. AuthorMap automatically generates and outputs the R and P values of the networks, indicating best fit to scale-free or random distributions.

We performed more analysis on each network to help justify our results from the Powerlaw package and support our stance on the hypothesis being tested. Using NetworkX built-in network analysis functions, we collected data that would support scale-free or random network behavior. The degree centrality average was calculated for each network, by averaging the value from every node but excluding all zero-values to allow for more accuracy. This way, we were able to compare each network's degree centrality value. Our program also calculated the median degree centrality as an alternative metric to average. Betweenness centrality was another measure of centrality calculated by our tool. We were able to use the betweenness centrality value to compare where and how many bottlenecks occurred in each network.

The clustering coefficient was also calculated for every node, and then we found the average clustering for each search query, which allowed us to perform comparisons across networks. The last metric calculated was the number of communities each network had. All the data collected for each network within the Broad List were exported into a CSV file which allowed for easy data analysis and graphing. Using the seaborn package within Python, the csv was read and then a multitude of graphs were created to allow for exploration of correlation between any of the metrics or any patterns seen within the field.

3.4.2 Narrow List

Similar to the Broad List, the "Narrow List" of search terms was also derived from PubMed's Biological Science Disciplines Mesh Terms. The terms in the Narrow List are one level lower in the hierarchical structure than those present in the Broad List, making them slightly more specific than the Broad List terms. This allowed us to explore slightly smaller fields of approximately 500,000-1,000,000 publications each. The Narrow List contained 64 terms, from 'anatomy-artistic', to 'toxicogenetics'.

Using the programmed AuthorMap tools that works systemically, the Narrow List was given as input to the tool and the same process as described in section 3.4.1 to the Broad List was performed. Data was extracted from PubMed, a network of last authors for all 64 specific fields was built, and network analysis was performed all systematically. A CSV file was exported and data analysis and graphing was done in Python using the seaborn package.

3.4.3 Case Studies

To further explore a few of the Narrow List fields and the results acquired previously, we selected three different fields based on some criteria. We selected two interesting fields that didn't follow the same trends in the network statistical analysis as the majority of the data: artistic anatomy and comparative physiology. The last field we selected did follow the normal trends and was picked based on interest in the field and its current popularity at WPI: neurobiology. With these three fields, the goal of these case studies are to look at non-metric descriptors (year, model organisms, h-index, etc.) to better understand the network structure as a whole. By studying the spread of knowledge from one last author to another, we can have a better idea of what forms this unique network type/shape.

We decided to focus on a variety of non-metric descriptors of particular nodes within each network. These descriptors include the year, the publishing journal impact factor, model organism worked with, and the authors h-index. The certain nodes were selected based on their betweenness centrality (or bottleneck value) that was gathered by using the AuthorMap tool designed earlier in the project. Betweenness centrality is a measure of the fraction of shortest paths going through any given node (Barthélemy, 2004). This value for each node shows the influence of that last author on flow of information through the entire network (more information in section 3.2.4). Four research questions were proposed to help guide our investigation on the interaction between all these variables:

1. If the author is more influential, has a large h-index, will the node likely have a large influence over the flow of information within our network?
2. In the spread of knowledge from one last author to another, does the bridging author influence their first degree connections with the type of model organism used?
3. Based on the bridging author, are their first degree connections within the network publishing in journals with comparable impact factors?
4. Are collaborations between last authors based on work completed in certain years or rise in popularity of the field?

These research questions will help shine light on the answers we desire of how the network gained its unique node distribution and shape. The process to collect all this information to answer these questions used a variety of tools including AuthorMap, Cytoscape, the PubMed website, and Scopus.

The first step was to determine the influential authors within the three fields of interest. By running our AuthorMap tool using the NetworkX betweenness centrality function, the tool exported a sorted list of dictionaries containing the last author name and their betweenness centrality. If the betweenness centrality is larger than zero, that author can be referred to as a bottleneck within the network. Their betweenness value reflects how much they are affecting the flow of information. A larger betweenness centrality pertains to a node with many shortest paths running through it, and a smaller betweenness centrality means the opposite. We picked the top

three authors that had the largest betweenness centrality and also scrolled to the bottom of the sorted list and picked the author with the smallest betweenness centrality for comparative purposes. Having the smallest bottleneck allowed us to contrast the three largest bottlenecks studied.

All four bottlenecks were being studied within the three fields of interest and all steps at this point are repeated for every bottleneck in each field. All the bottlenecks were located within the network using the Cytoscape visualization tool. By searching for the last authors name in the search bar, the last author was highlighted along with it's first degree connections. All of the first degree connection names were recorded in a tabular format to keep records organized and allow for easy comparison between the bottleneck author and all their connections.

With all the authors names of interest, we can now incorporate the PubMed website and Scopus. Through these two sources, we can gather the non-metric descriptors for both the bottlenecks author and all of their first degree connections. To begin this big data collection, an advanced search was performed in PubMed specifying the field and the last authors name. The search results returned are organized by the "Best Match" papers (which is the search our tool performs using the API). An example of a search query and the result can be seen below in the top of Figure 11.

Figure 11: An example PubMed query and search result. An example of one publication for the author in the field (top). An example if the author has more than more publication (bottom).

If a last author only has one publication within the queried field it will immediately pop up. For some authors, they will have multiple papers within the field and a list of papers will be returned to you. In this case, the first paper that is sorted by "best match" is extracted for our records. For example shown in the bottom of Figure 11 above, notice that "Best Match" is selected for the sort by technique and the first paper titled "Glia Maturation factor.." would be the paper in which we would extract information from. From PubMed the non-metric descriptors that are recorded are the year published and the model organism worked with from the title/abstract. This information can be recorded into the table corresponding to the author you are researching. The title is copied and then pasted into the document search on the Scopus website, specifying the search by article title using the drop down, to finish gathering the information. The paper is most likely to be the only search result retrieved unless it has a general title, in this case you would then scan the search results looking for the paper with the correct last author.

From the search results in Scopus, one can click upon the source hyperlinked to the right of the title (Figure 12, top) and this will bring you to another page revealing the journal impact

factor. Record this information in the same table with the year and model organism. Using the back arrow to return to the search results, then click on the last author (Figure 12, bottom) and this will bring you to the personal authors page in Scopus. On the author's personal page, the h-index is shown and can be also be recorded in the table.

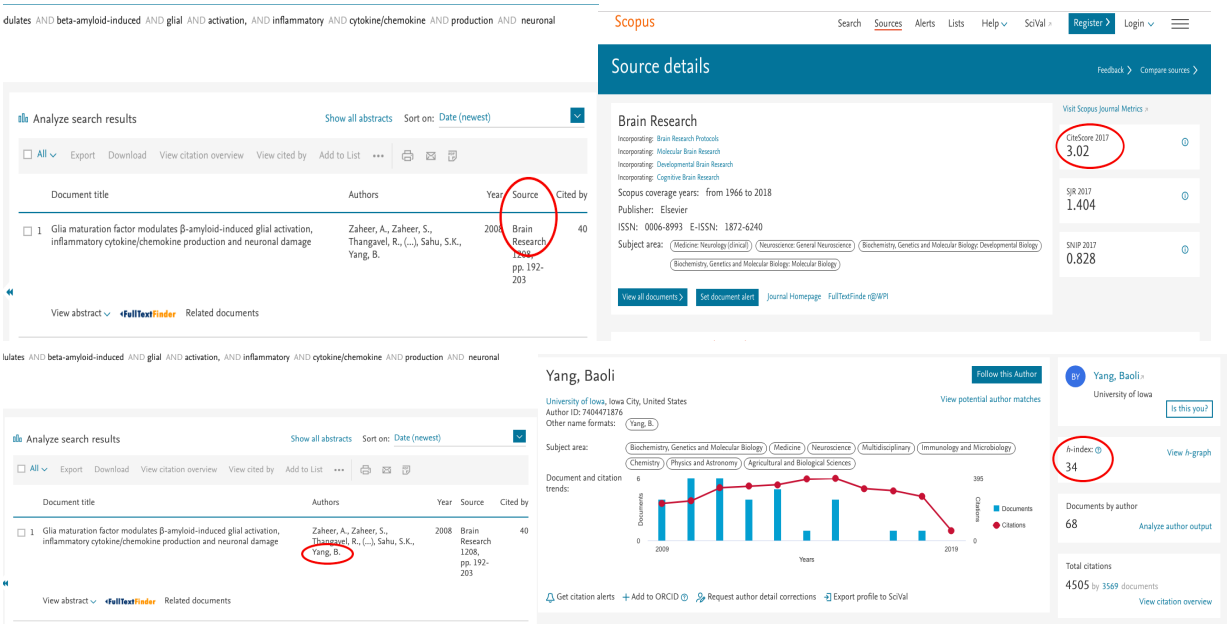

Figure 12: Showing Scopus search results and hyperlinks to click on to retrieve information. Accessing the specific journals impact factor (top). Accessing the last authors specific page (bottom).

Every table should represent each identified bottleneck author for all three fields, including the bottleneck and their first degree connections and all the information collected. Each field has four master tables. With all this information pooled together, we can then compare the data and start to answer our research questions.

Based on the data collected and some interesting findings, we further looked into the correlation between the numerical values of the non-metric descriptors. For example, if there is any correlation between h-index and betweenness centrality, journal impact value and betweenness centrality, or h-index and betweenness centrality. Since we did not have a large enough sample size from the case studies alone, thirty random last authors were picked along the distribution of betweenness centralities from the entire network. Each author was then searched in PubMed and Scopus retrieving the h-index and journal impact factor. This data was then collected and stored within a CSV file and then graphed using the ggplot2 library in R.

4 Results

The data that was extracted from the designed authorMap Python tool will be discussed within this section. All results from the Broad List, Narrow List, and the chosen case studies are presented and analyzed in detail. Initially, a great amount of curiosity led to testing a variety of hypotheses to discover what type of network was being built with every PubMed query. It was discovered that all networks: Broad List (10) and Narrow List (64), all had node distributions that best fit a power law distribution pertaining to a scale-free network. When performing network analysis, some of the metrics including clustering coefficient, degree centrality supported this hypothesis and others did not. The three case studies: artistic anatomy, neurobiology, and comparative physiology were chosen due to interesting results affecting our hypothesis stance. Each case study was investigated to discover more details about the field outside of the network type including the most common year & journal name, organisms studied, and the bottlenecks within the given field.

4.1 Broad List

The Broad List consisting of 10 terms with very large field sizes were queried to gather a portion (maximum of 500,000) of the most relevant papers to analyze to construct a social network. The network was built using last authors of every paper extracted from the search, and drawing connections to other last authors in the field. Many different metrics were tested when analyzing the network and these all can be seen below in Figure 13 & 15. The raw data that the graphs reflect that were collected from AuthorMap can be seen in Supplemental Figure 1.

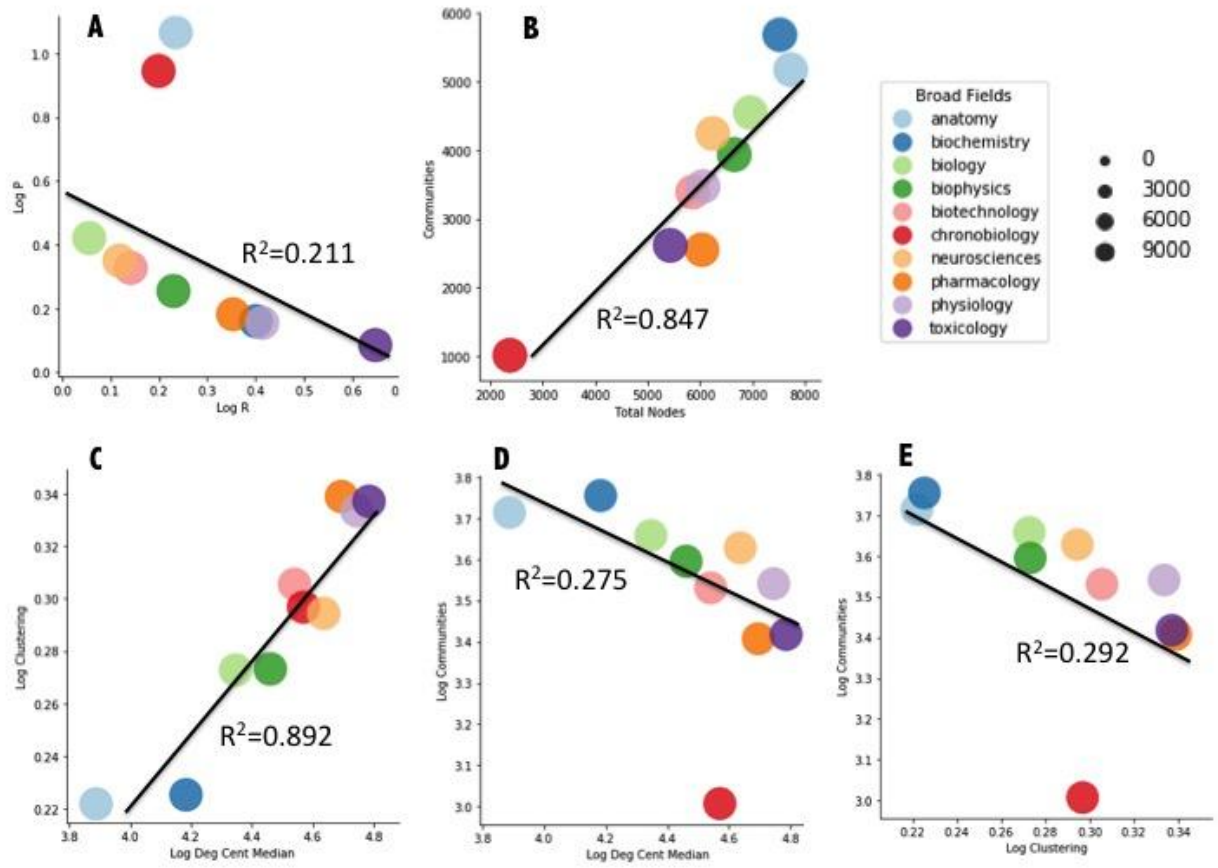

Figure 13: Broad List Results: each dot colored to represent field, and dot sized by total nodes within the created network(number of last authors). **A.** Log R versus log P values **B.** Total nodes versus communities **C.** Log degree centrality median versus log clustering coefficient **D.** Log degree centrality median versus log communities **E.** Log clustering coefficient versus log communities

In Figure 13 above, a variety of graphs are presented to display all data extracted from the 10 different networks constructed using our AuthorMap tool. Figure 13A shows the results from the power-law package that we used to test our network against model networks demonstrating scale-free and random distribution. The X axis is showing R, maximum likelihood values, which from the raw data were absolute valued and then calculated logarithm with base 10. The Y axis is the p value which shows if the fitting is statistically significant; an acceptable p value is ~0.01. All Broad List terms were within a range of negative R values before taking an absolute value, suggesting that they are all better fit to the power law distribution. None of the ten graphs had a small enough p-value for this fitting to scale-free to be accepted and statistically significant. The two outlier dots; anatomy (light blue) and chronobiology (light blue), had the best fit (most negative R value) to scale free and had the smallest two p values closest to being statistically significant with values of 0.08 and 0.1 respectively. Toxicology shown as the purple dot had the

least negative R value and had the highest p value, supporting a not well fit to the power-law distribution but it was a slightly better fit than log-normal, indicating that it behaved more as a scale-free graph than a random graph. No exact pattern can be seen from Figure 13A besides a similarity between a majority of the fields with two outliers.

To better understand how communities are located using the Louvain algorithm in NetworkX, we were curious if it has a correlation with the network size. Figure 13B shows a direct correlation between the two metrics as predicted. The larger the network size, the more communities are possible and are seen. The largest field within the Broad List is anatomy which is shown in light blue and is located in the upper right hand corner proving that it has the most communities. This pattern stay true throughout the graph; the second largest field is biochemistry which has the second highest number of communities as seen in the figure. Lastly, the smallest field within this list is chronobiology which is located in the lower left hand corner because it also has the smallest amount of communities. This was taken into account when moving forward with the Narrow List.

Figure 13C, shows the pairwise relationship between the two metrics degree centrality and clustering coefficient. Within the figure, a lot can be noted about the pattern displayed. Exponential growth can be seen as well as clustering of similar fields. Biology and biophysics can be seen clustering (light green and green) as well as chronobiology, biotechnology and neuroscience (light red, red, and orange). Values that are more to the right of the graph show a lower median degree centrality and values that are more to the top show a lower clustering coefficient. For example, toxicology is located in the upper right hand corner of the graph portraying the smallest clustering coefficient as well as the lowest median for degree centrality out of the ten Broad List terms. This makes sense due to this network was the least fit to scale- free distribution, thus would have a low values for both metrics because degree centrality and clustering coefficient pertain to a scale-free network. In this graph, there is one outlier which is consistent with the others: anatomy shown in light blue. Anatomy has the lowest log value for degree centrality and clustering coefficient, which means it has the highest true value. Both of these support the R and p values of the artistic anatomy network being scale-free.

The graph within the bottom row located in the middle, Figure 13D, is studying a relationship between the degree centrality median and the number of communities. The field with the highest median and largest number of communities is the most scale free network by the powerlaw package- once again, anatomy. Chronobiology who also was close to being statistically significant scale-free network has a lower median and few communities due to the network size. We expected for chronobiology to have a similar degree centrality median due to the R values being close in value. No other pattern can be induced from this graph, the fields that have a high number of communities will be closer to the top of the graph, and the fields with a high median will be to the left of graph. Most fields are represented in Figure 13D in the top half and to the right half. This shows that most of the terms from the Broad List have a high number of the communities, which makes sense because of the size of the field, and a low degree centrality median.

Degree centrality and number of communities have no notable relationship, but the number of communities with the clustering coefficient does. Understanding that to have a community, more clustering would have to occur in the network. Thus having more communities in a network would cause for a higher clustering coefficient. There are three outliers, two that strongly follow that statement and one that does not. Anatomy and biochemistry demonstrate very high clustering coefficients and also are the two fields with the largest number of communities. Chronobiology, the smallest field within the broadlist returning the least amount of last authors for the network also has a similar clustering coefficient to fields with thousands of more clusters. Other than these three, the rest of the seven fields in the list all have similar community sizes and clustering coefficients and are grouped in the central area of the graph.

Scatterplot matrices are helpful to better study pairwise relationship between all metrics, and if there is any relationship what is the nature of it. Also, this type of visualization can allow identification of outliers and clusters within the raw dataset. Every combination of every metric is shown with a scatterplot as well as a histogram on the diagonal of the matrix. For example, in Figure 14 below the histogram in the total nodes column, is showing the distribution of total nodes values of every field within the Broad List. The diagonal histogram shows the distribution of the variable in the corresponding column.

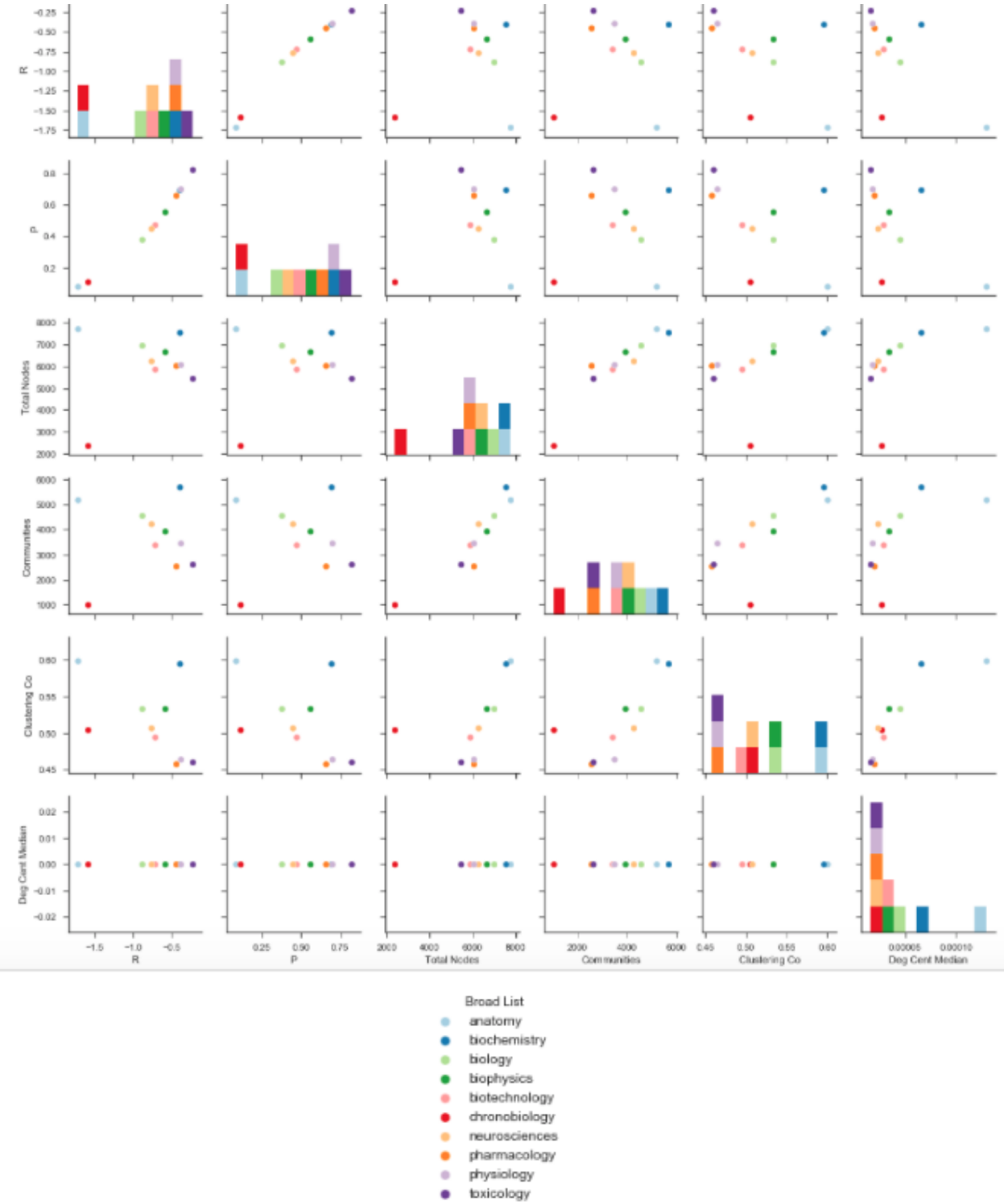

Figure 14: Broad List results presented within a scatterplot matrix to look for relationships between all computed metrics of each field, each dot represents one of the broad fields seen in the legend. Correlation between a variety of metrics is noted: total nodes & communities, R and P. The same outliers can be noted in each scatterplot.

Above in Figure 14, a scatterplot matrix was constructed for all metrics on the Broad List terms, each graph has different axes with raw values from our original dataset. As detected earlier in the individual graphs shown in Figure 13, some pairwise relationships were noted that are also detected in this visualization. In the matrix, a relationship between total node and number of communities as well as R and P is seen. In almost every dataset, there are two outliers the red dot and the light blue dot: chronobiology and anatomy respectfully. Excluding these outliers, the other eight data points seem to cluster together in no particular pattern within every graph combination. This is seen within the histograms for each metric, in most cases, light blue and red are not stacked with others and stand alone whereas the other terms mostly stack together or next to each other. The R and P histogram helps support our hypothesis, making it easier to see the data. You can see that the most negative R values are anatomy and chronobiology as well as the smallest p values. These two fields are stacked within the histogram and outliers when looking at the other fields. The scatterplot matrix helps summarize all the Broad List data collection/analysis and allows for easy comparison between metrics but does not allow for specificity of numbers as it is hard to calculate a scale with so many scatterplots.

Due to the large size of these fields it is impossible to claim whether these network statistics are true for the entire broad field or just the portion that we sampled. We sampled a max of 500,000 papers extracting from an organized list that sorts all of the millions of papers from most relevant to least. To better test our hypothesis of network types we dove into the next hierarchy of the PubMed MeSH terms. In the next section of the results chapter, we discuss the same data collection and analysis technique but on the 64 fields that are smaller and more specific.

4.2 Narrow List

To attempt at sampling a whole field with the AuthorMap tool, we ran the systematic tool on the Narrow List. The Narrow List consisted of 64 terms that were more specific and had less papers to query on PubMed. Below in Figure 15, all network analysis is shown in the form of scatter plots. Each dot represents a Narrow List term but is colored by the Broad List term that sits above it in the MeSH term hierarchy. The dots are also sized based on the total nodes within the individual network. The colors are shifted in these scatter plots compared to the colors in Figure 13 above in the Broad List results. The raw data that the graphs reflect below that were collected from AuthorMap can be seen in Supplemental Figure 2.

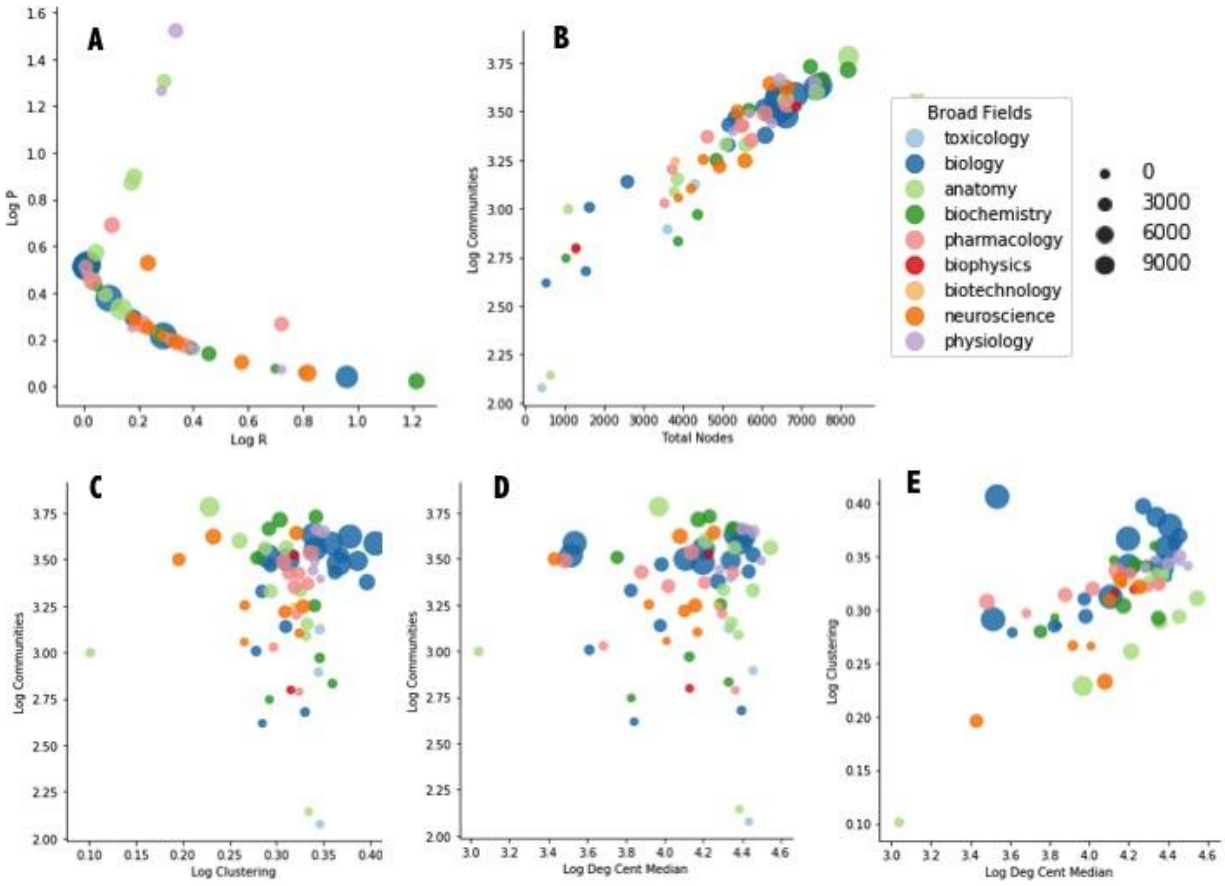

Figure 15: Narrow List Results: colored by hierarchical structure of term under the broad term and dot sized by total nodes within the created network (number of last authors). **A.** Log R versus log P values **B.** Total nodes versus communities **C.** Log degree centrality median versus log clustering coefficient **D.** Log degree centrality median versus log communities **E.** Log clustering coefficient versus log communities

The first graph on the top left shows the log R versus log P values. To create this graph, we took the absolute value of the computed R values, and then took the logarithm of those positive values. In this panel, fields in the upper left hand corner of the graph have the most negative R values and the most statistically significant P value, meaning that they are the most likely to be scale free. The two purple data points in this corner are from the physiology field. They are comparative physiology which has an R value of -2.169 and a P value of 0.030, and psychophysiology which has an R value of -1.924 and a P value = 0.054. This is interesting because in the Broad List graphs, physiology did not appear to be scale free, but here, these data points definitely do. The other green data point in this top left corner is from anatomy, which is consistent with the data we have seen thus far about anatomy consistently being an outlier. On the other end of the graph (with a very low R value and very high P value) is a Narrow List item from the

biochemistry field. This implies that this field is closer to a random graph than a scale-free, and is consistent with what was observed earlier in the Broad List data for biochemistry.

The second graph, on the top right, shows the number of total nodes in a network versus the logarithm of the number of communities in that network. There was also an additional aspect added to the graph to show additional information - data point diameter is also representative of total field size. Through this, it is also clear that the largest field also has the largest number of communities, which was expected. The largest field shown in this panel is comparative anatomy, which is consistent with the data seen in the analysis of the Broad List. Comparative anatomy also has the highest total number of communities. Oppositely, there was one toxicology field and one anatomy field that each had very low numbers of communities.

The third graph, on the bottom left, shows the logarithm of the node degree clustering versus the logarithm of the total number of communities. Here there is a clear trend in the clustering coefficient data, where most clustering coefficients fall in the range of 0.4-0.5, but there is one clear outlier, which is anatomy artistic at 0.79. This data point can be seen as a small light green point extremely close to the y axis - this means that it isn't a large field, and doesn't have many communities, but it does have a high clustering coefficient, making it an outlier. Larger fields shown in this graph, such as narrow terms that fall under biology, have a higher number of communities but lower clustering coefficients. Connecting back to a pattern that was observed in the previous graph, the toxicology and anatomy fields with low total communities can be seen as the two data points closest to the x axis. All other fields that follow the normal trend are seen clustered in the upper right hand corner.

The graph in the middle of the bottom row shows the logarithm of the median degree centrality versus the logarithm of the total number of communities. Again, artistic anatomy can be seen as an outlier here, shown as the light green dot next to the y axis with the greatest median degree centrality. All of the remaining points are clustered in the upper right hand side of the graph, indicating greater average number of communities and lower average degree centrality. To the left of this large cluster, there is a small, very densely packed cluster of four data points: two from biology, one pharmacology, and one neuroscience. This is interesting because it includes several of the biggest fields in the network.

Finally, the fifth graph shows the logarithm of the median node degree centrality versus the logarithm of the clustering coefficient. Here, there is some clear clustering occurring based on groups of broad fields. The biology (blue) is clustered at the highest in the uppermost top right corner. Under that cluster are anatomy (light green), biochemistry (dark green), and physiology (lilac) which are all also tightly clustered below the biology cluster, which means that they have better clustering coefficient. One outlier is present on the lower left hand corner - anatomy artistic - with high clustering coefficient and high median degree centrality. This was not surprising as we have seen anatomy artistic act continuously as an outlier throughout these results.

A scatter plot matrix was then used to summarize all the data collected for the Narrow List networks. In Figure 16 below, the scatter plot matrix for the Narrow List terms reveals a lot of interesting patterns within the data including the distribution of values for each metric. The same

two pairwise relationships from the Broad List data are noted here as well: total nodes and number of communities as well as R and P values. In mostly every scatter plot there is clustering of every field with minimal outliers. The consistent outlier throughout the entire matrix is one light green dot pertaining to a sub field of anatomy. In some graphs a purple dot, a sub field of physiology, serves as an outlier in all graphs with R on the Y axis. This is due to the large negative R value the field has but it is fairly consistent with all other values in the other metrics. Lastly, in median and average degree centrality versus clustering coefficient a dark orange and blue dot are outliers along with the light green dot. The orange and blue dots pertain to the broad fields of neuroscience and biology respectfully.

Figure 16: Narrow List results presented within a scatterplot matrix to look for relationships between all computed metrics of each field. Each field is colored by their categorized Broad List field, the legend is shown above. The same correlation that was seen in the Broad List scatterplot matrix is supported here.

Studying the diagonal histograms, there seems to be uniform and similar distributions for all metrics besides total nodes and the number of communities. These two values vary between all Narrow List terms grouped under the Broad List. For example, the broad field of anatomy has 10 narrow fields under the hierarchy. One of these fields is a tiny network with only 1,097 last author (anatomy artistic), whereas, another field is one of the largest networks with 8,210 last authors (anatomy comparative). This leads for an unequal distribution of values within the anatomy Broad List group. Due to this unequal distribution in the total nodes metric, this pattern will be carried to the number of communities distribution as these have a pairwise relationship. Overall in the data visualization summary, fields of interest from Figure 16 can easily be followed and conclusions can be drawn.

4.3 Conclusion

In conclusion, these results figures have shown that overall, there are clear trends in each plot that most of the fields follow. However, there have been several fields that are consistent outliers, such as anatomy artistic. There have also been several other fields that have not performed as expected. For example, the graphs that are statistically the most scale free according to R and P values are from the physiology field, but none of them were outliers in terms of median degree centrality or clustering coefficient. We expected that the graphs that would statistically be the most scale free would also have a high degree centrality and high clustering coefficient, as these are trademark characteristics of a scale free network. Additionally, the anatomy field was supported to be scale free via its degree centrality, number of communities, and clustering coefficient - however, it does not have a statistically significant P value for its R value. Because these fields performed so unexpectedly, we will be investigating them further in additional case studies below.

4.4 Case Studies

This section reflects the deeper findings within particular networks using the AuthorMap tool in conjunction with other sources. By involving non-metric descriptors and moving away from network analysis, we were able to study the transfer of knowledge from one last author to another as well as study any correlation between outside measures about the authors. Our research questions demonstrate the curiosity to better understand how the network was formed and if there are any patterns that can characterize the collaboration between two authors in a particular field. By studying each bottleneck we will answer the four research questions:

1. If the author is more influential, has a large h-index, will the node likely have a large influence over the flow of information within our network?
2. In the spread of knowledge from one last author to another, does the bridging author influence their first degree connections with the type of model organism used?
3. Based on the bridging author, are their first degree connections within the network publishing in journals with comparable impact factors?
4. Are collaborations between last authors based on work completed in certain years or rise in popularity of the field?

4.4.1 Case Study 1: Artistic Anatomy

Artistic anatomy is a small field that focuses on studying anatomy for artistic purposes represented by drawing, paintings, or sculptures. Most published papers within this field are atlases of an organism's anatomy. This field was interesting and selected as a case study because it demonstrates expected scale-free characteristics but is not a statistically signifcant scale-free network with a p-value of 0.59. All of the raw network analysis data collected by authorMap can be seen below in Table 2. A complete table of raw statistics for all 64 fields to use for comparison can be found in Supplemental Figure 2. Artistic Anatomy’s network data when compared to the other narrow fields was always an outlier with the highest degree centrality median and clustering coefficient which is expected in a scale-free network.

Table 2: Network Analysis Data for Artistic Anatomy Network

R	P	Total Nodes	Communities	Degree Cent. Avg.	Clustering Co	Degree Cent. Median
-0.537	0.591	1097	995	0.001	0.792	0.001

To further explore this unique network type that better fits a node distribution of scale-free although not significant, looking at the most and least influential authors and their connections will allow for us to better understand the structure and how the network was formed. Below in Figure 17 is the full artistic anatomy network, with the three largest bottlenecks highlighted in yellow.

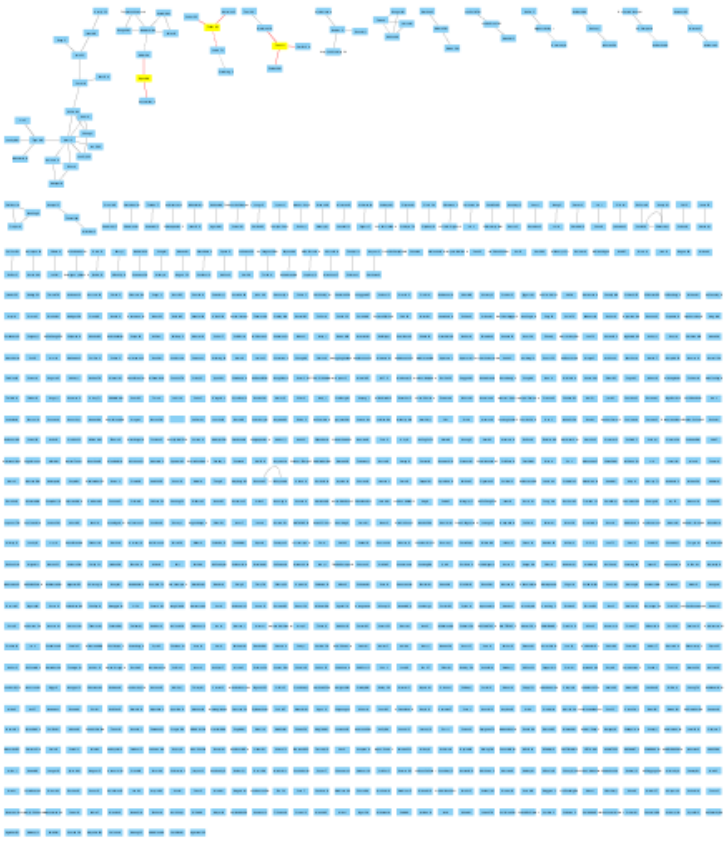

Figure 17: Structure of Artistic Anatomy Full Network with the Three Largest Bottlenecks Highlighted in Yellow

This figure is to be used for comparative purposes and zoomed in images of the bottlenecks and their connections can be found in the supplemental figures. The network is small with little connectivity, and consists of many authors that have never collaborated with another last author in the field, so the bottlenecks that are present seem minor to other bottlenecks in other fields but major within this specific network. The four bottlenecks, and their betweenness centrality further investigated in artistic anatomy are shown below in Table 3.

Table 3: Betweenness Centrality values for the four bottlenecks studied in this section

Name of Last Author	Betweenness Centrality
Bryan, R. Nick	9.926
Tomic, Irina	8.272
Tubbs, Richard Shane	8.272
Fischl, Bruce R.	0.0001

4.4.1.1 Largest Bottleneck: Bryan, R. Nick

The largest bottleneck in the artistic anatomy network is Bryan, RN with a betweenness centrality value of 9.93. In Supplemental Figure 3, a zoomed in screenshot of the original artistic anatomy network, one can see the last author and all of his first degree connections highlighted in Cytoscape. This last author has two first degree connections; Bryan RN has collaborated with Miller, GA as well as Davatzikos, C. The table below lists the bottleneck author first in italics and then his two connections. Each column represents a different non-metric descriptor: year, h-index, journal impact factor, and model organism.

Table 4: AA Bottleneck 1: Bryan, RN. & Connections Non-Metric Descriptors. Preservation of organism can be noted from this bottleneck.

Name	Year	H-index	Journal Impact Factor	Organism
Bryan, RN	1996	62	1.29	Humans
Miller, GA	1997	2	0.75	Humans
Davatzikos, C	2016	72	6.15	Humans

As we look to answer our research questions for the field of artistic anatomy, there is no pattern seen in studying this bottleneck alone. Some observations that are made are: model organism stays consistent across all three connections, journal impact factor and h-index vary tremendously from last author to last author, and the year stayed consistent with one connection but not the other. Another noted observation was although Bryan, RN has such a high h-index he did not publish his work in a journal with a high impact factor which contrasts with Davatzikos, C. who has a high h-index as well as published in a popular journal. The bottleneck author also has a high h-index which may be necessary to have a high betweenness centrality and be considered influential enough to affect the flow of the network information.

4.4.1.2 2nd Largest Bottleneck: Tomic, Irina

Studying the top three bottlenecks allows us to see if there are any trends that make an author influential as well as if there is any spread of knowledge from the bottleneck author to their connections in a variety of situations. The second bottleneck author, Irina Tomic from Serbia has three first degree connections within the artistic anatomy network. She and her connections can be visually seen in Supplemental Figure 4. The table below reflects the names of her connections in addition to all the non-metric descriptors.

Table 5: AA Bottleneck 2: Tomic, I. & Connections Non-Metric Descriptors. Preservation of organisms, journal impact factor, and h-index can be noted.

Name	Year	H-index	Journal Impact Factor	Organism
Tomic, I	2017	2	1.947	Humans
Djordjevic, D	2010	2	1.57	Humans
Starcevic A	2014	4	1.57	Humans
Cetkovic, M	2012	6	1.15	Humans

Irina Tomic having the second largest betweenness centrality with a value of 8.271 lacks a large h-index. With an h-index of only 2, it seems that to be considered a bottleneck within a network h-index is not always a factor. Within Table 5, we can see a spread of knowledge by use of the same model organism as well with consistency demonstrated in year, h-index, and the journal impact factor across all the last authors. All authors worked with humans, published in a journal with an impact factor from 1.0-2.0 from the year 2012-2017, and have an h-index within a range of four. These connection form a cluster within the artistic anatomy network, seen in Supplemental Figure 4, which makes sense as they are so similar in a variety of characteristics that could cause them to cluster with Tomic, I serving as the central pivot connecting them all.

4.4.1.3 3rd Largest Bottleneck: Tubbs, Richard Shane

With a betweenness centrality value identical to the last (8.272), Richard Tubbs is the third bottleneck to be investigated in the field of artistic anatomy along with his primary connections. He has three connections that can be seen in Supplemental Figure 5 and below in Table 6.

Table 6: AA Bottleneck 3: Tubbs, RS. & Connections Non-Metric Descriptors. Consistency in the model organisms can be seen across all four authors.

Name	Year	H-index	Journal Impact Factor	Organism
Tubbs, RS	2017	37	1.24	Humans
Salter, EG	2006	18	1.91	Humans
Lewis, TL	2014	10	12.41	Humans
Wartman, C	2007	3	2.40	Humans

The bottleneck author, Richard Tubbs, has a high h-index value but again doesn't necessarily have a large journal impact factor. One of his connection Lewis, TL who has a lower

h-index published in a better journal with a higher journal impact factor of 12.41. Looking at his connections no patterns can truly can extracted besides that all four last-authors worked with the same model organism: humans. Consistency lacks looking at the other non-metric descriptors.

4.4.1.4 Smallest Bottleneck: Fischl, Brue R.

To contrast any trends seen in the larger bottlenecks looking at the smallest bottleneck was important. Bruce Fischl is the author with the smallest bottleneck, not including authors that have no betweenness centrality (not connected to anyone or only one other). Bruce had a betweenness centrality of 0.0001. This last author from Harvard Medical School also had three first degree connections that can be seen in Supplemental Figure 6 or listed below in Table 7.

Table 7: AA Bottleneck 4: Fischl,BR. & Connections Non-Metric Descriptors. Contrasted with the large bottlenecks, the h-indexes in the smallest bottleneck are the largest noted throughout the entire case study thus no correlation between betweenness centrality and h-index.

Name	Year	H-index	Journal Impact Factor	Organism
Fischl, B	2007	89	6.13	Humans
Lein, ES	2017	1	3.30	Humans
Boas, DA	2012	88	3.30	Variety
Miller, MI	2010	70	6.75	Humans

Fischl being the smallest bottleneck has the largest h-index we have seen thus far in our data collection. This example further suggests that the betweenness centrality may not be based on the h-index of the author. Two out of his three connections also have very high h-indexes but the other only has an h-index of one. Although that author, Lein, has a small h-index, the work was still published in a journal with a high impact journal. Fischl, with a large h-index, also published in a journal with a high impact factor and so did his connections. The consistency in journal impact factors is noted. No similarity in publishing year was found but all three of his connections did work with humans and one expanded and worked with a variety of organisms that included humans. No large differences other than the h-index value of Fischl were discovered in this attempt to contrast the largest and the smallest bottleneck.

4.4.1.5 Studying Correlation

When studying the large and small bottlenecks, several instances suggested a lack of correlation between betweenness centrality and other metrics. To determine whether this is a significant finding, a larger sample of nodes with different betweenness centralities were chosen across the distribution to have a total sample size of 30 bottlenecks. Throughout looking at the top three bottlenecks and the smallest, questions about if the h-index measure is directly related to the betweenness centrality within the network arose. Below in Figure 18, the graph shows there is no correlation seen between the h-index of an author and their effect on the flow of information in a network based on the thirty bottlenecks sampled. Each dot represents an author which is sized based on journal impact factor and colored based on the percent of work within the given field; this parameter was extracted from Scopus.

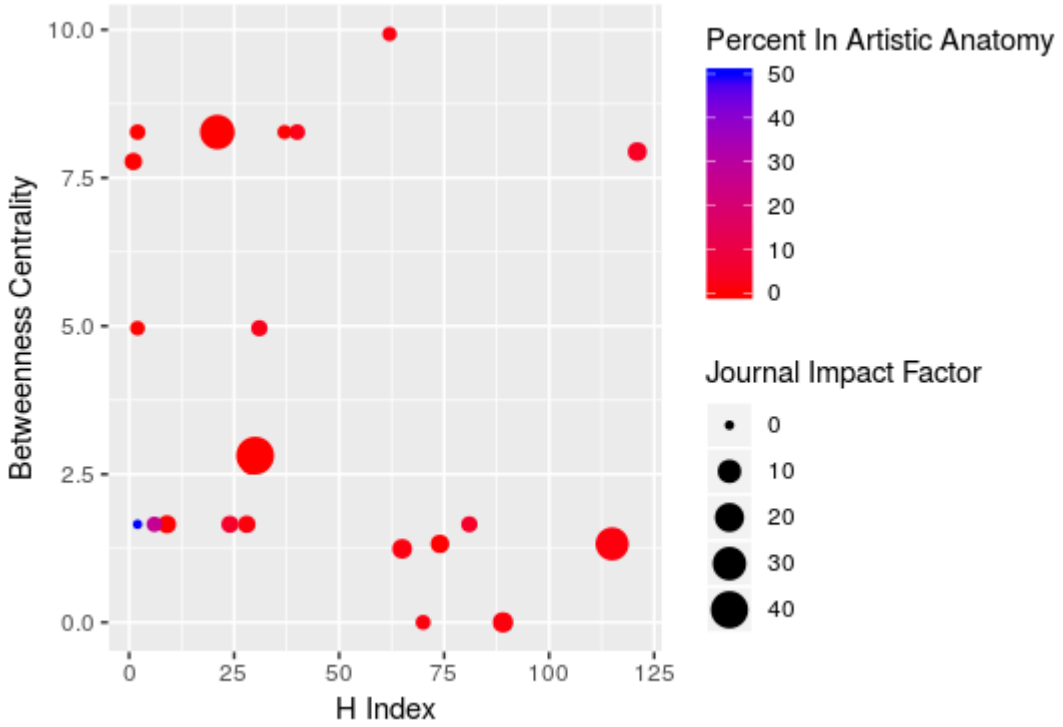

Figure 18: Correlation of H-index and Betweenness Centrality in Artistic Anatomy. No correlation can be extracted.

A reason why there is no correlation between h-index and betweenness centrality, is that h-index does not pertain to the importance of the author within that particular field but in general. For example, the smallest bottleneck author in the artistic anatomy field, Bruce Fischl, has an h-index of 89, but has only one paper within the artistic anatomy community; his high h-index is largely due to his extensive work within the neuroscience community. This can be seen in Figure 18 by the coloring of percent within the field and how most dots are at the lower end of the gradient showing a smaller percent of their work is done in this community. There are only two more blue circles in the bottom left and both have a low h-index and betweenness centrality.

To further see if there was any other correlation among the metrics collected for the thirty bottlenecks a Pearson Correlation test was performed in R studio. Below in Figure 19, there is a heat map showing if there is negative or positive correlation between all five metrics. The correlation value can be read through the color of the box as well as the actual correlation value is recorded in each box. The more red the box is the more negative the correlation between the two metrics, the more blue the box is the more positive the correlation is. Purple serves as a middle color.

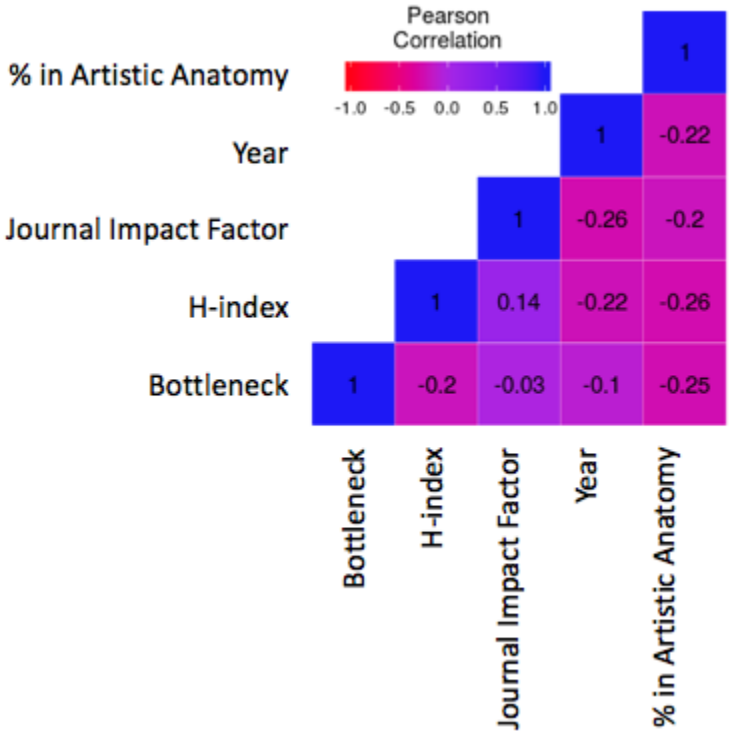

Figure 19: Heatmap showing Pearson Correlation between all non-metric descriptors using the same data from the thirty bottlenecks. Only positive correlation is between the journal impact factor and h-index. No correlations were statistically significant.

The Pearson Correlation provided no support that there is a positive correlation between the h-index and the bottleneck with a negative correlation value of -0.2. The only two metrics with positive correlation was between the journal impact factor and h-index, with a value of 0.14 (very small). The p-values for each correlation measure can be seen below in Table 8. There was no statistically significant values found ($p \leq 0.01$). The one positive correlation found was not significant with a p value of 0.509.

Table 8: Significance of Pearson Correlation between non-metric descriptor. No significant correlations within the non-metric descriptors in the field of artistic anatomy.

	Bottleneck	H-index	Journal Impact Factor	Year	% in AA
Bottleneck	NA	0.370	0.883	0.661	0.254
H-index	0.370	NA	0.509	0.324	0.224
Journal Impact Factor	0.883	0.509	NA	0.235	0.348
Year	0.661	0.324	0.235	NA	0.322
% in AA	0.254	0.223	0.348	0.322	NA

Overall, artistic anatomy had no statistically significant correlation found among comparing the five different non-metric descriptors. Contrary to previous beliefs, this correlation study helped support that there is no correlation between h-index and betweenness centrality. The only positive correlation noted, with a very small correlation value of 0.14, was between h-index and journal impact factor. Will any patterns be seen across all three case studies in correlation?

4.4.1.6 Summary

With a mission to answer the four research questions for the field, a lot of information was collected about the bottleneck authors and their first degree connections within the network made by the tool AuthorMap. The first research question posed was the studying if there is any correlation or consistency between the h-index and the betweenness centrality. Through all four authors studied, it was rare to see authors only collaborating with other last authors with a similar h-index. In a separate study, there was no correlation found between the two values of h-index and betweenness centrality. An author can have a very high h-index but a low betweenness centrality or vice versa.

The same approach was used to study the journal impact factor. Do connecting authors publish in similar journals with comparable impact factors or can this not be characterized by collaboration? By looking at our four authors and their connections, in only one case (Tomic, I) all four authors published in a journal with an impact factor in the range of 1.0-2.0. This pattern was not seen in the other cases to this extent, some had mostly similar impact factors and then an outlier. There are many journals that a scientist can publish in today and this may play a role. In the separate study, no correlation between betweenness centrality and the journal impact factor could be seen. The bigger the bottleneck is not directly influenced by the journal impact factor or the authors h-index.

There were some patterns noted with the spread of knowledge pertaining to model organisms. Mostly all authors that connected worked with the same model organism but it may

also be that humans makes up the largest model organism population for this field. Below in Table 9 represents the top 5 model organisms declared by PubMed overall and their showing in artistic anatomy. Due to the lack of large populations of other model organisms, this pattern noted for this field may have been a fluke.

Table 9: Model Organisms Populations within Artistic Anatomy Field

C.elegans	Zebrafish	Mice	Drosophila	Yeast	Humans
0	2	43	4	0	1084

Lastly, the years of publication were studied within artistic anatomy. Some consistency was noted that authors were publishing usually with the same decade of one another. To study if this is influenced by a rise in popularity within the field, data was extracted from PubMed to give us the total number of publications within the field within each year. Below in Figure 20 you can see a bar chart representing the trend of publications from 1960-2019.

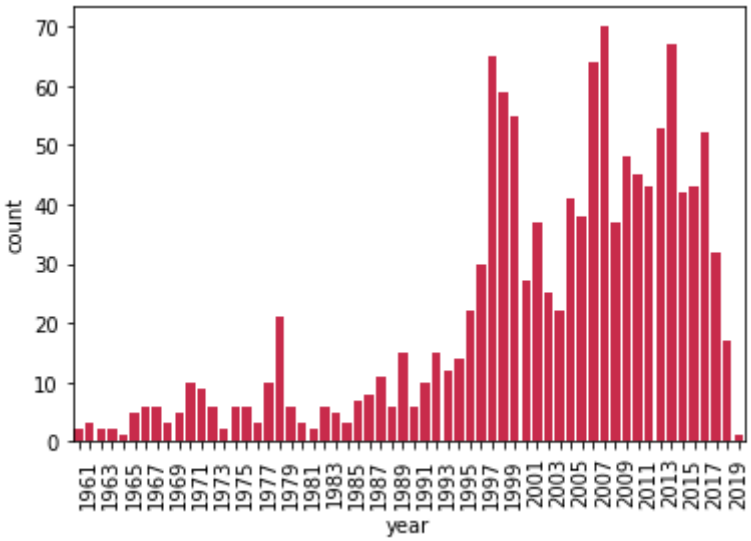

Figure 20: Number of Publications per year in Artistic Anatomy

There is a noted rise in popularity of the field from 1997 on but it is not constant from year to year. All the publication years within this case study were mostly from 2010 to 2018, which is definitely more popular than anything before 1997 but one cannot truly say the bottleneck is connected to certain years of popularity. Overall, the case study allowed us to better understand the field of artistic anatomy in relation to its interesting network statistics. The collaboration seen within the network can only be described from one of non-metric descriptors studied: model organism. In the artistic anatomy field, a majority of authors work with humans as the organism and this allows for a spread of knowledge from the bottleneck author to his connections.

4.4.2 Case Study 2: Comparative Physiology

Comparative physiology was a field of interest selected due to its contradicting network statistics found by the AuthorMap tool. Comparative physiology is studying and exploiting the different functional characteristics between organisms; it is a study of diversity across species. All network statistics can be found in Table 10 below. A complete table of raw statistics for all 64 fields to use for comparison can be found in Supplemental Figure 2. This network was the only network constructed that was shown to be scale-free based on the node distribution analyzed by the powerlaw package.

The R value is -2.169, negative meaning the best fit is to the first distribution (scale-free) and a large value meaning it has a large maximum likelihood that it fits the scale-free node distribution example. The p value is 0.03 which would allow for acceptance of the hypothesis that this network follows scale-free characteristics. Within a scale-free network we would expect to see a high clustering coefficient, high degree centrality average and median, and a large number of communities. The only scale-free characteristic that this field matches is a high number of communities but not in comparison with some of the other fields. The clustering coefficient and degree centrality was also not as high as expected, it mostly matches the other sixty plus fields tested. The network analysis data that was expected for a statistically significant scale-free graph was not seen in the comparative physiology network.

Table 10: Network Analysis Data for Comparative Physiology

R	P	Total Nodes	Communities	Degree Cent. Avg.	Clustering Co.	Degree Cent. Median
-2.169	0.030	7361	4451	4.61E-05	0.447	3.48E-05

To better understand these network statistics and how this network was formed, four bottleneck authors were examined to find any patterns that may uncover how each connection within the network was built. Below in Figure 21 shows the comparative physiology network with the top three bottlenecks highlighted in yellow.

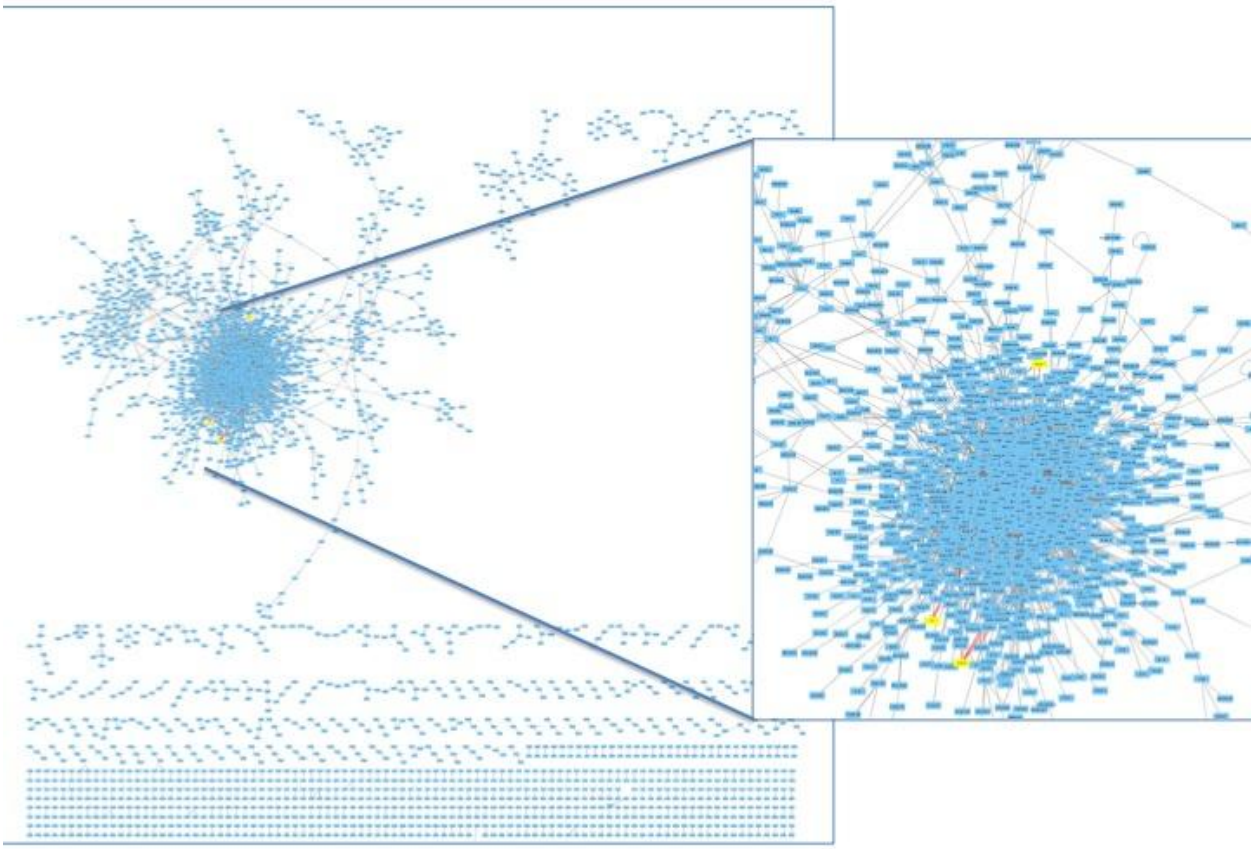

Figure 21: Zoomed in structure of the comparative physiology network (6,049 binary nodes not shown) with a zoomed-in view at the most dense part of the network and the three largest bottlenecks (corresponding to last authors Yu, Yang, and Dong, respectively) highlighted in yellow.

This figure is to be used for comparative purposes and zoomed in images of the individual bottlenecks and their connections can be found in the supplemental figures. In the network you can see a hairball structure, where a majority of the bottlenecks are found. The four bottlenecks, and their betweenness centrality further investigated in comparative physiology are shown below in Table 11.

Table 11: Betweenness Centrality values for the four bottlenecks studied in this section

Name of Last Author	Betweenness Centrality
Yu, Tao	9.939
Yang, Bin	9.932
Dong, Sijun	9.916
Chen, Feng	0.0001

4.4.2.1 Largest Bottleneck: Yu, Tao

The largest bottleneck within the entire network is Tao Yu, with a betweenness centrality of 9.94. Tao is an author from a medical university in China and has six published papers with the field of comparative physiology. Within the network he has collaborated with three other last authors, the structure of the connections can be seen in Supplemental Figure 7. Below in Table 12, the three first degree connections are listed along with their non-metric descriptors.

Table 12: CP Bottleneck 1: Yu, Tao & Connections Non-Metric Descriptors. All papers published within last decade- no other consistency among connections.

Name	Year	H-index	Journal Impact Factor	Organism
Yu, T	2012	2	3.40	Rabbits
Puhan, MA	2015	52	1.74	N/A
Huang, Z	2017	11	3.29	Plant
Jiang, Y	2016	63	4.12	Fruit

No spread of knowledge can be extracted from the table above. All three authors worked with a variety of different model organisms from rabbits to fruit with no overlap. The h-index also shows no consistency across all four authors and the bottleneck author has the lowest h-index with a value of 2. This result suggests that h-index doesn't necessarily correlate with betweenness centrality as was also demonstrated in the last case study. The journal impact factors are within a small range from one another, but not all within the same 1 point range. The only interesting and consistent aspect is that all the papers were published within the same decade and within five years of one another. This is similar to what we saw in the last case study and may be is in relation to the popularity of the field changing over time.

4.4.2.2 2nd Largest Bottleneck: Yang, Bin

The second largest bottleneck within the comparative physiology network is Bin Yang. Bin Yang was the last author on 23 publications within the comparative physiology query. The author is affiliated with Nankai University in China and has seven first degree connections. This can be seen in Supplemental Figure 8 as well as the table below. Table 13 lists all of the connections along with the information extracted about them from PubMed and Scopus.

Table 13: CP Bottleneck 2: Yang, Bin & Connections Non-Metric Descriptors. Most consistent non-metric descriptors across all 8 authors: year and journal impact factor.

Name	Year	H-index	Journal Impact Factor	Organism
Yang, B	2017	3	2.56	Bacteria
Liu, L	2018	24	4.48	Bacteria
Zhao, X	2016	8	2.04	Mice
Zhao, D	2015	15	2.63	Bacteria
Wang, J	2016	12	1.73	Zebrafish
Lei, T	2017	1	2.77	Humans
Wu, J	2017	9	0.47	Honey Bees
Perez-Enscio, M	2014	32	6.13	Pig

Again, the first observation noted is that the bottleneck author Bin Yang has a small h-index of 3 but is a bottleneck within the network, further supporting that there is no correlation between the two measures. The h-index also shows no consistency across all eight last authors ranging from 1-32. Another non-metric descriptor that shows no pattern across the eight authors is the model organism used. A lot of them are unique but all of the organisms are related to one another, as we know from evolution. Three of the eight authors are using a bacteria but the other five are using a variety of different organisms from pigs to honey bees. The year and journal impact factors are the two descriptors that are staying the most consistent. The journal impact factors are mostly consistent with a few outliers, most are within the range from 2.0-3.0. Among all eight of the authors, they have published within a few years of one another, with half of them publishing in 2017.

4.4.2.3 3rd Largest Bottleneck: Dong, Sijun

With a betweenness centrality of 9.92, Sijun Dong is the third largest bottleneck within the comparative physiology network. This author among the last two bottlenecks is from China and is affiliated with the National Institute of Urban Environment. Dong was the last author on 12 papers within the comparative physiology field, however only has two first degree connections in the network. Those connections can be seen visually in Supplemental Figure 9 or listed below in Table 14.

Table 14: CP Bottleneck 3: Dong, Sijun & Connections Non-Metric Descriptors. Preservation of the year across the three authors within this bottleneck.

Name	Year	H-index	Journal Impact Factor	Organism
Dong, S	2016	18	5.98	Mice
Wang, F	2017	9	3.15	Mice
Sun, Y	2017	16	5.81	Bacteria

Among the three authors the most stable measure is the year, all publishing from 2016-2017. The h-index when comparing all three authors is not identical but when comparing only Dong and Sun, they share comparable h-indexes and journal impact factors. Which may mean that within the comparative physiology field, h-indexes and journal impact factors are correlated. Dong and Sun however do not work with the same model organism but Wang and Dong do. One would expect a cluster of last authors who have collaborated to share more than one of these non-metric descriptors and in this bottleneck example the other sharing is within the year published.

4.4.2.4 Smallest Bottleneck: Cheng, Feng

To again contrast any of our finding within the larger bottlenecks, we looked at the smallest bottleneck. Feng Cheng last author who has the smallest impact of the flow of information within the network has a betweenness centrality of 0.0001. Cheng is affiliated with Nanjing Medical University which is also located in China as were all the other authors in this case study. Cheng was the last author on 65 papers within the comparative physiology field as this is his main field of study. In the network he has collaborated with five other last authors that can be seen in Supplemental Figure 10 or in Table 15 below.

Table 15: CP Bottleneck 4: Cheng, Feng & Connections Non-Metric Descriptors. Contrasted with the larger bottleneck, the same patterns are seen within organism and journal impact factor.

Name	Year	H-index	Journal Impact Factor	Organism
Chen, F	2017	18	1.89	Humans
Ren, L	2013	8	1.69	Humans
Kaufman, DB	2009	48	2.62	Mice
Kung, HF	2011	55	1.67	Humans
Berke, JD	2013	28	15.14	Humans
Chen, DJ	2009	14	1.91	Bacteria

The journal impact factor and model organism among the six authors stays relatively constant with two outliers in both categories. One author is the same outlier in both categories, Kaufman. With a slightly higher journal impact factor than the rest and using mice instead of humans in his studies labels Kaufman as an outlier. Chen uses bacteria instead of humans and Berke published in a journal with an impact factor of 15 while all the other authors published in journals with a range from 1.0-2.0. The h-index and the year are spread out and differ from author to author, no distinct patterns can explain why all these authors are connected to Chen based on the h-index and year. Although Chen was the smallest bottleneck, he has a larger h-index at 18 than two of the larger bottlenecks previously studied and tied with Dong. No other differences are noted between the larger and smaller bottlenecks. The majority of connections can be assumed to collaborated with Chen by using humans as the model organism and publishing within similar journals.

4.4.2.5 Studying Correlation

In order to determine whether there are any differences in the correlation previously studied from field to field based of the network characteristics is something we questioned. Following the technique stated within the methodology, we picked 30 bottleneck authors with a variety of betweenness centralities across the distribution of betweenness centralities. This will allow us to study if there are any relationships between h-index, journal impact factor, and betweenness centrality.

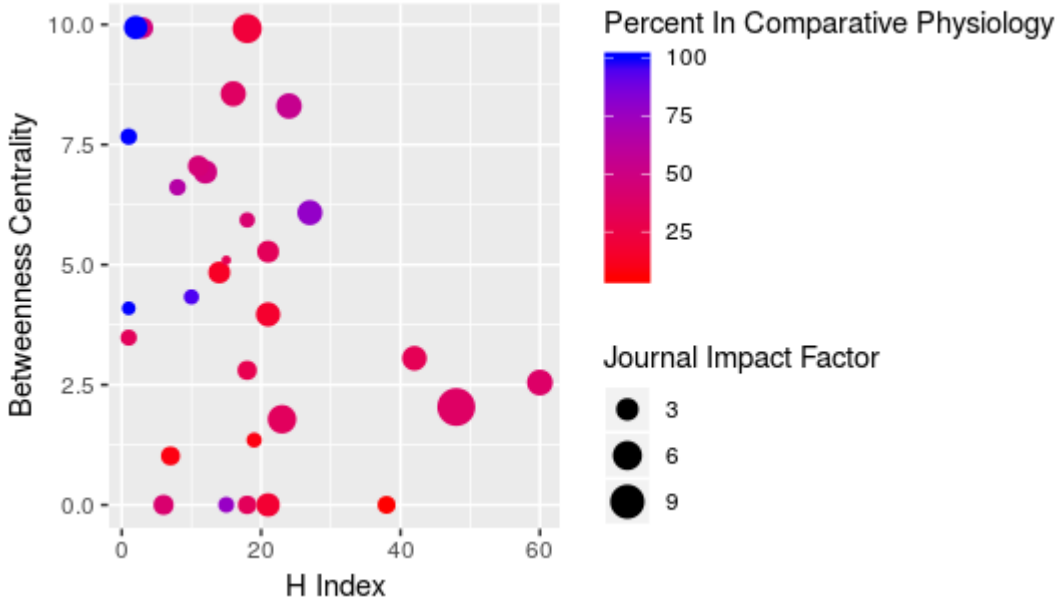

Figure 22: Correlation of H-index and Betweenness Centrality in Comparative Physiology. No correlation extracted between the four measures.

No correlation is seen between h-index and the betweenness centrality. For this field in particular, most authors within the field don't have an h-index above 30; but even authors with small h-indexes are just as likely to have large betweenness centrality as authors with large h-indexes. even authors with smaller h-indexes have the largest bottlenecks as seen within our case study above. Also, dots on the blue end of the spectrum are present in the figure above due to a lot of authors within this field work primarily in this field whereas in artistic anatomy many authors only had one paper in the field. Another interesting observation is that authors with larger h-indexes and small betweenness centralities don't have a large percent of publications in comparative physiology versus an author with a lower h-index and a high betweenness centrality. A reason for this pattern could be that comparative physiology is more segregated from other fields, an author that solely publishes papers about comparative physiology could have a lower h-index due to the lack of citations the paper is receiving because of the lack of overlap. Overall, some trends are noticed that were different from the artistic anatomy study.

A Pearson Correlation test was run across the five metrics and a heatmap was created to reflect these values in Figure 23 below. The figure shows the negative and positive correlation between every possible combination of the five metrics. H-index and betweenness centrality had a negative correlation value of -0.33. There were three positive correlations that were found in this case study versus the one found in artistic anatomy. The one that was found in artistic anatomy, between h-index and journal impact factor, was also found in comparative physiology but stronger. H-index and journal impact factor had a correlation value of 0.57. The second strongest positive correlation was between percent in the field and the betweenness centrality with a value of 0.38. Lastly, a small positive correlation of 0.07 was noted between the journal impact factor and the bottleneck value.

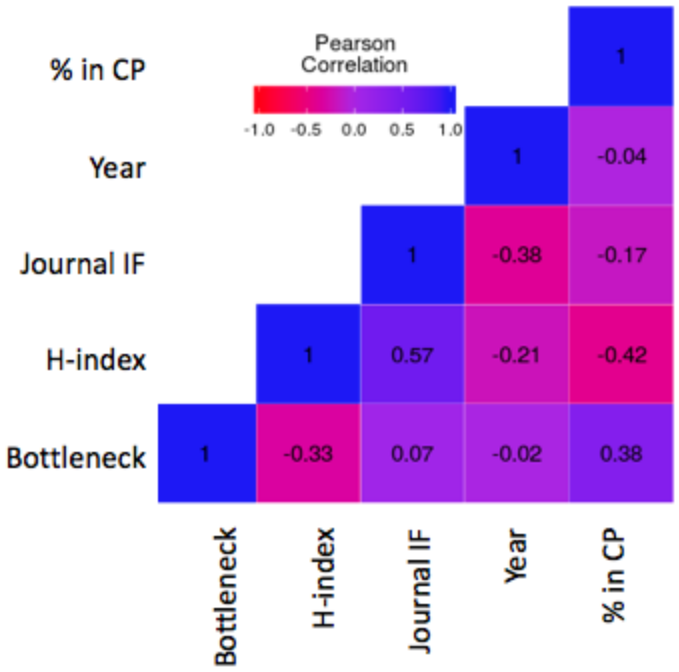

Figure 23: Heatmap showing the Pearson Correlation between the several non-metric descriptors in the comparative physiology field. With three positive correlations shown, the strongest being between journal impact factor and h-index; this was the only significant correlation (p= .001).

To see if any of these correlations, positive or negative, were statistically significant the p-values of the Pearson Correlation test were exported and can be found below in Table 16. The only statistically significant correlation was between journal impact factor and h-index with a p-value of 0.0011.

Table 16: P-Values for the Pearson Correlation heatmap. Only one statistically significant ($p \leq 0.01$) correlation between the h-index and journal impact factor was found.

	Bottleneck	H-index	Journal Impact Factor	Year	% in CP
Bottleneck	NA	0.076	0.700	0.932	0.037
H-index	0.763	NA	0.0011	0.256	0.022
Journal Impact Factor	0.700	**0.0011**	NA	0.037	0.372
Year	0.932	0.256	0.037	NA	0.822
% in CP	0.037	0.022	0.372	0.822	NA

Overall, comparative physiology had one statistically significant correlation found between h-index and journal impact factor. This means that authors with higher h-indexes were publishing in journals with high journal impact factors. This case study in addition to the artistic anatomy has shown that there is no correlation between h-index and betweenness centrality.

4.3.2.6 Summary

In conclusion, this case study varied in different ways from artistic anatomy and this may be due to the differences that were seen before in the network statistical analysis. Based on the four bottleneck authors studied there is no overarching patterns that were seen consistently among all cases. Starting with the non-metric descriptor h-index; there was no seen correlation between the h-index and the betweenness centrality. The two biggest bottlenecks within the field have h-indexes of two and three respectively. It was rare to see any sort of consistency with the h-index measure comparing the bottleneck author and his/her primary connections. One observation that was noted consistently in the individual bottlenecks examined is although the bottleneck author may not always have a large h-index, that author always collaborates with some authors that have larger h-indexes. The bottleneck author doesn't just collaborate with people that share the same influential measure.

This also stands true for model organisms within this field. An author that works with mice isn't only going to work with other authors that worked with mice. Below in Table 17, the distribution of papers working with the top five model organisms is shown. Comparing with artistic anatomy, this field uses a much larger range of organisms although humans is still the most popular. This observation is consistent with the name of the field itself; one might expect many different types of organisms to be represented.

Table 17: Model Organisms Populations in Comparative Physiology

C.elegans	Zebrafish	Mice	Drosophila	Yeast	Humans
2,335	2,517	113,648	8,497	6,546	661,996

It was seen in a few cases above where a majority of the connections do work with the same model organisms but never inclusive of all authors. For three out of four of the cases excluding Yu (largest bottleneck), this was true, where 50% of the connections did work with the same model organism as the bottleneck author supporting that there could be a potential spread of knowledge.

In most cases within this case study, all of the bottleneck authors and their primary connections were publishing in similar journals with comparable impact factors. In some situations there was outliers but a majority of the connections were within the same range of impact factors. This could potentially be due to the specialist journals within the field; many of the authors are publishing in the same journals that are unique to the field of comparative physiology which mostly have similar impact factors. For example, a paper about comparative physiology is not going to be

published in a journal about neuroscience but mostly likely in a medical journal about surgery or research.

The last research question pertains to the pattern of publishing years among each bottleneck. With an exception to the smallest bottleneck, Cheng, all the papers were published within the same decade similar to what was seen before in artistic anatomy. The third largest bottleneck, Dong, and his two connections published all within one year of each other. Below in Figure 24, shows the number of publications in each year for the field of comparative physiology.

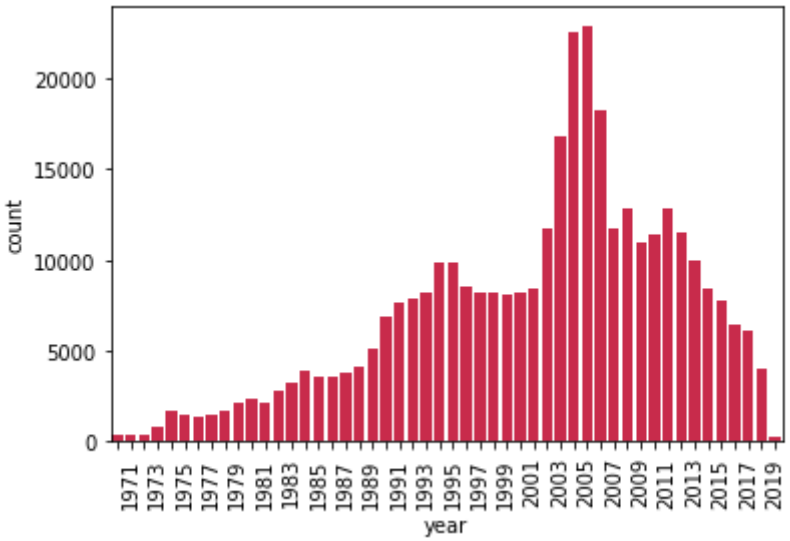

Figure 24: Number of Publications in Comparative Physiology

In the figure above, there is a clear rise in popularity from 2003-2007. This is not reflected within the bottlenecks we studied. The oldest paper found within our searches was 2009. We can conclude that the collaboration between last authors and betweenness centrality is not based on time or popularity within the field. There is potential that our tool could be bias, and collaborations could be based on time. Based on how we extract our information from the PubMed API, gathering the papers that are "most relevant", we also may be getting the most recent papers. This pattern can be seen across all three bottlenecks because most of the papers are published within the last decade. Since most of the bottleneck authors have multiple papers within the field, we could just be getting their most recent paper but they may have published previously within the popularity phase from 2003-2007.

In conclusion, this exploratory analysis of the field of comparative physiology showed that the network connections could be based on the journal impact factor and the model organism. There was no direct relationship to the h-index or the year which was a slightly different finding than the artistic anatomy case study. In both fields, organisms seems somewhat similar in a bottleneck cluster, but journal impact factor is only similar for comparative physiology. This could speak to the different network statistics that the two have and why one is a statistically significant scale-free network and the other is not. For example, another driving force for collaborations

between last authors may cause more communities which is seen when comparing artistic anatomy and comparative physiology. Each community in the comparative physiology network may be made up of authors working with the same model organism and publishing in journals with similar impact factors.

4.4.3 Case Study 3: Neurobiology

The third and final case study focuses on the field of neurobiology; it was chosen as a case study in addition to the other two because it represents the "norm" of all our data collected. Neurobiology was never an outlier in any of the network statistics and always followed the bulk of our data in any trends or correlations seen. Out of all the data that followed the "norm", neurobiology was selected because of the recent search at WPI for neuroscience faculty. It is a hot topic, and we wanted to learn more about it. Below in Table 18, the original raw data from the neurobiology network analysis is presented with the network best fitting a scale-free node distribution but the fitting was not statistically significant. The other values for clustering coefficient, degree centrality, and number of communities was average when compared with all the other networks in the Narrow List. A complete table of raw statistics for all 64 fields to use for comparison can be found in Supplemental Figure 2.

Table 18: Neurobiology Network Analysis Statistics

R	P	Total Nodes	Communities	Degree Cent. Avg.	Clustering Co	Degree Cent. Median
-0.659	0.510	6,054	3,325	4.64E-05	0.427	3.49E-05

The same methods performed on the last two case studies, artistic anatomy and comparative physiology, were done again, looking at the three largest bottlenecks and the smallest bottleneck for contrast. In the previous two case studies we found that certain non-metric descriptors could explain the collaboration between a bottleneck author and another last author. Hopefully, we wish to uncover the same information in this case study to better understand the structure of the network and if it can better be described by scale-free or random characteristics. Below in Figure 25 is the full neurobiology network for comparison of where the bottlenecks are found with relation to the whole.

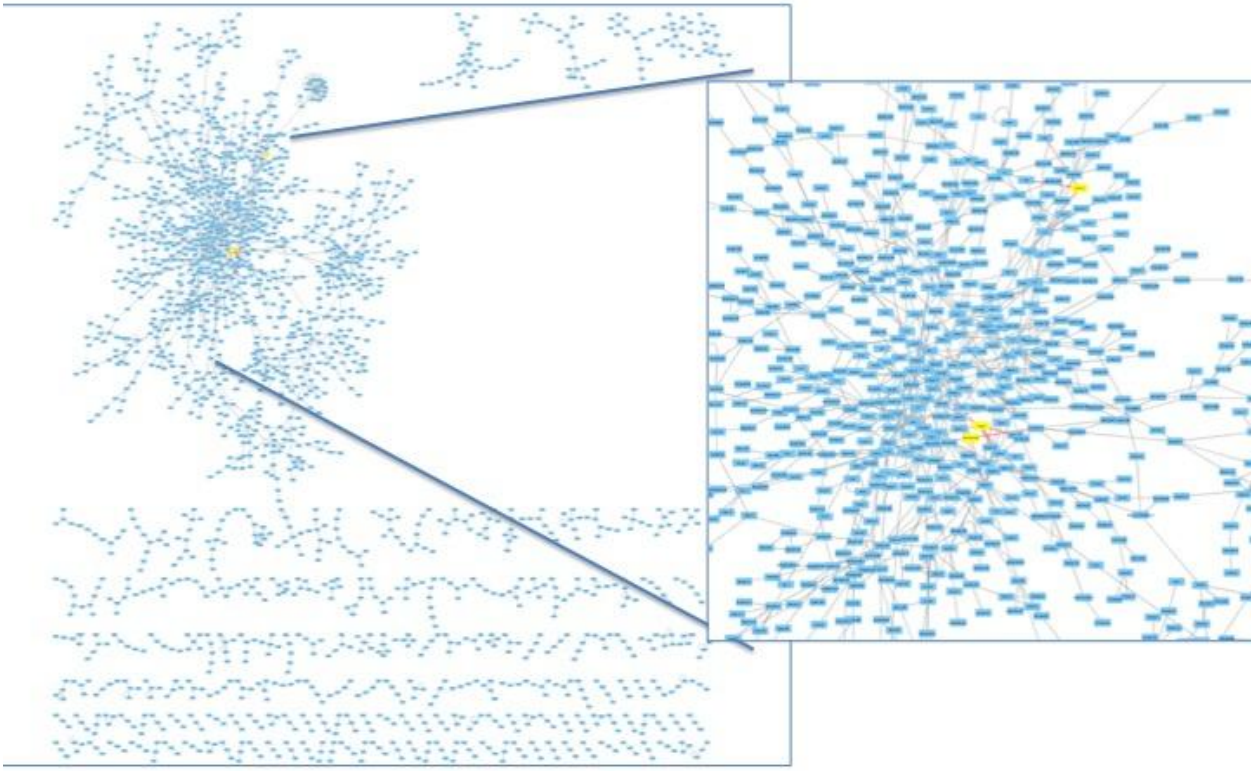

Figure 25: Zoomed in structure of the neurobiology network (3,000 binary nodes not shown) with a zoomed-in view at the most dense part of the network and the three largest bottlenecks (corresponding to last authors Marenda, Yang, and Xiao, respectively) highlighted in yellow.

In the network you can see a hairball structure, where a majority of the bottlenecks are found. The three largest bottlenecks are highlighted in yellow within the network. The four bottlenecks, and their betweenness centrality further investigated in neurobiology are shown below in Table 19.

Table 19: Betweenness Centrality values for the four bottlenecks studied in this section

Name of Last Author	Betweenness Centrality
Marenda, Daniel R.	9.955
Yang, Yi	9.793
Xiao, Hai	9.791
Bai, Yun	0.0001

4.4.3.1 Largest Bottleneck: Marenda, Daniel R.

With a betweenness centrality of 9.96, Daniel Marenda is the largest bottleneck within the neurobiology network outputted by the AuthorMap tool. Marenda is affiliated with the Department of Biology at Drexel University. He has been cited 482 times by 332 documents that he was an author on. Marenda is a last author on five papers within the neurobiology field. Within the network he has seven first degree connections, the connections can be seen in Supplemental Figure 11 and also listed below in Table 20.

Table 20: NB Bottleneck 1: Marenda, Daniel R. & Connections Non-Metric Descriptors. Consistency seen in none of the non-metric descriptors across the eight authors, no reason can be deduced why these authors collaborated.

Name	Year	H-index	Journal Impact Factor	Organism
Marenda, DR	2014	14	2.77	Drosophila
Shen, X	2014	28	1.85	Humans
Racke, MK	2006	59	1.85	Humans
Zhang, T	2017	13	12.51	Humans
Moreno, RA	2015	20	1.41	N/A
Busatto, GF	2012	42	6.16	Humans
Zhang, Y	2017	9	5.97	Variety
Carvalho, AF	2018	5	5.08	Humans

At the first glance, no immediate conclusions can be drawn from the table above that contains the eight authors and all their non-metric descriptors. The bottleneck author does not have the highest h-index and is connected to four other last authors that have higher h-indexes. Marenda also is the only author to work with Drosophila, this knowledge is not spread to his primary connections as a majority of them work with humans. In addition, the journal impact factor and year have no consistency between all eight authors. Some of the primary connections share some metrics, but none of the connections directly match the bottleneck author. The author with the most in common with Marenda is Shen; both of them published in 2014, Shen has double the h-index of Marenda; Marenda has a larger journal impact factor by ~1.0 and works with Drosophila while Shen works with humans. It can not be stated that there is any reason why these authors collaborated with Marenda based on the four non-metric descriptors we studied.

4.4.3.2 2nd Largest Bottleneck: Yang, Yi

Yi Yang, a scholar from Hangzhou Key Laboratory of Medical Neurobiology in China, is the second largest bottleneck within the neurobiology network with a betweenness centrality of 9.79. Yang has an h-index of 17 and has been cited 2,330 times by 2,251 papers. In the specific field of neurobiology, he is a last author on 15 papers. Within the network outputted by the AuthorMap tool, Yang has collaborated with five last authors. These connections can be seen within a network screenshot in Supplemental Figure 12 or listed below in Table 21.

Table 21: NB Bottleneck 2:Yang, Yi & Connections Non-Metric Descriptors. Preservation of two different model organism across connections within this bottleneck.

Name	Year	H-index	Journal Impact Factor	Organism
Yang, Y	2016	17	5.17	Humans
Racke, MK	2006	59	1.85	Humans
Qian, Z	2017	21	3.03	Mice
Miller, LC	2010	22	7.23	Humans
Ross, TJ	2015	30	4.93	Humans
Stein, EA	2000	58	3.87	Mice

Yang published his most cited paper in 2016 using humans as a model organism in a journal with an impact factor of 5.17. None of his connections published in 2016 and most published a decade or more before. A majority of the collaborators incorporated into their study the same model organism that Yang worked with, humans, and the other two worked with mice. All of the authors that Yang worked with have large h-indexes, and Yang with the largest betweenness centrality actually has the smallest influential factor at 17. Lastly, the journal impact factor was studied and it is noticed there isn't any outliers but very spread out from 1.0-8.0. These authors did not connect based on year, h-index, or journal impact factor, it was based on the knowledge of neurobiology applied to the organism.

4.4.3.3 3rd Largest Bottleneck: Xiao, Hai

The last bottleneck to be studied was Hai Xiao. Xiao is from China and was affiliated with First Affiliated Hospital of Gannan Medical University. Xiao has not published many papers in neurobiology or in other fields. How can this be if he is a bottleneck? He has only been cited 21 times by 18 documents. In the network, Xiao is connected with four other last authors. These connections can be see in Supplemental Figure 13 or listed below in Table 22.

Table 22: NB Bottleneck 3: Xiao, Hai & Connections Non-Metric Descriptors. Interesting outliers can be studied within this bottleneck but the outlier in every category is not the same author. A majority of authors are working with the same model organism.

Name	Year	H-index	Journal Impact Factor	Organism
Xiao, H	2018	3	2.23	Mice
Jing, J	2017	26	4.01	Aplysia
Krogan, NJ	2018	79	31.40	Humans
Qi, Y	2018	10	1.29	Mice
Li, X	2017	4	3.16	Variety (mice included)

Some observations noted from the table above was that one author served as an outlier to the rest, and it was not the bottleneck author. Krogan, with an h-index of 79 and publishing in a Cell with a 31.40 impact factor speaks to how influential the author must be. Xiao only has an h-index of 3 and published in a journal with a 2.23 impact factor and Xiao and Krogan do not work with the same model organism. Krogan worked with humans while Xiao and two of his other connections worked with mice. The only connection between Xiao and Krogan based on these non-metric descriptors is the year published, both in 2018. Qi also published in 2018 and the remaining two authors published in 2017, only a few years apart from one another. The closest author to match Xiao's non-metric descriptors is Li; their h-index vary by one point as well as their journal impact factor and both worked with mice. Why do only some of the connections show a spread of knowledge and why do bottlenecks authors only collaborate with authors similar to them? Every field has demonstrated different patterns and trends among their bottleneck authors and connections. In this case study it seems to be most similar to comparative physiology.

4.4.3.4 Smallest Bottleneck: Bai, Yan

To contrast any trends seen within the larger bottlenecks in the neurobiology field, we look to the smallest bottleneck. Yan Bai has a betweenness centrality 0.0001. Bai is affiliated with The Third Military Medical University in China. This author has been cited 2,116 by 1,976 documents. Within the neurobiology network Bai has collaborated with three other last authors.

Table 23: NB Bottleneck 4: Last author Yan Bai and Connections Non-Metric Descriptors. Contrasting with the larger bottlenecks, the most correlation across all non-metric descriptors is seen as well as the smallest bottleneck having the largest h-index across all four studied.

Name	Year	H-index	Journal Impact Factor	Organism
Bai, Y	2006	24	3.16	Mice
Zheng, X	2014	9	1.22	Mice
Li, X	2017	4	3.16	Variety including Mice
Wang, N	2017	4	2.28	Mice

This cluster of last authors seems the most similar out of the top three largest bottlenecks. All three collaborators worked with mice, so a spread of knowledge is seen there. All four journal impact factors are similar with Bai and Li publishing in the same journal, Neuroscience Bulletin, thus an identical impact factor. Bai has the largest h-index, and the other authors h-indexes are comparable with Li and Wang both having an h-index of four. When comparing with the top three bottlenecks, smaller numbers are noticed for h-index and the journal impact factor. All of Bai's connections published recently but he published in 2006. So this collaboration is not based off of time but more surely off of model organisms and publishing in similar journals.

4.4.3.5 Studying Correlation

Based on the contrast of the h-index in the smallest bottleneck and the larger bottlenecks, a correlation study was produced. Thirty bottlenecks with a variety of between centralities based on an even spread across the distribution were selected and their h-index, percent in neuroscience, and journal impact factors were extracted from various sources.

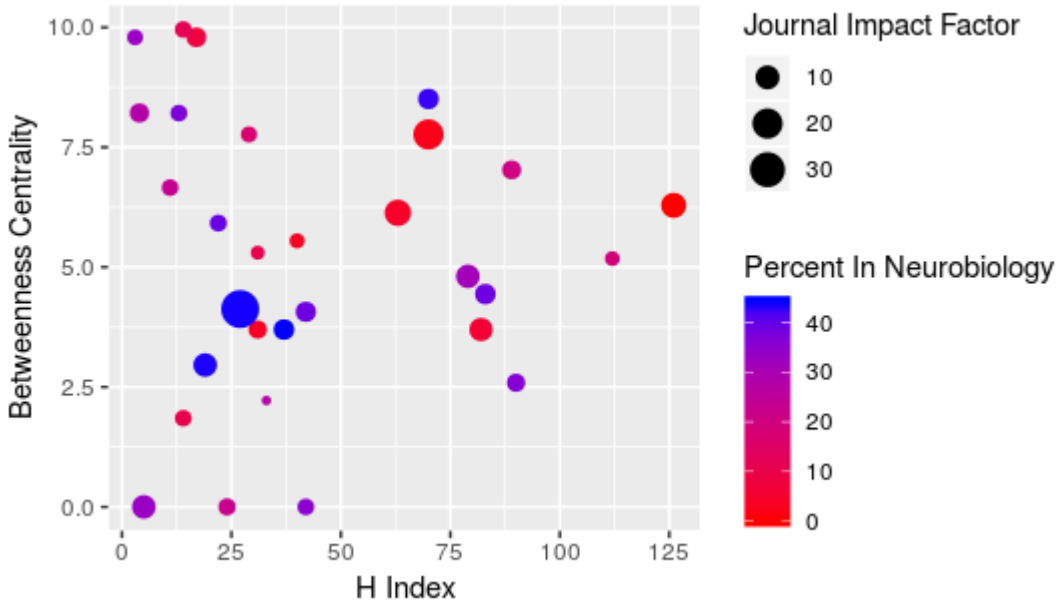

Figure 26: Correlation of H-index and Betweenness Centrality in Neurobiology. No correlation can be noted between h-index and betweenness centrality.

In Figure 26 no correlation can be noted between h-index and betweenness centrality within the field of neuroscience that was not seen in either of the other two fields. Some of the highest betweenness centralities are found in the top left corner, a position meaning they have a low h-index. The h-indexes in this field were much higher than the average seen in comparative physiology or artistic anatomy. The journal impact factor is shown by the size of the dot, and no correlation is seen with betweenness centrality or with h-index. Our author that published in the most influential journal only had an h-index of 25. Something very interesting about this graph that is absent in most of the other is the presence of scientists who publish in this field and it is their main field of work. More blue and purple dots are seen in this graph versus red dots. This could be because neurobiology is a more established field or a variety of other reasons.

To better quantify the correlation seen above in Figure 26, a pearson correlation test was run on the raw data in R. A heatmap of the correlation values can be found below in Figure 27. The same gradient scale is used from the scatterplot above. The correlation value between h-index and betweenness centrality is -0.03, so there is no correlation between these two factors with a correlation value of almost 0. Similar to comparative physiology there are three positive correlation values found within the heatmap. The strongest positive correlation value is between percent in neurobiology and the year the paper was published with a value of 0.48. The other two positive correlation values were small between percent in neurobiology and journal impact factor as well as h-index and journal impact factor with values of 0.08 and 0.13 respectfully.

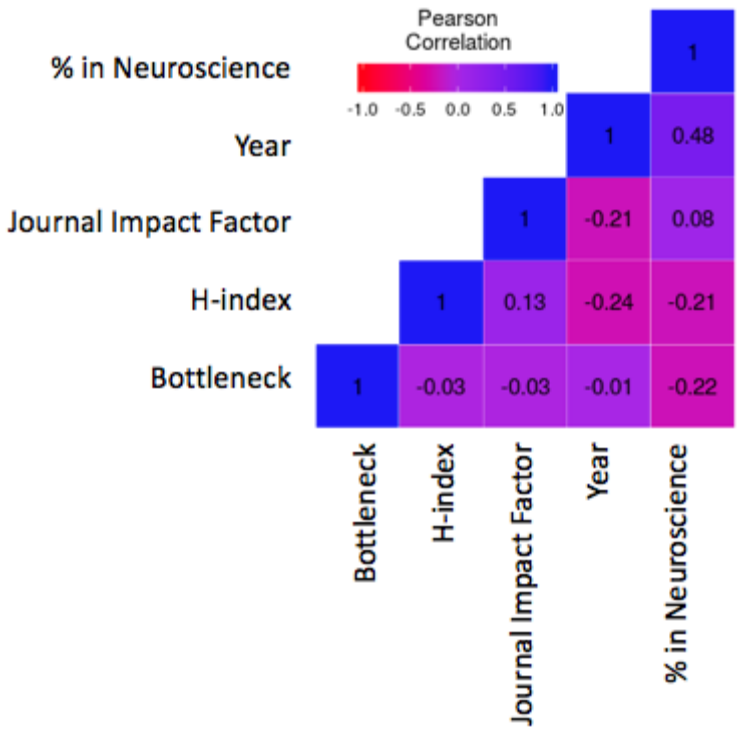

Figure 27: Heatmap showing the Pearson Correlation between the several non-metric descriptors in the neurobiology field. With three positive correlations shown, the strongest being between the percent in neuroscience and year.

Table 24: P-Values for the Pearson Correlation heatmap. Only one statistically significant ($p \leq 0.01$) correlation between the percent in neurobiology and year was found.

	Bottleneck	H-index	Journal Impact Factor	Year	% in NeuroB
Bottleneck	NA	0.857	0.867	0.948	0.251
H-index	0.857	NA	0.482	0.206	0.275
Journal Impact Factor	0.867	0.482	NA	0.268	0.684
Year	0.948	0.206	0.268	NA	**0.007**
% in NeuroB	0.251	0.275	0.684	0.007	NA

The p-values were extracted for every correlation value. Seen above in Table 24, there is only one statistically significant correlation between percent in neurobiology and the year. The p-value between these two metrics was 0.007.

Overall, the neurobiology case study allowed us to determine that every life science field is different with correlation between the non-metric descriptors studied. The only patterns that could be extracted were that there was a positive correlation between h-index and the journal impact factor in all case studies but only in comparative physiology was the correlation statistically significant. As well as there is no correlation between h-index and betweenness centrality in all three cases which went against our initial hypothesis. Neurobiology was the only field with a statistically significant correlation between percent in neurobiology and the year published. More correlation studies would need to be completed to see if any consistent patterns arise.

4.4.3.6 Summary

During the exploratory analysis of neurobiology, the field that followed the "normal" trend in data during network analysis, we collected information about the non-metric descriptors for the major bottlenecks within the field to answer our four research questions. The first question focused on h-indexes within the network in conjunction with betweenness centrality. Throughout our case studies we saw no evidence that any correlation was present, as our three largest bottlenecks had an h-index of 14, 17, and 3. Using the smallest bottleneck as contrast with an h-index of 24 helps supports the lack of correlation. To better explore more bottlenecks the correlation study was completed, and the statistical analysis supports our other two case studies that also showed no correlation between the two measures. This case study as well showed that authors don't always collaborate with authors that have similar h-indexes.

Looking at the journal impact factors, we studied the correlation in addition to the spread of knowledge. For the third time it was shown that an author does not have to have a large betweenness centrality or h-index to publish in a high impact journal. To make sense of this scenario, new young investigators just starting their labs may publish in high impact factor journals but potentially could not have accumulated enough articles to have a high h-index. Within the neurobiology field, not all connections publish in comparable journals. There are usually some outliers within a cluster but they are usually all within a small range of one another.

Although the journal impact factor was not shared from author to author, the model organism was. In all three of the largest bottlenecks, a majority if not all authors were working with the same model organism. This same pattern was seen within the smallest bottleneck as well. Below in Table 25, the top five organisms declared by PubMed are presented along with the number of papers in neurobiology using them.

Table 25: Model Organisms Populations within Neurobiology Field

C.elegans	Zebrafish	Mice	Drosophila	Yeast	Humans
596	916	19720	2219	314	34141

The most common model organisms are humans and mice within the entire neurobiology field and that is reflected within in our cases studies, as the majority is usually working with one of the two. Lastly, the years of publications were studied. Less consistency was found in neurobiology than in the other two case studies with exception to the second largest bottleneck which all articles were published from 2017-2018. In the other two case studies a lot of the publications were grouped by the bottleneck and published within a decade of each other. A majority of the neurobiology publications studied are from 2014- 2017 but spread out among the four studied bottlenecks. Below in Figure 28, one can see that this is expected based on the popularity of the field.

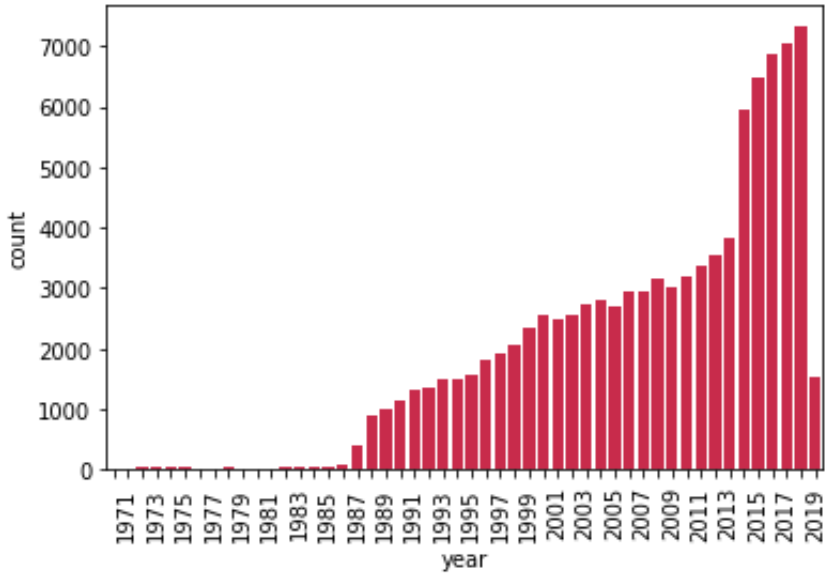

Figure 28: Number of Publications per year in Neurobiology

The field of neurobiology became a hot topic in 2014 and that is mirrored by the largest bottlenecks within this field. Overall, this case study allowed us to better understand the edges within the network and to notice any patterns or trends within the non-metric descriptors to help explain the type of network. Neurobiology collaborations are seem to be based on the model organisms which is the same seen across all three case studies in large bottlenecks and in small bottlenecks.

4.4.4 Case Study Conclusion

In conclusion, the case studies helped better understand the structure of the three networks studied and the potential reason(s) for the collaboration between last authors. Through studying the three largest bottlenecks and contrasting with the smallest bottleneck, we were able to categorize the connections based on non-metric descriptors and study correlation between the metrics. All three networks, despite their varying network statistics, the bottlenecks and connections were driven through the use of model organisms. The bottleneck author were more likely to collaborate with other last authors that experimented with the same organism. In addition to model organisms within the only statistically significant scale-free network, comparative physiology, the collaborations between last authors can also be categorized by the journal impact factor. This may be the reason why the structure of this network varies from the other two and is scale-free.

Through the correlation study, in artistic anatomy and comparative physiology, there was no correlation between h-indexes and betweenness centrality as well as between journal impact factor and betweenness centrality. One important observation that could be supported with further work is the potential discovery that a bottleneck author does not need to be the most “influential” within a network, where influence is represented by the h-index. A pattern was noted in each case study that a bottleneck connects more important last authors and their hubs together. Usually all of their connections had higher h-indexes which originally was not expected. In short, these case studies gave us more insight into the structure and characteristics of several of the networks built by our tool and gave us better understanding of the nature of scale-free networks.

5 Future Work & Final Conclusions

In conclusion, through this project we were able to gain a better understanding of scale-free networks and how authors in the life sciences interact and collaborate. We built a systematic network-building tool in Python using the PubMed API (GitHub: https://github.com/hnorthcott/authorMap-). We connected this program to Cytoscape for network visualization, and implemented NetworkX for network analysis. In conclusion, we found that a majority of our networks demonstrated scale-free characteristics, but not all of them could be statistically significant by the PowerLaw package. In addition, we performed several case studies on three life science fields of interest based on their characteristics demonstrated by the tool: artistic anatomy, comparative physiology, and neurobiology. These case studies helped us understand how several of the non-metric descriptors (such as h-index, journal impact factor, and experimental organism) can drive whether authors collaborate. We were surprised to discover that there was no correlation between h-index and betweenness centrality as well as journal impact factor and betweenness centrality. However, these case studies did help us to better understand the characteristics of a scale-free network.

For other researchers continuing this work, we recommend first and foremost that a UI be added to our network analysis tool. This will allow for better interactivity. Additionally, we recommend updates to the network visualization in Cytoscape. More dimensions should be added to the dictionary so that the network can be color-coded based on other characteristics, such as h-index or journal impact factor. This would add dimensionality to the network visualizations and therefore allow users to get a better understanding of the full picture of the network.

Supplemental Information

Broad List	R	P	Total Nodes	Communities	Degree Cent	Clustering Co	Median
anatomy	-1.717779924	0.08583676476	7728	5174	0.0002542955194	0.5999090297	0.000129315918
biochemistry	-0.3972819186	0.6911595794	7523	5684	0.0001276715345	0.5951750064	6.55E-05
biology	-0.8789972351	0.379402775	6959	4550	3.69E-05	0.5336223943	4.50E-05
biophysics	-0.5885810072	0.5561423777	6654	3936	2.21E-05	0.533097729	3.46E-05
biotechnology	-0.7214324497	0.4706434876	5860	3392	8.04E-06	0.4947410278	2.88E-05
chronobiology	-1.582412134	0.1135555165	2375	1017	1.45E-06	0.5048305151	2.69E-05
neurosciences	-0.761049053	0.4466277689	6243	4244	9.31E-06	0.5076491225	2.31E-05
pharmacology	-0.4433112246	0.657540638	6037	2553	1.53E-06	0.4581265531	2.02E-05
physiology	-0.3855194384	0.6998526076	6056	3473	6.06E-06	0.4638147705	1.80E-05
toxicology	-0.2252516786	0.8217834885	5433	2615	3.02E-06	0.4600799107	1.64E-05

Figure 1: Screenshots of the final results of raw data collected from our AuthorMap tool from the Broad List. Positive R values suggest a better fit to the log-normal / random distribution, while negative R values suggest a better fit to the power law / scale free distribution; however, none of these values reached statistical significance.

Field	BroadField	R	P	Total Nodes	Communities	Degree Cent. Avg.	Clustering Co	Degree Cent. Median
anatomy artistic	anatomy	-0.5367428162	0.5914452763	1097	995	0.001293961513	0.7921428571	0.0009124087591
anatomy comparative	anatomy	-0.729304007	0.4658157218	8210	6069	0.0001882225218	0.5909693596	0.000107492206
anatomy cross-sectional	anatomy	-0.9249313291	0.355001629	7401	3982	0.0001215589295	0.5487907432	6.12E-05
anatomy regional	anatomy	-1.529296058	0.1261910664	6585	3629	9.15E-05	0.515520953	4.36E-05
anatomy veterinary	anatomy	-1.495891572	0.134681913	5607	2131	7.63E-05	0.508670936	3.50E-05
cell biology anatomy	anatomy	-1.108276272	0.2677425197	6632	3633	6.50E-05	0.489117338	2.84E-05
embryology	anatomy	-0.8335307876	0.4045454417	5102	2145	5.84E-05	0.4723676708	4.97E-05
histology	anatomy	-1.965117679	0.04940061885	3866	1420	5.40E-05	0.464883252	4.53E-05
neuroanatomy	anatomy	-1.025517326	0.3051191543	3785	1223	5.12E-05	0.4660792825	4.17E-05
ostology	anatomy	-0.7240785973	0.4690174834	644	139	5.14E-05	0.4629716298	4.12E-05
carbohydrate biochemistry	biochemistry	-0.3493091241	0.7268572494	5683	3219	0.0003888430325	0.52571736	0.0001759633996
chemistry bioinorganic	biochemistry	-0.1998819091	0.8415729394	1037	557	0.0003381914836	0.509933257	0.0001487873828
histocytochemistry	biochemistry	-0.5150657069	0.6065071054	8201	5168	0.0001525493187	0.4970506341	6.70E-05
immunochemistry	biochemistry	-0.06116740356	0.9512258893	7552	4625	0.0001003185926	0.5104428878	4.46E-05
metabolomics	biochemistry	-1.021404507	0.3070628311	4380	933	1.00E-04	0.4507738955	7.45E-05
molecular biology	biochemistry	-1.023924078	0.3058711363	7249	5370	7.42E-05	0.4549726469	5.87E-05
neurochemistry	biochemistry	-0.4052991789	0.68525761	4855	1782	6.87E-05	0.4565033895	5.14E-05
proteomics	biochemistry	-0.8921885298	0.3722918891	3883	680	6.58E-05	0.4367696199	4.67E-05
botany	biology	-1.022456767	0.3065647647	6419	3858	5.75E-05	0.4383394056	4.06E-05
cell biology, biology	biology	-0.9913639802	0.3215078819	5179	2697	5.13E-05	0.4340849579	3.68E-05
computational biology	biology	-1.020346809	0.3075640108	6866	3854	0.0003740230295	0.3929639925	0.0002913752914
cryobiology	biology	-1.019287074	0.3080666987	2598	1376	0.0002608639284	0.4899820082	0.0001056524036
developmental biology	biology	-1.019822433	0.307812682	6378	3599	0.0001572038194	0.4556300792	6.31E-05
ecology	biology	-1.021920562	0.3068184994	7479	4284	0.0001022777002	0.4572752804	4.29E-05
exobiology	biology	-1.017825909	0.3087606971	1545	477	9.79E-05	0.4672459614	4.02E-05
genetics	biology	-1.024128416	0.3057746241	6627	3000	8.39E-05	0.4301322048	6.35E-05
laboratory animal science	biology	-0.4918735396	0.622808745	6100	2385	7.62E-05	0.4012336119	5.32E-05
microbiology	biology	-0.1097200853	0.9126313708	6250	3100	6.38E-05	0.4104861853	4.56E-05
natural history	biology	-0.5119748017	0.6086686452	7392	4194	5.31E-05	0.4184912243	3.90E-05
neurobiology	biology	-0.6589419065	0.5099330742	6054	3325	4.64E-05	0.4273309696	3.49E-05
parasitology	biology	-0.8086534455	0.4187145135	6558	3293	0.0003362189108	0.5119314011	0.0003051571559
photobiology	biology	-0.98918939	0.3225704885	1635	1016	0.000268183555	0.5269237109	0.0002442002442
radiobiology	biology	-1.04390781	0.2965280446	5167	2129	0.0001847758353	0.5193509612	0.0001496669909
sociobiology	biology	-1.04390781	0.2965280446	534	415	0.0001770345109	0.519023023	0.0001439055979
synthetic biology	biology	-1.040350998	0.2981768588	5333	2946	0.0001277396218	0.5083818822	0.0001040257984
zoology	biology	-1.035440851	0.3004630764	6439	3150	9.90E-05	0.4876557058	7.79E-05
bionics	biophysics	-1.035154581	0.3005967262	1292	628	9.59E-05	0.4839692119	7.42E-05
electrophysiology	biophysics	-1.023924078	0.3058711363	6889	3345	7.47E-05	0.4797354542	5.91E-05
biomimetics	biotechnology	-0.5831560442	0.5597882584	3815	1754	6.80E-05	0.4803081116	5.31E-05
cognitive neuroscience	neuroscience	-0.5831560442	0.2965280446	5585	1766	6.10E-05	0.4694065445	6.96E-05
neuroanatomy	neuroscience	-0.2652996285	0.7907786507	5395	3157	0.0004012330752	0.6373126654	0.0003705762461
neurobiology	neuroscience	-0.1524714521	0.8788151051	6633	4194	0.0001694943244	0.5855416287	8.31E-05
neurochemistry	neuroscience	-0.664411786	0.5064267915	4525	1790	0.0001465706991	0.5419195663	0.00012083862
neuroendocrinology	neuroscience	-0.5200878357	0.6030023563	3893	1136	0.0001201015381	0.5424473304	9.78E-05
neuropathology	neuroscience	-0.5831560442	0.5597882584	4938	1650	0.0001013229384	0.4909539507	7.88E-05
neuropharmacology	neuroscience	-0.6421125313	0.52080012	4214	1268	8.66E-05	0.4738776699	6.76E-05
neurophysiology	neuroscience	-0.4580938166	0.6468850411	6223	4383	6.91E-05	0.4771100584	5.58E-05
biopharmaceutics	pharmacology	-0.5136621638	0.6074882095	3736	1596	6.44E-05	0.4772601389	5.05E-05
chemistry, pharmaceutical	pharmacology	-0.1897342524	0.5402670357	5481	2663	5.67E-05	0.4745817605	4.44E-05
ethnopharmacology	pharmacology	-0.664411786	0.5597882584	1279	614	5.55E-05	0.4740295689	4.32E-05
neuropharmacology	pharmacology	-0.9270670292	0.3538917335	6091	3086	0.0003452013005	0.4928121389	0.0003278151123
pharmacoepidemiology	pharmacology	-0.9719012859	0.33109966	3541	1068	0.000268228271	0.5050485927	0.0002073613271
pharmacogenetics	pharmacology	-1.269694781	0.2041933726	5505	2680	0.000171621743	0.4856170297	0.0001320132013
pharmacognosy	pharmacology	-0.4274182014	0.6690747525	5745	2248	0.0001322988156	0.4793224332	9.57E-05
pharmacology, clinical	pharmacology	-0.6124092313	0.5402670357	6651	3430	0.0001007441063	0.4609736595	7.26E-05
psychopharmacology	pharmacology	-0.4787559193	0.6321122793	4632	2345	8.50E-05	0.4640235509	6.22E-05
electrophysiology	physiology	-0.3946330293	0.6931137341	6267	2757	7.05E-05	0.4582140943	5.20E-05
endocrinology	physiology	-0.1897342524	0.8495173788	5269	2482	6.24E-05	0.4498856058	4.57E-05
neurophysiology	physiology	-1.018994079	0.3082057774	6467	4593	5.28E-05	0.453802426	3.98E-05
physiology, comparative	physiology	-2.169116034	0.03007387571	7361	4451	4.61E-05	0.4485122577	3.48E-05
psychophysiology	physiology	-1.924080819	0.05434445388	5694	3083	4.12E-05	0.4562494733	3.16E-05
ecotoxicology	toxicology	-0.1585033835	0.8740601547	4311	1330	3.95E-05	0.4505884234	4.64E-05
forensic toxicology	toxicology	-0.9687775208	0.3326562046	3615	783	3.82E-05	0.4520289925	3.49E-05
toxicogenetics	toxicology	-0.8514919027	0.3944961594	423	119	3.80E-05	0.4505422067	3.68E-05

Figure 2: Screenshots of the final results of raw data collected from our AuthorMap tool of the Narrow List

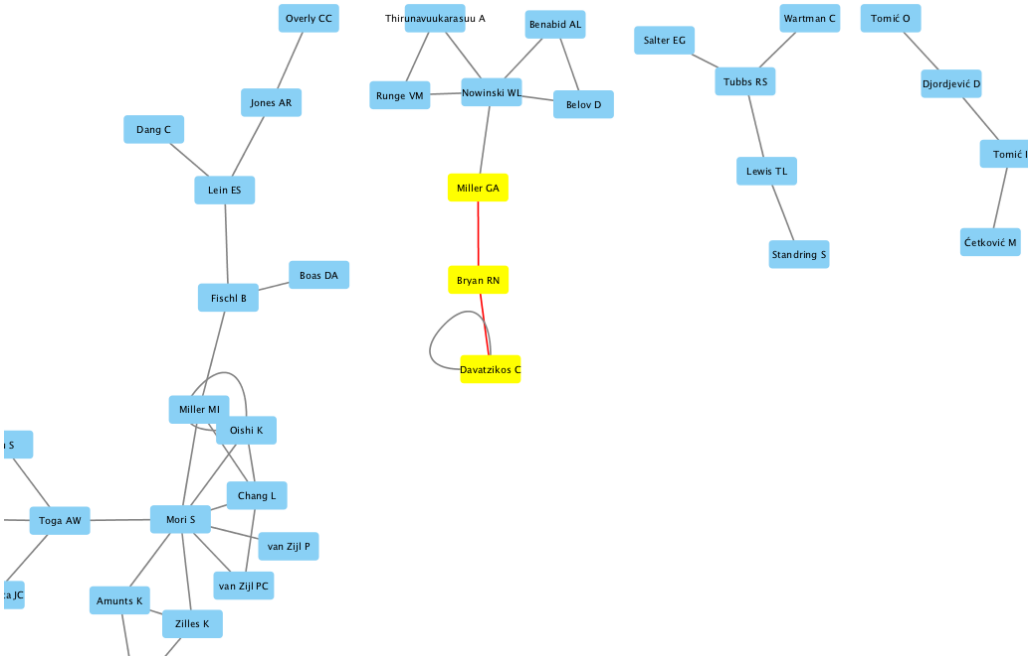

Figure 3: Artistic Anatomy Bottleneck 1- Bryan, RN. and Connections

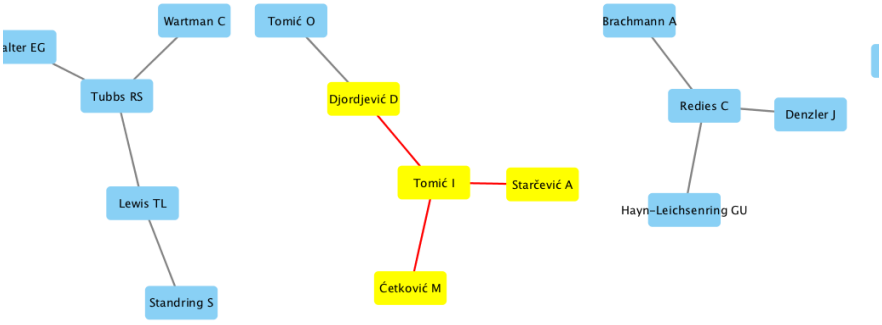

Figure 4: Artistic Anatomy Bottleneck 2- Tomic, I. and Connections

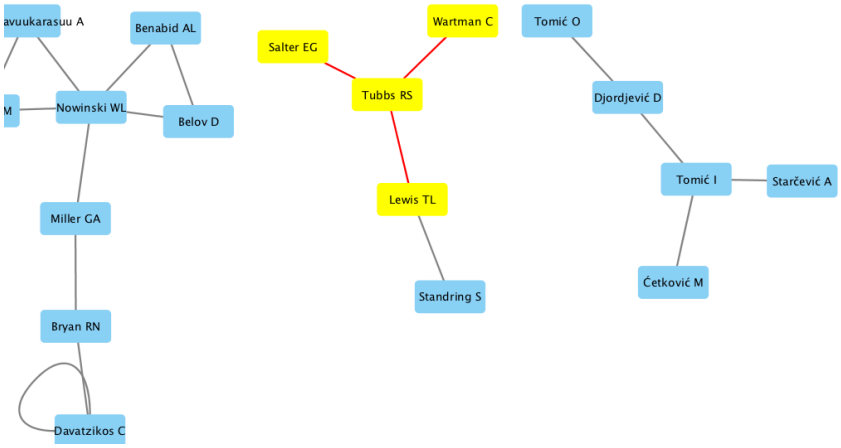

Figure 5: Artistic Anatomy Bottleneck 3- Tubbs, RS. and Connections

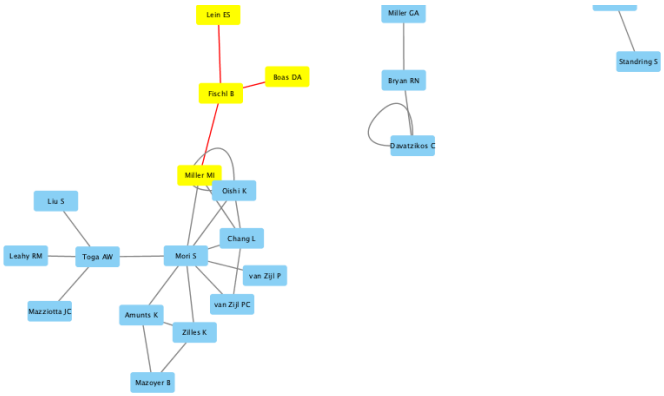

Figure 6: Artistic Anatomy Bottleneck 4- Fischl, B. and Connections

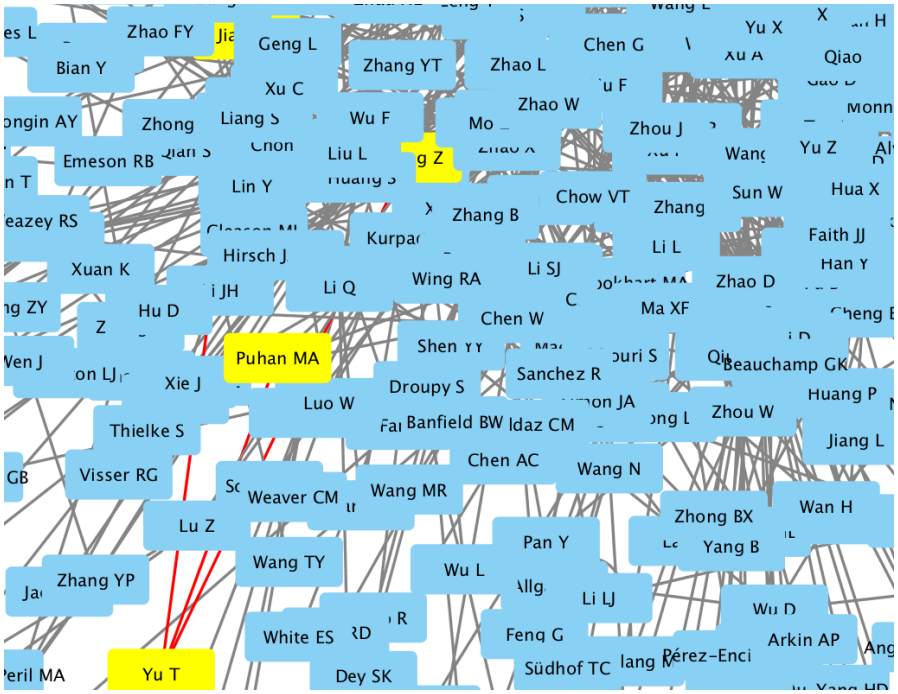

Figure 7: Comparative Physiology Bottleneck 1- Yu,T. and Connections

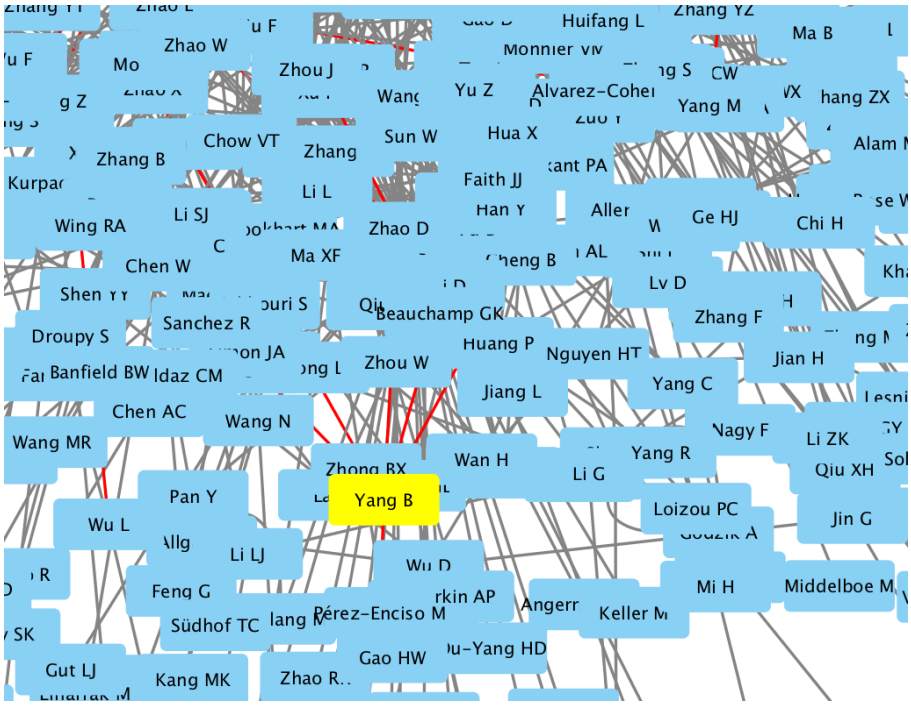

Figure 8: Comparative Physiology Bottleneck 2- Yang, B. and Connections

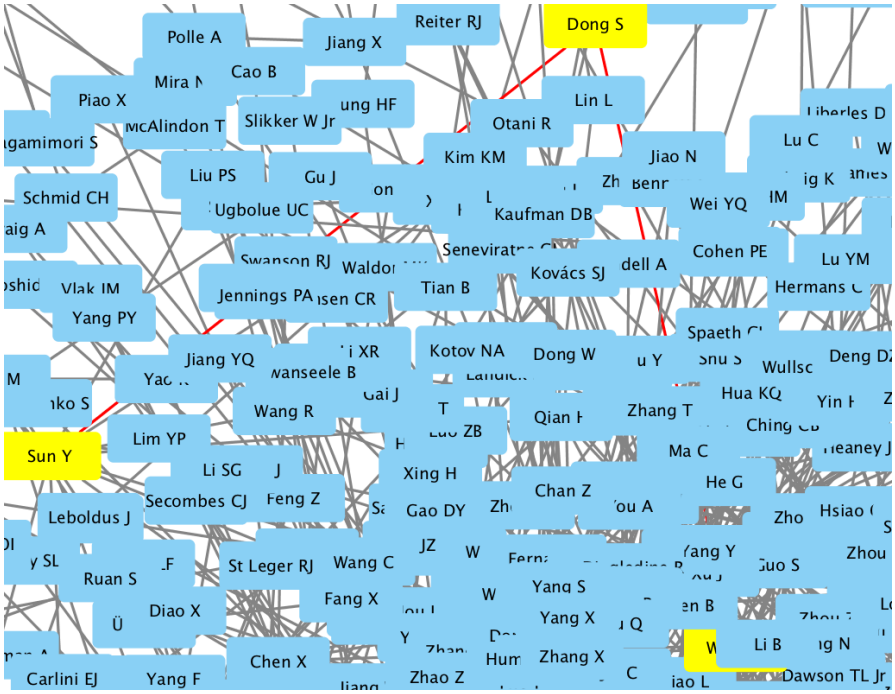

Figure 9: Comparative Physiology Bottleneck 3- Dong, S. and Connections

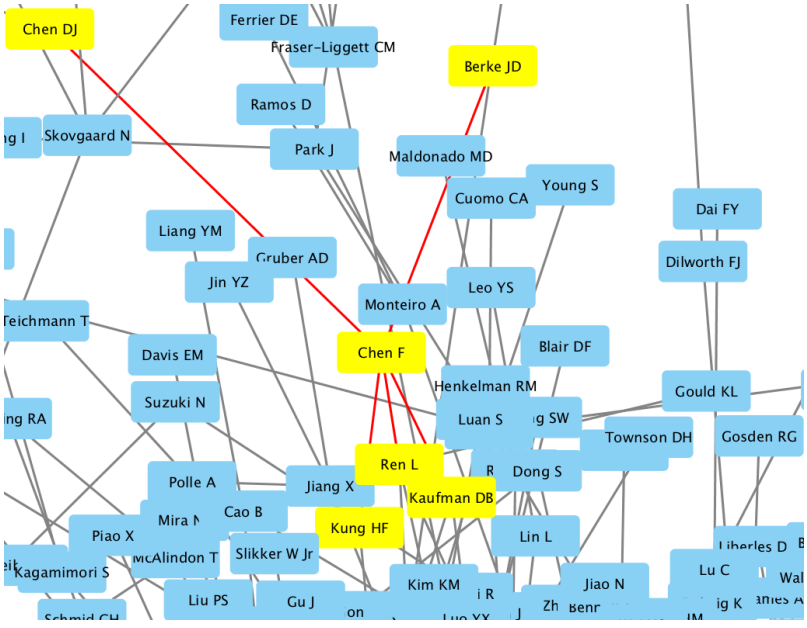

Figure 10: Comparative Physiology Bottleneck 4- Chen,F. and Connections

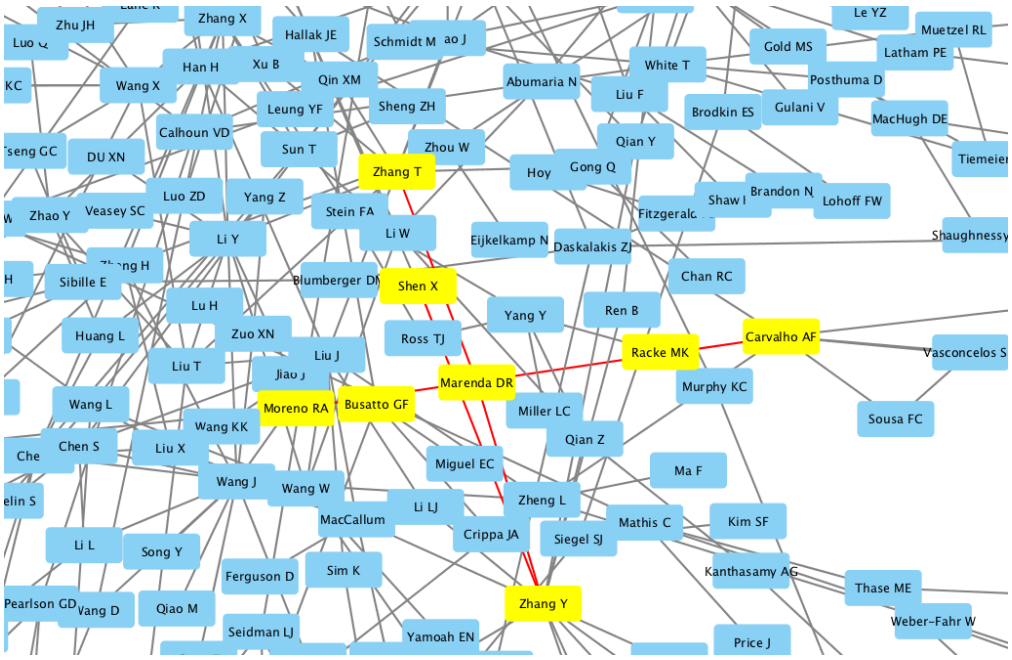

Figure 11: Neurobiology Bottleneck 1- Marenda, DR. and Connections

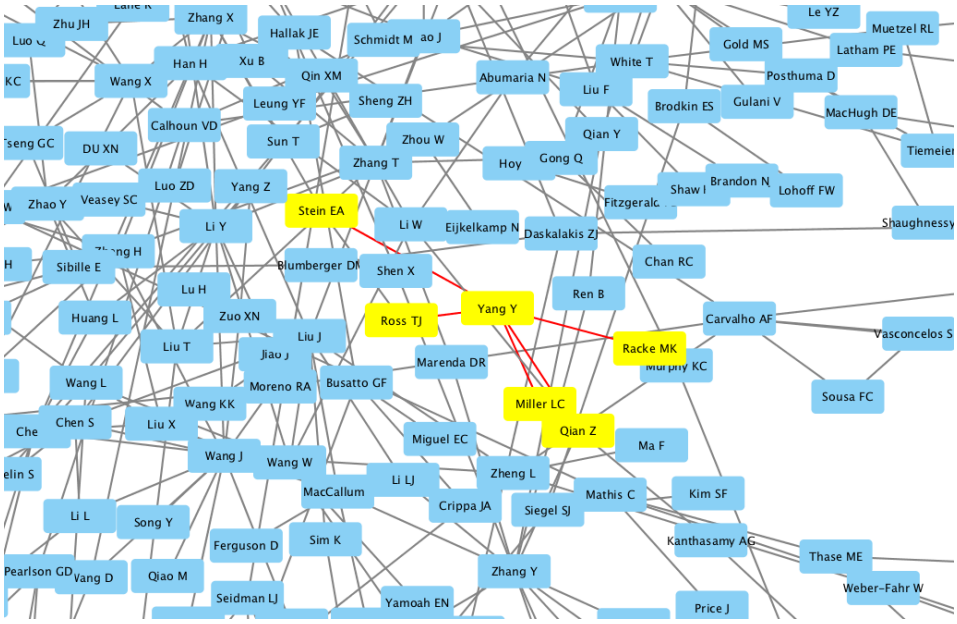

Figure 12: Neurobiology Bottleneck 2- Yang,Y. and Connections

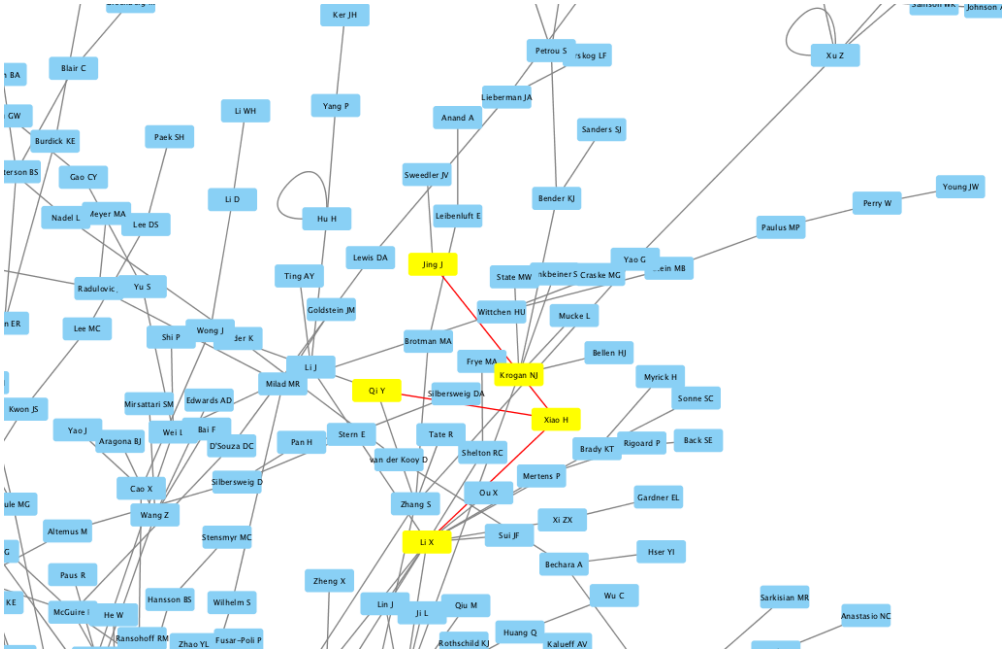

Figure 13: Neurobiology Bottleneck 3- Xiao, H. and Connections

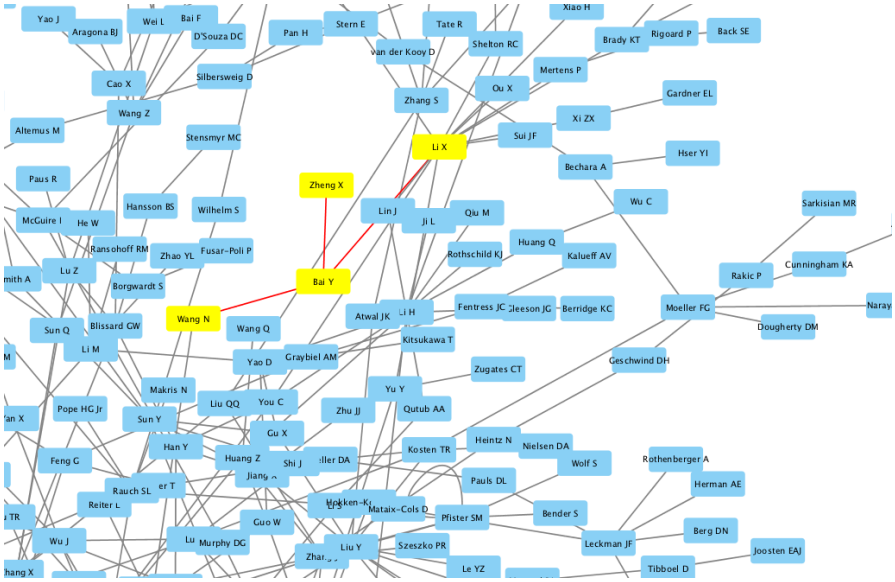

Figure 14: Neurobiology Bottleneck 4- Bai, Y. and Connections

References

Alstott, J., Bullmore, E., Plenz, D. (2014). Powerlaw: a Python package for analysis of heavy-tailed distributions. `PLoS ONE 9(1): e85777 <http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0085777>

Aynaud, T. (2010). Community API. Retrieved from https://python-louvain.readthedocs.io/en/latest/api.html.

Barabási, A., & Pósfai, M. (2016). *Network science*. Cambridge: Cambridge University Press.

Barthélemy, M. Betweenness Centrality in Large Complex Networks. Eur. Phys. J B 38, 163 (2004).

BBC. (2014) How do Search Engines work? Available from: https://www.bbc.com/bitesize/articles/ztbjq6f.

Blondel, V. D., Guillaume, J., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. Journal of Statistical Mechanics: Theory and Experiment, 2008(10). doi:10.1088/1742-5468/2008/10/p10008.

Chen H, Sharp BM. Content-rich biological network constructed by mining PubMed abstracts. BMC Bioinformatics. 2004;5:147. doi: 10.1186/1471-2105-5-147.

Clauset, A., Cosma Rohilla, S., & Newman, M. J. (2009). POWER-LAW DISTRIBUTIONS IN EMPIRICAL DATA. *SIAM Review,* 661-703. Retrieved February 6, 2019, from https://arxiv.org/pdf/0706.1062.pdf.

Cohen KB, Hunter L (2008) Getting Started in Text Mining. PLoS Comput Biol 4(1): e20. https://doi.org/10.1371/journal.pcbi.0040020.

Douglas S. M., Montelione G. T., Gerstein M. PubNet: a flexible system for visualizing literature derived networks. Genome Biology. 2005;6(9) doi: 10.1186/gb-2005-6-9-r80.R8.

Fairchild, G., & Fries, J. (2012, January 24). Lecture Notes: Social Networks: Models, Algorithms, and Applications. Retrieved February 6, 2019, from http://homepage.divms.uiowa.edu/~sriram/196/spring12/lectureNotes/Lecture3.pdf.

Graham, W. C. (2013, November 17). Random Networks. Retrieved February 6, 2019, from http://www.patternsinnature.org/Book/RandomNetworks.html.

Hagberg, A., Schult, D., and Swart, P. (2008). "Exploring network structure, dynamics, and function using NetworkX", in Proceedings of the 7th Python in Science Conference (SciPy2008), Gäel Varoquaux, Travis Vaught, and Jarrod Millman (Eds), (Pasadena, CA USA), pp. 11–15.

Klaus, A., Yu, S., & Plenz, D. (2011). Statistical Analyses Support Power Law Distributions Found in Neuronal Avalanches. *PLoS ONE, 6*(5). doi:10.1371/journal.pone.0019779.

National Center for Biotechnology Information. PubChem Database. Gene=672, https://pubchem.ncbi.nlm.nih.gov/gene/672 (accessed on Mar. 22, 2019).

NCBI. (n.d.). Biological Science Disciplines - MeSH. Retrieved February 6, 2019, from https://www.ncbi.nlm.nih.gov/mesh?term=Biological Science Disciplines).

Rzhetsky A, Seringhaus M, Gerstein MB (2009) Getting Started in Text Mining: Part Two. PLoS Comput Biol 5(7): e1000411.

Sayers E. (2010). A General Introduction to the E-utilities. In: Entrez Programming Utilities Help [Internet]. Bethesda (MD): National Center for Biotechnology Information (US). Available from: https://www.ncbi.nlm.nih.gov/books/NBK25497/.

Shaheen, J. (2017). Fitting a Power Law: Comparing the Degree Distribution of a Synthetic Network to a Collected One. Retrieved February 6, 2019, from http://www.josephshaheen.com/fitting-power-law-comparing-degree-distribution-synthetic-network-collected-one/2506?fbclid=IwAR2IY3dg-qs3_-EOz1w2EJ4CF906UGh0sGbEi3R8Ril42_aW5Bi3GBurxDU.

Singer, Emily. "Biology's Big Problem: There's Too Much Data to Handle." *Wired.com*, Wired, 3 June 2017, www.wired.com/2013/10/big-data-biology/.

Srinivasan, P. (2001). MeSHmap: a text mining tool for MEDLINE. Proceedings of the AMIA Symposium, 642–646.