SEARCH ALGORITHMS FOR PROMOTION OF NOVEL BIOMEDICAL RESEARCH

By

Partha Sarathi Mukherjee

A Thesis Submitted to the Graduate

Faculty of Rensselaer Polytechnic Institute

in Partial Fulfillment of the

Requirements for the Degree of

MASTER OF SCIENCE

Major Subject: COMPUTER SCIENCE

Examining Committee:

Boleslaw K Szymanski, Thesis Adviser

Malik Magdon-Ismail, Member

Mukkai Krishnamoorthy, Member

Rensselaer Polytechnic Institute
Troy, New York

March 2017
(For Graduation May 2017)

CONTENTS

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGMENT

I would like to thank my advisor Professor Boleslaw Szymanski for his guidance and encouragement. He gave me the chance to work on the Synergy research project and support my Master's research. His ideas and his enthusiasm spurred me forward. It has been my greatest pleasure and honor to work with him.

I would also like to thank Dr. Chris Gaeteri for his contributions. He has been a fountain of ideas which has driven this project forward. Synergy is a group project and I would not have been able to contribute as much as I did without my other team members - Konstantin Kuzmin, Xiaoyan Lu and Juntao Zhuang. I would like to thank them profusely because they have been the best team that I have been a part of.

Lastly, I would like to thank my family and friends - especially Aritra and Atriya - for their support and encouragement.

ABSTRACT

A major problem plaguing the field of biomedical research is the tendency to perform safe, incremental research.This implies that biomedical researchers tend to collaborate with other researchers who are already “close” to them in their co-authorship networks or citation networks. We propose a novel method to find new potential collaborators in the Synergy research project. The idea is to include information from molecular networks and propose researchers working on “nearby” molecules as potential collaborators. Based on this idea, we have built a software application where a biomedical researcher could input a list of molecules of her choice and find several ranked lists of potential collaborators as output. The underlying network is the Synergy network - a multilayer network formed from data in the PubMed database. This Master’s thesis describes the algorithms to find and sort these potential collaborators. These have been implemented in Java in the Synergy software application. The algorithms output the results within minutes even with tens of millions of author nodes and publication nodes in the network. Several potential collaborators are identified with sample lists of molecules in this thesis. Finally, this thesis validates the results found by comparing the probability of molecules which serve as research topics of two co-authors as being neighbors to the probability of molecules chosen at random being neighbors. It is found that the former is considerably higher than the latter.

1. INTRODUCTION

Public calls for return on investment in biological research, such as the Cancer Moonshot (Lowy & Collins, 2016) [1] and the National Alzheimer's Plan Act ("National plan to address Alzheimer's disease", 2012) [2] that targets a preventative drug by 2025, do not seek incremental scientific advances. Instead they call for transformative insights that will substantially improve patient care. Big data resources in biology may be one path to creating such insights, as seen in efforts to extract actionable research directions, such as European programs on organizing large-scale biological data (Crosswell & Thornton, 2012) [3] and collaborative endeavors across National Institutes of Health (NIH), like the Big Data to Knowledge trans-NIH initiative (Margolis et al., 2014) [4] or scientific community attempts to create large-scale metabolic models (Thiele et al., 2013) [5]. However, in the face of unprecedented data sources and public calls for transformational scientific research that makes use of these resources, the expert consensus is that the field of biology increasingly favors "safe" research that does not challenge the status quo of the field (Alberts, Kirschner, Tilghman, & Varmus, 2015) [6].

According to Smalheiser, Perkins, and Jones (2005) [7], there are two extreme cases of how collaboration is established. One is a passive approach when one side of the relationship, the supplier side, assumes a "vendor model" by providing only a minimal set of well-defined resources to the receiver who is typically the initiator of the collaboration. The other extreme case is an active model where two parties are fully engaged in the collaboration, carry equal responsibility, and receive equal credit for the work. There is also a wide range of possibilities between those extremes which can be potentially very productive but are quite difficult to initiate due to uncertainty associated with the need to agree upon many essential details. As a proposed solution, Smalheiser et al. (2005) [7] introduce a set of guidelines which describe several possible engagement levels (the minimal level and a number

Portions of this chapter previously appeared as (Kuzmin, K., Lu, X., Mukherjee, P. S., Zhuang, J., Gaiteri, C., & Szymanski, B. K. (2016). Supporting novel biomedical research via multilayer collaboration networks. *Applied Network Science*, *1*(1), 11.)

of higher levels) that can be used by a supplier and a receiver to negotiate the terms of the collaboration. Our goal is also to support potential collaborations that can emerge from the middle area between two extremes.

To support this goal, a tool - Synergy is proposed which can help to immediately reverse the trend towards incremental research, but does not require "high-risk" efforts by young scientists. This is done by mining the structure of multilayer molecular and authorship networks in search for rational innovative partnerships, which have been shown to generate high-quality scientific findings (Wuchty, Jones, & Uzzi, 2007) [8]. The current relationship of publications to molecular networks is that publications generally pertain to "popular molecules" and rarely connect to less studied ones (Rzhetsky, Foster, Foster, & Evans, 2015) [9]. A more efficient way to explore biochemical relationships entails moving away from popular topics and exploring additional subjects. Accordingly, award-winning scientists show a preference for exploring emerging topics (Uzzi, Mukherjee, Stringer, & Jones, 2013) [10] and novel relationships between them (Foster, Rzhetsky, & Evans, 2015) [11]. We utilize molecular networks to promote innovative, unbiased science, while minimizing career risk; we identify and connect researchers whose topics of study are "nearby" in molecular networks. Essentially, when molecules A and B interact biophysically, we suggest that researchers of molecule A and B should interact scientifically to explore their related interests.

By mirroring molecular organization in science, historical bias in the shape of citation networks or collaboration networks is decreased. For instance, one scientist may have negative findings related to a molecule in the context of cancer, while those results can be useful to another scientist who studies interacting molecules in schizophrenia. The links between their research, which share no overlapping keywords, can only be found through the structure of molecular networks, which connect the molecules they study. These collaboration recommendations not only make use of molecular networks, but are resistant to historical bias and can be updated as new or specialized molecular data become available. In short, by following paths in molecular data, rational scientific communities can be constructed, as researchers are alerted to the hidden potential in their existing research. This idea was first expressed in the paper by Kuzmin, Gaiteri, and Szymanski (2016) [12].

The remainder of this thesis is organized as follows. A survey of the relevant research on multilayer networks and related fields is provided in chapter 2. In chapter 3, the Synergy network and its implementation are described. The ranking methods used to rank the authors in the search results are described in chapter 4. In chapter 5, implementation of the various ranking methods is described. The results are presented in chapter 6. In chapter 7, the validation of the results is described. Finally, in chapter 8, the contribution is summarized and the future work is discussed. The core original contributions of the author of this thesis are presented in chapters 4, 5, 6 and 7, based on the sub-sections "Network analysis and mining" and "Performance evaluation" and the section "Validation" of the paper by Kuzmin et al. (2016a) [23] of which the thesis author was the sole author, while other sections of the thesis present material to which author contributed with other authors of the paper by Kuzmin et al. (2016a) [23].

2. RELATED WORK

This chapter describes the research related to the Synergy multilayer network. One of the key concepts of the Synergy network proposed by Kuzmin et al. (2016b) [12] is to establish new collaborative links between different types of entities (molecules, authors, publications, etc.) It is a fusion of two concepts - multilayer networks and collaboration networks.

2.1 Multilayer Networks

The idea of combining several different but related datasets into a single multilayer network is widely used in complex systems. De Domenico et al. (2013) [13] define multilayer networks as networks which contain entities with different sets of neighbors in each layer. The applications of multilayer networks are mostly found in sociology and social information systems. A comprehensive review by Boccaletti et al. (2014) [14] contains a detailed description of the properties and structural and dynamic organization of networks that represent different relationships as layers. Such networks have shown utility in economics, technical systems, ecology, biology and psychology. Multilayer networks originate from many experimental sources and model organisms. In many omics analyses it is now standard to project results into these networks structures, to identify the overall functional role of the results. Many free and commercial online tools are available for this purpose (For example, see the papers by Krämer, Green, Pollard, & Tugendreich (2013) [15] and Mostafavi, Ray, Warde-Farley, Grouios, & Morris, (2008) [16]).

Multilayer networks often contain nodes and edges of various types. The Synergy network is an example. Thus the Synergy network is also a Heterogeneous Information Network (HIN). Gong et al. (2012) [17] give a formal mathematical definition and describe a Social-Attribute Network(SAN) which is also an example of a heterogeneous network. HINs are widely used to model and study different

Portions of this chapter previously appeared as (Kuzmin, K., Lu, X., Mukherjee, P. S., Zhuang, J., Gaiteri, C., & Szymanski, B. K. (2016). Supporting novel biomedical research via multilayer collaboration networks. *Applied Network Science*, *1*(1), 11.)

types of networks in various fields, like social sciences, biology, medicine, and transportation, as well as across fields (e.g., scientific collaboration networks). The fact that heterogeneous networks include different types of entities and relationships in many cases significantly simplifies the process of mapping the properties of objects being studied to the attributes of network entities, as compared to homogeneous networks. For example, in the dblp computer science bibliography database [18] one node can represent either a publication or an author. Publications and authors are connected with relationships, such as a "co-author" relationship, and pairs of relationships, like "cite" and "cited-by", and "publish" and "published-by".Even though multilayer networks and HINs have different terminologies they can be essentially treated as networks with multiple types of nodes. At the same time HINs highlight different types of relationships among the nodes.

2.2 Collaboration Networks

The earliest work on collaboration networks by Newman (2001a) [19] defines these networks as networks in which a scientist/author is represented by a node. Two author nodes are joined with an unweighted edge only if they have been coauthors in a publication. Such networks can be used to explore social connections among scientists. The study of these networks includes calculating various network measures - see the papers by Newman (2001a) [19] and Newman (2001b) [20]. These measures include means and distributions of the number of edges, clustering coefficient, average distances between scientists in a network, and centrality measures like closeness and betweenness centrality.

Recent work by Bian et al. (2014) [21] goes beyond such traditional metrics. The networks themselves are slightly different — the edges are weighted based on the number of collaborative grants awarded to the relevant pair of scientists, instead of coauthorship. Multi-year grants are counted for every fiscal year. On these enhanced networks, the "leaders", or the most influential scientists are identified by various centrality measures and rank aggregation techniques. Furthermore, new collaborations are suggested using the Random Walk with Restart (RWR) algorithm. However, this research does not take into account connections between scientists who might be working on related topics but who might not have collaborated, something which is addressed in the collaborators recommended by the Synergy software application.

3. SYNERGY NETWORK AND SOFTWARE APPLICATION

This chapter describes the concept of the Synergy network and how it has been implemented with real world data from the PubMed Central database [22] and with the help of various software tools to develop the Synergy software application.

3.1 The Network Structure

The Synergy network is a multilayer network with four layers - biomedical researchers, publications, molecules with common names and unique molecules(Ensemble ID molecules). Each node in the first layer represents a biomedical researcher, henceforth referred to as the author. Each author node is connected by unweighted inter-layer edges to the nodes in the next layer - publications.

Each node in the second layer represents a publication. There are no intra-layer edges in this layer. Each publication node is connected by unweighted inter-layer edges to author nodes. These author nodes represent the authors who published the publication. Each publication node is also connected by inter-layer edges to the molecules mentioned in its abstract.

The node in the third layer represents a molecule identified by its common name mentioned in publications. These common names are alpha-numeric with a minimum length of three. However, common names of molecules are not unique. Different authors might refer to the same molecule with different common names. Moreover, different authors might have used the same common name to describe different molecules, though this is rarer. Therefore, there is the fourth layer to represent molecules uniquely.

Each node in the fourth layer represents an unique molecule. The unique identifier for each node is the Ensemble ID, which is the unique identifier for molecules in the biomedical domain. This layer has weighted intra-layer edges. Two molecules

Portions of this chapter previously appeared as (Kuzmin, K., Lu, X., Mukherjee, P. S., Zhuang, J., Gaiteri, C., & Szymanski, B. K. (2016). Supporting novel biomedical research via multilayer collaboration networks. *Applied Network Science*, *1*(1), 11.)

are connected by an intra-layer edge if they participate together in one or more biochemical reactions. The edge weight reflects the number and the importance of the biochemical reactions in which the two connected molecules participate in. A high edge weight indicates that these two nodes are "close". Each node in this layer also has unweighted inter-layer edges to the common name molecule nodes based on the common names used to refer to it in the publications.

3.2 Data Source and Formation of the Network

The data source for the Publication layer and the Author layer is the PubMed Central database [22]. Each paper was inserted as a node in the publication layer of the Synergy network. For each of the authors of the paper, at first, we check if there is an author node with identical attributes (identifying information). If there exists such a node, the publication and the pre-existing author node are connected via an inter-layer edge. If not, the author node is first created and then it is connected to the publication like before.

The common name nodes and the Ensemble ID nodes are created from the data described in the paper by Mostafavi et al. (2008) [16]. Biological molecules participate together in various biochemical reactions in the human body. Two molecules which participate together in one particular reaction has an edge connecting them in biological networks. Usually, these edges are unweighted. However, some of these reactions are more important than others, according to biomedical researchers. Thus different reactions should have different weights. The edge weight between any two molecules is thus the sum of the weights of the various reactions in which the two molecules participate together. The nodes and edges of these two layers in the Synergy network were formed from these edges and a list of mappings of Ensemble IDs to one or more common names.

3.3 Architecture

The Synergy software application consists of three layers - the Web Interface, the Middle Layer and the Graph Database Layer. These layers are described in the subsections below.

3.3.1 Web Interface

The Web Interface serves as the interface between the user and the Synergy software application. The user types in one molecule or a list of molecules as input in the Web Interface. On click of a button, she sees the various ranked lists of potential collaborators based on the several ranking algorithms as output. The web interface has been developed with Hyper-Text Markup Language(HTML)- version 5, Cascading Style Sheets (CSS) - version 3 and JavaScript - version 1.7. The webserver softwares are Nginx 1.4.6 and Tomcat 7.0.52.

3.3.2 Middle Layer

The middle layer is a Java (version 1.7) software program. It receives the input from the Web Interface, queries the Neo4j database based on the input and processes the result. The search and ranking algorithms are also implemented in this layer. This layer returns the list of authors to the Web Interface as the output. This layer is described in greater detail in chapter 5.

3.3.3 Graph Database Layer

The entire network described in section 3.1 is stored in a native graph database, Neo4j, version 3.0.8. The numbers of each type of nodes and edges in the Synergy network are given in the tables 3.1 and 3.2.

Table 3.1: Node Count.

Node Type	Count
Author	27,080,319
Publication	24,358,442
Molecule(Common name)	59,477
Molecule(Ensemble ID)	19,264

The author and publication nodes contain multiple attributes or properties. However, the molecule nodes, both common name and Ensemble ID, only have a single property - the common name and the Ensemble ID respectively. The properties and their descriptions for the author and publication nodes are given in the tables 3.3 and 3.4 respectively.

Table 3.2: **Edge Count.**

Edge Type	Count
Author-Publication	88,132,307
Publication-Molecule(Common name)	3,674,750
Molecule(Common name)-Molecule(Ensemble ID)	18,037
Molecule(Ensemble ID)-Molecule(Ensemble ID)	7,290,094

Table 3.3: **Author Node Attributes.**

Attribute	Description
AUID	Unique identifier integer assigned to node during Neo4j node creation
Author_Initials	First name initial of author
Author_ForeName	Full first name, including middle name
Author_AffiliationInfo	List of strings describing the affiliations of the author
Author_LastName	Full last name of the author

Table 3.4: **Publication Node Attributes.**

Attribute	Description
JournalISOAbbreviation	Abbreviation of the Journal by ISO 4 standard
JournalIssue_Issue	Sequential number of the journal issue
JournalIssue_PubDate_MedlineDate	Publication date of the journal
JournalISSN	Unique identifier for the journal
JournalTitle	Name of the journal
JournalIssue_Volume	Sequential number of the journal issue's volume
ArticleTitle	Title of the Article
AbstractText	Contents of the Abstract
PMID	Unique identifier for the publication

4. RANKING ALGORITHMS

Using the power of the Synergy software application, an author can enter the molecule(s) of interest and find a list of potential collaborators researching on the molecules which are "one hop" away (directly connected by a single intra-layer edge at the Ensemble ID node layer) from the molecule(s) of interest. Henceforth, the molecules which are one hop away are referred to as the "neighbors". This chapter describes how these potential collaborators are ranked.

4.1 Number of Publications on Neighbors

The number of publications on neighbors by an author is henceforth indicated by n_{PC}. The first approach is to put authors with higher n_{PC} before the authors with lower n_{PC}. This ensures that the most relevant prolific authors are at the top. However, this approach often promotes authors who might have a higher total number of publications on all molecules (this statistic is indicated by n_{TOTAL}), including the neighbors. Alternatively, it would be desirable to find authors who are more dedicated on researching the neighbors.

4.1.1 Promoting Dedicated Authors

A modification of the described ranking method promotes authors who research more exclusively on the neighbors. To accomplish this, a new measure is introduced - normalized publication count r_{PC}.Its formulation is given in equation 4.1:

$$r_{PC} = \frac{n_{PC}}{n_{TOTAL}} \tag{4.1}$$

Authors with higher r_{PC} is ranked higher than authors with r_{PC} in this approach.

This work constitutes original contributions of the author of the thesis, any portion of this chapter previously appeared as (Kuzmin, K., Lu, X., Mukherjee, P. S., Zhuang, J., Gaiteri, C., & Szymanski, B. K. (2016). Supporting novel biomedical research via multilayer collaboration networks. *Applied Network Science*, *1*(1), 11.) Sub-section "Network analysis and mining" were written by the author.

4.2 Number of Neighbors

Focusing on the number of publications on neighbors by an author has one big disadvantage. Consider this scenario - one of the neighbors is a “popular”, that is, a well-researched molecule with considerably more publications on it than other neighbors. Authors who are more prolific on the popular neighbor will have higher ranks. However, finding such authors will not promote novel research on the neighborhood. In fact, using n_{TOTAL} as a ranking criteria will produce similar results in this scenario.

The solution to this problem is to consider a different ranking criteria - the number of neighbors on which the author has published, m_{NEI}. Authors who have higher m_{NEI} are ranked higher. This ranking method achieves a higher “coverage” of the neighborhood.

4.2.1 Weighted Number of Neighbors

Using m_{NEI} as a ranking criteria ignores one aspect, not all neighbors are equal. Some neighbors are connected via higher weight edges to the searched molecule(s) than others. Authors who have published on these neighbors should be ranked higher. To account for this, a new measure is introduced, w_{NEI}, which is the sum of the weights of each edge connecting the neighbor to the searched molecule(s). Authors who have higher w_{NEI} are ranked higher.

4.3 Combining These Ideas

The previous sections in this chapter describe several ranking criteria. Each of them have their own advantages and disadvantages. Therefore, it would be ideal to combine these ranking criteria to form one composite criterion. Four of these composite criteria are given in the subsections below.

4.3.1 Non-normalized and Unweighted Ranking

This ranking criterion involves first sorting the authors based on the number of neighbors m_{NEI} on which they have researched on, as described in section 4.2. Since the weights of the edges connecting the neighbors to the searched molecule(s) are not accounted for here, the word "unweighted" is used in the name. For the authors who have researched on the same number of neighbors , the sorting is based on the number of publications on the neighbors n_{PC}, as described in section 4.1. Since the number of publications is not normalized by the total number of publications of the author, the word "non-normalized" is used in the name.

4.3.2 Normalized and Unweighted Ranking

This ranking criterion, like in the previous subsection 4.3.1, involves first sorting the authors based on the number of neighbors m_{NEI} on which they have researched on. For the authors who have researched on the same number of neighbors, the sorting is different. Here it is based on the normalized number of publications on the neighbors r_{PC}, as described in subsection 4.1.1. Since the number of publications is normalized by the total number of publications of the author, the word "normalized" is used in the name.

4.3.3 Non-normalized and Weighted Ranking

This ranking criterion involves first sorting the authors based on the weighted number of neighbors w_{NEI} on which they have researched on, as described in subsection 4.2.1. Since the weights of the edges connecting the neighbors to the searched molecule(s) are accounted for here, the word "weighted" is used in the name. For the authors who have researched on the same weighted number of neighbors , the sorting is based on n_{PC}, same as the secondary sorting criterion described in 4.3.1.

4.3.4 Normalized and Weighted Ranking

This ranking criterion, like in the previous subsection 4.3.3, involves first sorting the authors based on the weighted number of neighbors w_{NEI} on which they have researched on. For the authors who have researched on the same weighted number of neighbors , the sorting is based on r_{PC}, same as the secondary sorting criterion described in 4.3.2.

5. IMPLEMENTATION OF SEARCH AND RANKING ALGORITHMS

The Synergy network is queried to find the potential collaborators researching on neighbor molecules. This chapter describes how those query results are processed. It also describes how the ranking algorithms described in chapter 4 are implemented using the Java programming language.

5.1 Object Oriented Approach

An object oriented approach has been used to represent the node types in the Synergy network with classes. The table 5.1 gives this correspondence. Out of these, ENSGMolecule and Publication classes are mainly for future use. There is another class SynergyMiddleLayer, which has static functions and is used in an imperative paradigm. This is the class which contains the main function and hence serves as the starting point of the code execution.

Table 5.1: Node Type and Corresponding Java Class

Node Type	Java Class
Author	Author
Publication	Publication
Molecule(Common name)	Molecule
Molecule(Ensemble ID)	ENSGMolecule

5.1.1 Author Class

This class contains instance variables corresponding to the attributes of the author nodes - author ID, first name, last name, initials, affiliation. In addition, this class contains instance variables which store other important information about the

This work constitutes original contributions of the author of the thesis, any portion of this chapter previously appeared as (Kuzmin, K., Lu, X., Mukherjee, P. S., Zhuang, J., Gaiteri, C., & Szymanski, B. K. (2016). Supporting novel biomedical research via multilayer collaboration networks. *Applied Network Science*, *1*(1), 11.) Sub-section "Network analysis and mining" were written by the author.

author - her total number of publications n_{TOTAL}, the number of publications on the neighbors n_{PC}, the weighted number of neighbors on which she has published w_{NEI} and a hashtable to store the number of publications corresponding to each neighbor. Finally, this class contains the comparators which allow us to sort the author objects by the ranking criteria described in section 4.3.

5.1.2 Molecule Class

This class contains instance variables to store the following information - the common name, the hashtable to store the edge weights corresponding to the neighboring molecules and the sum of edge weights. It also contains a comparator to compare Molecule objects based on their sum of edge weights.

5.2 Flow of the Program

This section describes the flow of the program, mainly driven by the function searchAuthorsThruNeighbors. The function has two boolean arguments - one to indicate if the ranking method uses weighting for the number of molecules or not and another variable to indicate whether the ranking method uses n_{PC} or r_{PC} for the secondary ranking criteria. In addition, the function has two list arguments - a list of common name molecule strings which would serve as the search terms and a list of common name molecule strings which would be avoided if they are found as neighbors of the search terms. The function also has two numeric arguments - a floating point number which serves as the threshold for the number of molecules w_{NEI} or m_{NEI} for an author to be considered and an integer which serves as the threshold for n_{PC} for an author to be considered.

The function begins by placing all the search terms in the list of molecules to avoid. This implies that if a pair of molecules in the search terms list correspond to neighboring Ensemble IDs, none of them would be counted among the neighbors for which we want to find the researching authors.

At this point, another function findTopNeighbors is called. This function populates a hashtable ht_{MOL} where each key is a neighbor common name molecule string NEI and the value is another hashtable. In the inner hashtable ht_{WEI}, each key is a search term $SRCH$ and the value is the edge weight connecting NEI to $SRCH$ through weighted molecule (Ensemble ID) edge(s). After this function has returned, Molecule objects are created corresponding to each neighbor. Each of these objects are placed in a list. The list is sorted based on the highest total sum of edge weights over all of the ht_{WEI} hashtables, using the comparator mentioned in subsection 5.1.2. Then this list is pruned to keep only the top 30 closest neighbors.

For each of these top neighbors, the function findAuthorsFromMolecule is called. This function populates a hashtable ht_{AUTH} where each key is an author ID and the value is another hashtable ht_{PUB}. In ht_{PUB}, each key is a publication ID representing a unique publication and the value is the set of neighbor common name molecule strings which are mentioned in the abstract of that publication. This function also populates another hashtable $ht_{AUTHINFO}$ where each author ID is the key and an Author object created from her information is the corresponding value.

The next step is to traverse through the hashtable ht_{AUTH}. For every author in the hashtable, w_{NEI} or m_{NEI} and n_{PC} are calculated from the information in ht_{PUB} and ht_{MOL}. The measures calculated depends on the ranking method used (these are described in section 4.3) — for the weighted ranking methods, w_{NEI} is calculated while for the unweighted ranking methods, m_{NEI} is calculated. n_{PC} is required for all ranking methods. However, if the ranking method is normalized, n_{TOTAL} is also required to be calculated, to find r_{PC} as per equation 4.1. If the author passes the threshold arguments, the Author object obtained from the $ht_{AUTHINFO}$ hashtable is placed in a list of selected authors $li_{SELECTAUTH}$.

After the hashtable ht_{AUTH} has been processed, the list $li_{SELECTAUTH}$ is sorted using the custom comparators in the Author class, based on the ranking method used. The information of all these authors are then displayed on standard output or written to a file.

5.3 Finding the Closest Neighbors

This section describes the process by which the closest neighbors are found by the findTopNeighbors function. This function has three arguments - the search term string $SRCH$, the hashtable ht_{MOL} mentioned in section 5.2 and the set of molecules to avoid. Data structures are passed by reference in Java, hence the existing information in the hashtable ht_{MOL} is preserved as it is called repeatedly for each of the search terms.

The function creates a Neo4j query to find all the common name molecules which are “directly” connected to $SRCH$. These connections would be three hops away - $SRCH$ would be connected via inter-layer edge(s) to one or more Ensemble ID molecules, which in turn would have connections to other Ensemble ID molecules via weighted intra-layer edges. These neighbor Ensemble ID molecules would be connected via inter-layer edges to other common name molecules. These common name molecules are the “neighbors” mentioned in chapter 4. The edge weight connecting the Ensemble ID nodes are also obtained through this query. As it is possible for $SRCH$ to be represented by more than one Ensemble ID molecule, the connection between $SRCH$ and its neighbor NEI could be through multiple weighted edges. In this case, the sum of these edge weights SUM_{EDGEWT} is considered. Thus, in the hashtable ht_{MOL}, corresponding to the key NEI, the value hashtable ht_{WEI} would have an entry with $SRCH$ as the key and SUM_{EDGEWT} as the value. The result of the Neo4j query is processed to extract the information described above.

5.4 Finding the Authors

This section describes the process by which the authors researching on a particular neighbor are found by the findAuthorsFromMolecule function. This function has three arguments - the two hashtables ht_{AUTH} and $ht_{AUTHINFO}$ and the neighbor common name string NEI. This function creates a Neo4j query to find all the authors and their information who research on NEI. These authors are two hops away from NEI - NEI is connected via inter-layer edges to publications which mention it and these publications, in turn, are connected via inter-layer edges to their publishing authors. The result of the Neo4j query is processed to extract the information required to populate the hashtables ht_{AUTH} and $ht_{AUTHINFO}$.

Another citation for the bibliography:[**?**]

6. RESULTS

This chapter presents the Synergy software application's results and their analysis. The Synergy software application has been run with several molecule lists as input.

6.1 Tables and Discussion - First Input List

One of these lists $INP1$ is given in the table 6.1, along with the corresponding Emsemble IDs. The Ensemble IDs corresponding to common name molecules $LUBB3$ and $U1 - C$ are not found. Therefore, there are no authors who research on the neighbors of these molecules in the results.

The top five authors recommended for collaboration using the ranking methods - non-normalized unweighted, normalized unweighted, non-normalized weighted and normalized weighted are given in the tables 6.2, 6.3, 6.4 and 6.5 respectively. The results are similar for the four ranking methods, with identical top three authors, even though Dr.Juri Rappsilber is second in the normalized methods but third in the non-normalized methods. Since the number or weight of neighbors is the primary ranking criteria, the two weighted methods are even closer with identical top four authors. The same is true for the unweighted methods.

The ranking methods have been compared through a plot 6.1. The results of the three ranking methods - normalized unweighted, non-normalized weighted and normalized weighted have been plotted against the results of the non-normalized unweighted ranking method. This has been accomplished by identifying the top 100 authors in the non-normalized unweighted ranking method and then finding and plotting their ranks in the other methods, if the ranks are within 210.

As in the tables, it is found that the two unweighted methods have some

This work constitutes original contributions of the author of the thesis, any portion of this chapter previously appeared as (Kuzmin, K., Lu, X., Mukherjee, P. S., Zhuang, J., Gaiteri, C., & Szymanski, B. K. (2016). Supporting novel biomedical research via multilayer collaboration networks. *Applied Network Science*, *1*(1), 11.) Sub-section "Performance evaluation" were written by the author.

Table 6.1: Sample Input List of Molecules $INP1$

Input Common Name	Corresponding Ensemble ID(s)
ANXA5	ENSG00000164111
CD44	ENSG00000026508
DNM3	ENSG00000197959
EPB41L3	ENSG00000082397
LMNA	ENSG00000160789
LUBB3	None
MSN	ENSG00000147065
PLCD1	ENSG00000187091
PLEC	ENSG00000178209
PPP1R7	ENSG00000115685
PTRHD1	ENSG00000184924
RTN4	ENSG00000115310
SFRP1	ENSG00000104332
SNRNP70	ENSG00000104852
SNRPB	ENSG00000125835
SNRPN	ENSG00000128739
U1-A	ENSG00000077312
U1-C	None

Table 6.2: Top Five Authors Recommended as Collaborators with $INP1$(Non-normalized Unweighted Ranking Method).

Name	Neighbor Molecules (Number Of publications)	Number Of Neighbor Molecules	Publication Count On Neighbor Molecules n_{PC}
Gideon Dreyfuss	EIF4A3(2), MAGOH(6), SNRPD1(1), SNRPD3(1), SNRPE(1)	5	7
Matthias Mann	EIF4A3(1), MAGOH(1), SNRPD1(1), SNRPD3(1), SNRPE(1)	5	2
Juri Rappsilber	EIF4A3(1), MAGOH(1), SNRPD1(1), SNRPD3(1), SNRPE(1)	5	2
Edouard Bertrand	EIF4A3(1), MAGOH(2), SNRPD1(1), SNRPD3(1)	4	3
Martin E Schwab	LINGO1(2), ROCK2(1), RTN4R(8)	3	11

Table 6.3: Top Five Authors Recommended as Collaborators with $INP1$(Normalized Unweighted Ranking Method).

Name	Neighbor Molecules (Number Of publications)	Number Of Neighbor Molecules	Normalized Publication Count r_{PC}
Gideon Dreyfuss	EIF4A3(2), MAGOH(6), SNRPD1(1), SNRPD3(1), SNRPE(1)	5	0.13
Juri Rappsilber	EIF4A3(1), MAGOH(1), SNRPD1(1), SNRPD3(1), SNRPE(1)	5	0.03
Matthias Mann	EIF4A3(1), MAGOH(1), SNRPD1(1), SNRPD3(1), SNRPE(1)	5	0.01
Edouard Bertrand	EIF4A3(1), MAGOH(2), SNRPD1(1), SNRPD3(1)	4	0.04
Amelie K Gubitz	SNRPD1(1), SNRPD3(1), SNRPE(1)	3	1.00

Table 6.4: Top Five Authors Recommended as Collaborators with $INP1$(Non-normalized Weighted Ranking Method).

Name	Neighbor Molecules (Number Of publications)	Sum Of Edge Weights On Published Neighbors	Publication Count On Neighbor Molecules n_{PC}
Gideon Dreyfuss	EIF4A3(2), MAGOH(6), SNRPD1(1), SNRPD3(1), SNRPE(1)	3.19×10^{-2}	7
Matthias Mann	EIF4A3(1), MAGOH(1), SNRPD1(1), SNRPD3(1), SNRPE(1)	3.19×10^{-2}	2
Juri Rappsilber	EIF4A3(1), MAGOH(1), SNRPD1(1), SNRPD3(1), SNRPE(1)	3.19×10^{-2}	2
Francis S Collins	LMNB1(1), LMNB2(1), MSH4(1)	3.06×10^{-2}	2
Catherine Tomasetto	EIF4A3(2), MAGOH(4), ROCK2(1)	2.88×10^{-2}	6

Table 6.5: Top Five Authors Recommended as Collaborators with $INP1$(Normalized Weighted Ranking Method).

Name	Neighbor Molecules (Number Of publications)	Sum Of Edge Weights On Published Neighbors	Normalized Publication Count r_{PC}
Gideon Dreyfuss	EIF4A3(2), MAGOH(6), SNRPD1(1), SNRPD3(1), SNRPE(1)	3.19×10^{-2}	0.13
Juri Rappsilber	EIF4A3(1), MAGOH(1), SNRPD1(1), SNRPD3(1), SNRPE(1)	3.19×10^{-2}	0.03
Matthias Mann	EIF4A3(1), MAGOH(1), SNRPD1(1), SNRPD3(1), SNRPE(1)	3.19×10^{-2}	0.01
Francis S Collins	LMNB1(1), LMNB2(1), MSH4(1)	3.06×10^{-2}	0.01
Fabien Alpy	EIF4A3(2), MAGOH(4), ROCK2(1)	2.88×10^{-2}	0.25

similarity. The blue points indicating the normalized unweighted ranks are present in two sections of the plot - bottom left and top right, indicating a positive relationship. However, the relationship is not exactly linear. The number of blue points is also far lower than 100 implying that not all of the top 100 authors in the non-normalized unweighted ranking method are present within the top 200 list of authors in the normalized weighted ranking method.

On the other hand, the two weighted methods produce quite different ranks. The red points, indicating the ranks in normalized weighted ranking method, are all over the plot, indicating little to no positive relationship with the non-normalized unweighted ranking method. The green points, indicating non-normalized weighted ranking method, are also all over the plot. In contrast to the other two ranks, their number is lower, indicating that fewer of the top 100 authors in the non-normalized unweighted ranking method are placed among the top 200 in the non-normalized weighted ranking method. Thus the two non-normalized ranking methods are drastically different from each other due to the weighting factor.

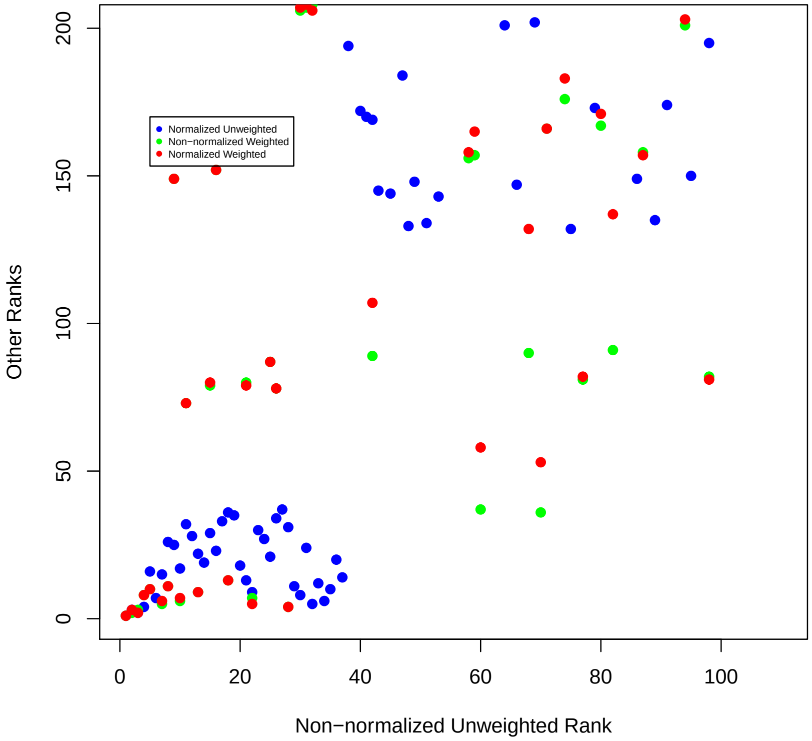

Figure 6.1: Plot of the normalized unweighted, non-normalized weighted and normalized weighted ranks of the top 100 authors from the non-normalized unweighted ranking method — $INP1$.

6.2 Tables and Discussion - Second Input List

Another input list $INP2$ is given in the table 6.6. The Ensemble ID corresponding to the common name $DA2IP$ is not found.

The top five authors recommended for collaboration with this input list are given in the tables 6.7, 6.8, 6.9 and 6.10. The results show more variability for $INP2$. Only one author is consistently present in all the four ranking methods. However, the weighted methods are much closer, with the identical five authors and their identical ranks. The unweighted methods are also closer but not as much, with the identical top three authors and swapped ranks.

The ranking methods have also been compared through a plot 6.2. The plot has been generated using the method described in 6.1 but with the results for $INP2$.

Table 6.6: Sample Input List of Molecules $INP2$

Input Common Name	Corresponding Ensemble ID(s)
CCDC85C	ENSG00000205476
CIC	ENSG00000079432
CSRP1	ENSG00000159176
DA2IP	None
FAM63A	ENSG00000143409
FURIN	ENSG00000140564
HMG20B	ENSG00000064961
IGFBP5	ENSG00000115461
ISYNA1	ENSG00000105655
KIF1C	ENSG00000129250
PADI2	ENSG00000117115
SLC38A2	ENSG00000134294
SNAP25	ENSG00000132639
STX1A	ENSG00000106089
STXBP3	ENSG00000116266
SV2B	ENSG00000185518
SYT1	ENSG00000067715
SYT12	ENSG00000173227
VGF	ENSG00000128564
ZBTB47	ENSG00000114853

The number of blue points, indicating the normalized unweighted ranking method, is closer to 100 here. The positive relationship between the two unweighted methods is more apparent here. However, green points, indicating non-normalized weighted ranks, are extremely few in number. The number of red points, indicating the normalized weighted ranking method, is considerably lower than the number of blue points, unlike for $INP1$. The red points are also all over the map. Thus the weighted methods, especially the non-normalized one, have produced dramatically different ranks, when compared with similar results for $INP1$. One probable reason for this phenomenon is that the popularity of the neighbors of the molecules listed in $INP2$ - most of the n_{PC} values are greater than 10 for the top 5 authors for $INP2$ with different ranking methods, while for $INP1$, it is mostly lower than 10. The change in the weighting scheme of the ranking method thus brings new prolific authors into the ranked list, especially for the non-normalized ones.

Table 6.7: Top Five Authors Recommended as Collaborators with $INP2$**(Non-normalized Unweighted Ranking Method).**

Name	Neighbor Molecules (Number Of publications)	Number Of Neighbor Molecules	Publication Count On Neighbor Molecules n_{PC}
Thomas C Sudhof	RAB3A(11), SLC17A7(1), STX1B(1), STXBP1(4), VAMP2(7)	5	24
Yang Shi	HDAC2(1), IGF1(1), KDM1A(6), PHF21A(3), RCOR1(2)	5	10
Gudrun Ahnert-Hilger	SLC17A7(2), SLC18A2(2), STX1B(1), STXBP1(1), VAMP2(2)	5	8
Robert H Edwards	SLC17A7(15), SLC18A2(9), SSLC18A3(1), VAMP2(1)	4	24
Bruno Giros	HDAC2(1), SLC17A7(4), SLC18A2(4), SLC18A3(2)	4	11

Table 6.8: Top Five Authors Recommended as Collaborators with $INP2$**(Normalized Unweighted Ranking Method).**

Name	Neighbor Molecules (Number Of publications)	Number Of Neighbor Molecules	Normalized Publication Count r_{PC}
Gudrun Ahnert-Hilger	SLC17A7(2), SLC18A2(2), STX1B(1), STXBP1(1), VAMP2(2)	5	0.18
Thomas C Sudhof	RAB3A(11), SLC17A7(1), STX1B(1), STXBP1(4), VAMP2(7)	5	0.12
Yang Shi	HDAC2(1), IGF1(1), KDM1A(6), PHF21A(3), RCOR1(2)	5	0.05
Noelia Fernandez -Castillo	CPLX1(1), RAB3A(1), STXBP1(1), VAMP2(1)	4	1.00
Bernard Lakowski	HDAC2(1), KDM1A(1), PHF21A(1), RCOR1(1)	4	1.00

Table 6.9: Top Five Authors Recommended as Collaborators with $INP2$**(Non-normalized Weighted Ranking Method).**

Name	Neighbor Molecules (Number Of publications)	Sum Of Edge Weights On Published Neighbors	Publication Count On Neighbor Molecules n_{PC}
Jerome I Rotter	IGF1(1), IGFALS(1), KDM1A(1), NOTCH3(1)	3.77×10^{-2}	4
Romano Regazzi	CPLX1(1), RAB3A(3), SLC17A7(1), VAMP2(4)	2.80×10^{-2}	9
Gudrun Ahnert-Hilger	SLC17A7(2), SLC18A2(2), STX1B(1), STXBP1(1), VAMP2(2)	2.73×10^{-2}	8
Robert H Edwards	SLC17A7(15), SLC18A2(9), SSLC18A3(1), VAMP2(1)	2.72×10^{-2}	24
Bruce M Psaty	IGF1(2), IGFALS(1), PAPPA2(1), RCOR1(1)	2.70×10^{-2}	5

Table 6.10: Top Five Authors Recommended as Collaborators with $INP2$**(Normalized Weighted Ranking Method).**

Name	Neighbor Molecules (Number Of publications)	Sum Of Edge Weights On Published Neighbors	Normalized Publication Count r_{PC}
Jerome I Rotter	IGF1(1), IGFALS(1), KDM1A(1), NOTCH3(1)	3.77×10^{-2}	0.01
Romano Regazzi	CPLX1(1), RAB3A(3), SLC17A7(1), VAMP2(4)	2.80×10^{-2}	0.22
Gudrun Ahnert-Hilger	SLC17A7(2), SLC18A2(2), STX1B(1), STXBP1(1), VAMP2(2)	2.73×10^{-2}	0.18
Robert H Edwards	SLC17A7(15), SLC18A2(9), SSLC18A3(1), VAMP2(1)	2.72×10^{-2}	0.28
Bruce M Psaty	IGF1(2), IGFALS(1), PAPPA2(1), RCOR1(1)	2.70×10^{-2}	0.01

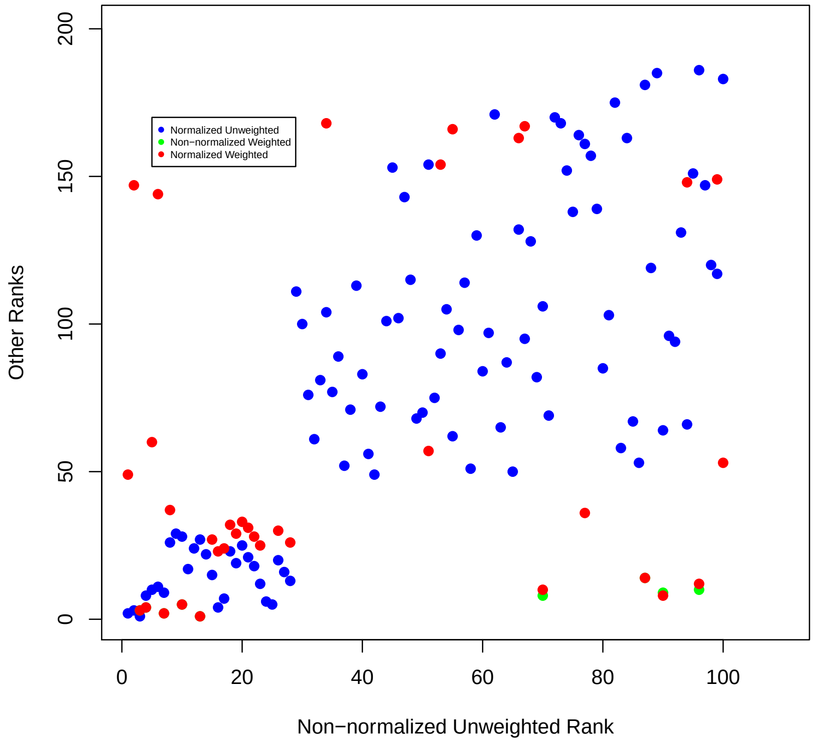

Figure 6.2: Plot of the normalized unweighted,non-normalized weighted and normalized weighted ranks of the top 100 authors from the non-normalized unweighted ranking method — $INP2$.

6.3 Running Time

The most time consuming step in this project is running a Neo4j query. Normalized methods, which require the calculation of n_{TOTAL} for every author in the ranked list with a separate query for each, take much more time than non-normalized methods. The other factor affecting run time is the number of authors in the ranked list, which decreases for high thresholds and less popular neighbors. The running times with the four ranking methods for $INP1$ and $INP2$ are compared with bar graphs in Figures 6.3 and 6.4 respectively.

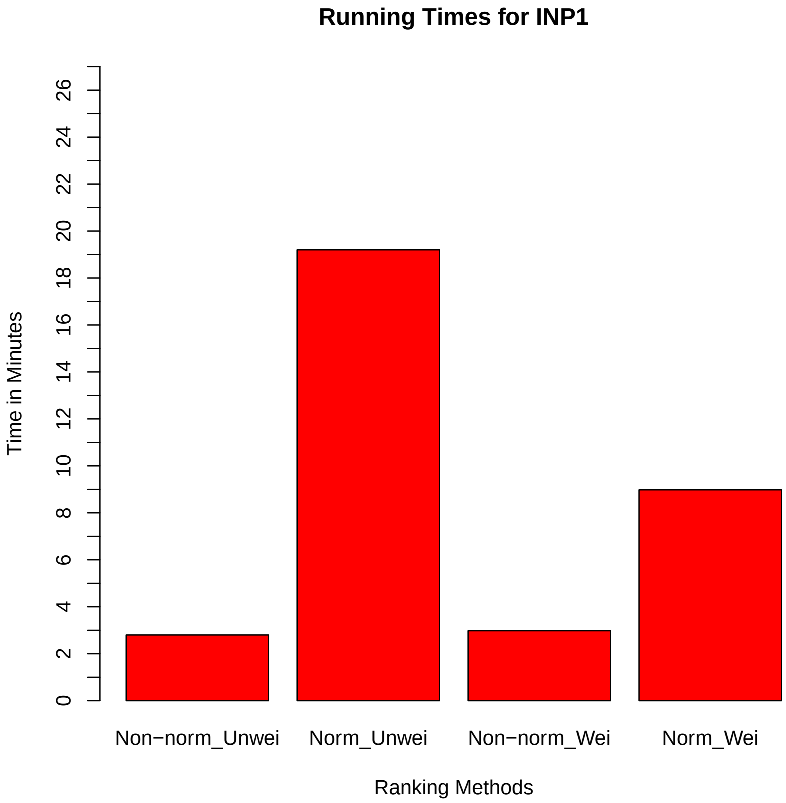

Figure 6.3: Comparison of the running time produced by the non-normalized unweighted, normalized unweighted, non-normalized weighted and normalized weighted ranking methods with $INP1$.

The asymptotic running time is calculated for the processing performed on the results returned by Neo4j - the neighbors of a searched molecule or the authors researching on a neighbor molecule. The number of neighbors selected is 30. The size

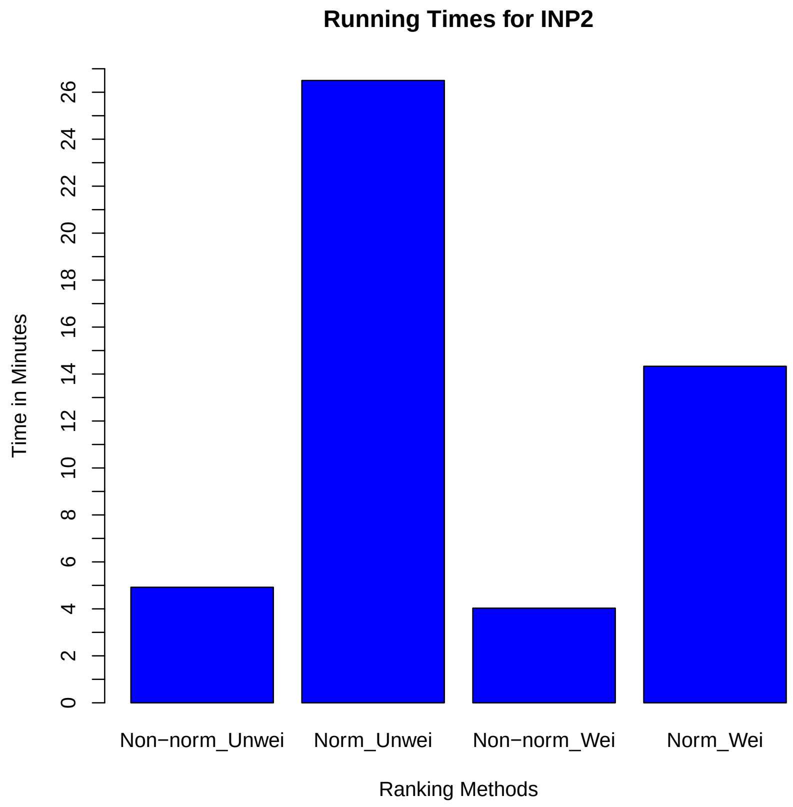

Figure 6.4: Comparison of the running time produced by the non-normalized unweighted, normalized unweighted, non-normalized weighted and normalized weighted ranking methods with $INP2$.

of the input molecule list rarely exceeds 100. The number of Ensemble ID molecule nodes connected to a common name molecule node are mostly one and never exceed five in the worst case. The number of common name molecule nodes connected to an Ensemble ID molecule node, on the other hand, are often more than one but their number never exceeds 16 in the worst case. Hence, all these small constants are ignored in the asymptotic running time calculations. The asymptotic running time is thus a function of the maximum number of keys in the hashtables ht_{MOL} and ht_{AUTH}. These numbers are the number of common name molecule nodes N_{CM} and the maximum number of author nodes N_A connected to a common name molecule node through the inter-layer edges (author-publication) and (publication-common name molecule) respectively. Therefore, the running time is O(N_{CM} log(N_{CM}) + $N_A$log(N_A)). The logarithmic terms originate from the sorting operations performed on the lists containing the molecule objects from ht_{MOL} and the author objects from ht_{AUTH} respectively.

7. VALIDATION

The collaborators recommended by the Synergy software application are often different from the collaborators found by searching the co-authorship or citation networks. This should promote innovative research. However, it would also be desirable to maintain continuity with the current research trajectories of the authors, to promote innovation with minimal disruption. This chapter describes how the collaborators recommended by the Synergy software application helps to achieve this goal of continuity and thus maintains a balance between the two different, often competing goals.

7.1 Molecular Connection Comparison

The Synergy software application recommends authors based on their publications on directly connected molecules or neighbors. Therefore, to establish continuity, the direct connectivity of the molecules which are the subjects of research for authors and their coauthors are explored in this section. To put this concretely, the hypothesis is that if $Author_A$ researches on molecule $SRCH$ and her coauthor $Coauthor_A$ researches on MOL_B when she is not collaborating with $Author_A$, the probability of $SRCH$ and MOL_B being directly connected is greater than the probability of two molecules chosen at random are directly connected. Note that the term "connectivity" here refers to connectivity at the Ensemble ID molecule node layer.

7.1.1 Connectivity of Molecules Chosen Randomly

To find the connectivity of molecules chosen at random, 1000 common name molecules are selected at random. They are then divided into two groups of 500 each. A pair is formed by selecting one molecule from one group and the other

This work constitutes original contributions of the author of the thesis, any portion of this chapter previously appeared as (Kuzmin, K., Lu, X., Mukherjee, P. S., Zhuang, J., Gaiteri, C., & Szymanski, B. K. (2016). Supporting novel biomedical research via multilayer collaboration networks. *Applied Network Science*, *1*(1), 11.) Section “Validation” were written by the author.

molecule from the other group. There are 250,000 possible pairs. Out of them, 10,300 pairs have an edge between them at the Ensemble ID molecule node layer.

7.1.2 Connectivity of Molecules from Author-Coauthor Pairs

To find the connectivity of molecules researched by authors and their coauthors, 1000 authors with more than a threshold number of publications, five, are selected at random. For each such author $Author_A$, a set of five molecules Set_A are selected on which she has published the highest number of papers. Five of her collaborators are also selected with whom she has published the highest number of publications. For each of these collaborators $Coauthor_A$, a set of five molecules Set_C are selected on which $Coauthor_A$ has published the highest number of papers, excluding the papers $Coauthor_A$ has published with $Author_A$. At this point, Set_A and Set_C are compared. If there are molecules present in both sets, they are removed from both sets. Then, the number of possible pairs and the actual number of edges for these pairs are computed where each pair have one molecule from Set_A and the other from Set_C. There are 14,760 possible pairs. Out of them, 1,735 pairs have an edge between them.

7.2 Results

The numbers are presented in a tabular form in table 7.1. When the Fisher's exact test is performed, the odds ratio is found to be 3.09 at the 95% confidence interval with a p value less than 2.2×10^{-16}. Thus molecules researched by an author and her coauthors are significantly more likely to be neighbors than molecules chosen at random.

Table 7.1: Validation Contingency Table.

	Random Molecule Pairs	Author-Coauthor Molecule Pairs
Non Neighbors	239,670	13,025
Neighbors	10,330	1,735

8. FUTURE WORK AND CONCLUSION

Future work on the Synergy software application and the Synergy network would involve enhancements in several areas. One of them is the disambiguation of author nodes. Many authors often have the same first name and last name. The PubMed database might not have the affiliation or other identifying information to distinguish these authors. Consequently, these authors have been fused to a single author node in the Synergy network. The current and future research is focusing on how to disambiguate or separate such composite author nodes into their constituents.

Another enhancement would be to include another network as a layer to the Synergy network - the disease network. A disease node would be connected to molecules via inter-layer edges if the molecules are affected by the disease or otherwise interact with the disease. The disease node could also be connected to publications via inter-layer edges if the publications focus on the disease. A further enhancement would be the reduction in the running time by targeting the bottleneck - time taken to run a query on the graph database Neo4j.

To summarize, this thesis proposes algorithms to find and sort potential collaborators using the Synergy multilayer network. The idea is to find authors who are working on molecules directly connected to the molecule of interest of an interested author. The algorithms run within minutes even with tens of millions of authors and publications. These algorithms often help discover new authors who are not found by searching traditional co-authorship or citation networks. Collaborations with authors found by this novel recommendation technique is likely to promote innovative research as opposed to safe, incremental research. But collaborating with these authors also help maintain the continuity in the research of the interested authors because molecules researched by an author and her co-authors are more likely to be neighbors.

Portions of this chapter previously appeared as (Kuzmin, K., Lu, X., Mukherjee, P. S., Zhuang, J., Gaiteri, C., & Szymanski, B. K. (2016). Supporting novel biomedical research via multilayer collaboration networks. *Applied Network Science*, *1*(1), 11.

LITERATURE CITED

[1] Lowy, D. R., & Collins, F. S. (2016). Aiming high—changing the trajectory for cancer. *New England Journal of Medicine*, *374*(20), 1901-1904.

[2] DHSS (2012) *National plan to address Alzheimer's disease, U.S. Department of Health and Human Services.* Retrieved from https://aspe.hhs.gov/national-plans-address-alzheimers-disease. Accessed 2 Nov 2016

[3] Crosswell, L. C., & Thornton, J. M. (2012). ELIXIR: a distributed infrastructure for European biological data. *Trends in biotechnology*, *30*(5), 241.

[4] Margolis, R., Derr, L., Dunn, M., Huerta, M., Larkin, J., Sheehan, J., ... & Green, E. D. (2014). The National Institutes of Health's Big Data to Knowledge (BD2K) initiative: capitalizing on biomedical big data. *Journal of the American Medical Informatics Association*, *21*(6), 957-958.

[5] Thiele, I., Swainston, N., Fleming, R. M., Hoppe, A., Sahoo, S., Aurich, M. K., ... & Thorleifsson, S. G. (2013). A community-driven global reconstruction of human metabolism. *Nature biotechnology*, *31*(5), 419-425.

[6] Alberts, B., Kirschner, M. W., Tilghman, S., & Varmus, H. (2015). Opinion: Addressing systemic problems in the biomedical research enterprise. *Proceedings of the National Academy of Sciences*, *112*(7), 1912-1913.

[7] Smalheiser, N. R., Perkins, G. A., & Jones, S. (2005). Guidelines for negotiating scientific collaboration. *PLoS Biol*, *3*(6), e217.

[8] Wuchty, S., Jones, B. F., & Uzzi, B. (2007). The increasing dominance of teams in production of knowledge. *Science*, *316*(5827), 1036-1039.

[9] Rzhetsky, A., Foster, J. G., Foster, I. T., & Evans, J. A. (2015). Choosing experiments to accelerate collective discovery. *Proceedings of the National Academy of Sciences*, *112*(47), 14569-14574.

[10] Uzzi, B., Mukherjee, S., Stringer, M., & Jones, B. (2013). Atypical combinations and scientific impact. *Science*, *342*(6157), 468-472.

[11] Foster, J. G., Rzhetsky, A., & Evans, J. A. (2015). Tradition and innovation in scientists' research strategies. *American Sociological Review*, *80*(5), 875-908.

[12] Kuzmin, K., Gaiteri, C., & Szymanski, B. K. (2016, January). Synergy Landscapes: A multilayer network for collaboration in biological research. In *International Conference and School on Network Science* (pp. 205-212). Springer International Publishing.

[13] De Domenico, M., Solé-Ribalta, A., Cozzo, E., Kivelä, M., Moreno, Y., Porter, M. A., ... & Arenas, A. (2013). Mathematical formulation of multilayer networks. *Physical Review X*, *3*(4), 041022.

[14] Boccaletti, S., Bianconi, G., Criado, R., Del Genio, C. I., Gómez-Gardenes, J., Romance, M., ... & Zanin, M. (2014). The structure and dynamics of multilayer networks. *Physics Reports*, *544*(1), 1-122.

[15] Krämer, A., Green, J., Pollard, J., & Tugendreich, S. (2013). Causal analysis approaches in ingenuity pathway analysis (ipa). *Bioinformatics*, btt703.

[16] Mostafavi, S., Ray, D., Warde-Farley, D., Grouios, C., & Morris, Q. (2008). GeneMANIA: a real-time multiple association network integration algorithm for predicting gene function. *Genome biology*, *9*(1), S4.

[17] Gong, N. Z., Xu, W., Huang, L., Mittal, P., Stefanov, E., Sekar, V., & Song, D. (2012, November). Evolution of social-attribute networks: measurements, modeling, and implications using google+. In *Proceedings of the 2012 ACM conference on Internet measurement conference* (pp. 131-144). ACM.

[18] DBLP (2016) *dblp: open bibliographic information on major computer science journals and proceedings* Retrieved from http://dblp.uni-trier.de/. Accessed 12 July 2016

[19] Newman, M. E. (2001). The structure of scientific collaboration networks. *Proceedings of the National Academy of Sciences*, *98*(2), 404-409.

[20] Newman, M. E. (2001). Scientific collaboration networks. II. Shortest paths, weighted networks, and centrality. *Physical review E*, *64*(1), 016132.

[21] Bian, J., Xie, M., Topaloglu, U., Hudson, T., Eswaran, H., & Hogan, W. (2014). Social network analysis of biomedical research collaboration networks in a CTSA institution. *Journal of biomedical informatics*, *52*, 130-140.

[22] PubMed Database (2017) Retrieved from https://www.ncbi.nlm.nih.gov/pubmed. Accessed 20 Mar 2017

[23] Kuzmin, K., Lu, X., Mukherjee, P. S., Zhuang, J., Gaiteri, C., & Szymanski, B. K. (2016). Supporting novel biomedical research via multilayer collaboration networks. *Applied Network Science*, *1*(1), 11.