Disease Network Construction and Analysis: A case study of Alzheimer's Disease

A thesis submitted during 2023 to the University of Hyderabad in partial fulfillment of the award of a Ph.D. degree in School of Computer and Information Sciences

by

Shailendra Sahu

Reg. No: 16MCPC15

School of Computer and Information Sciences
University of Hyderabad
(P.O.) Central University, Gachibowli,
Hyderabad – 500046
Telangana, India

2023

CERTIFICATE

This is to certify that the thesis entitled **"Disease Network Construction and Analysis: A case study of Alzheimer's Disease"** submitted by **Shailendra Sahu** bearing **Reg. No: 16MCPC15** in partial fulfillment of the requirements for the award of **Doctor of Philosophy** in **Computer Science** is a bonafide work carried out by him under my supervision and guidance.

The thesis is free from plagiarism and has not been submitted previously in part or in full to this or any other University or Institution for the award of any degree or diploma.

The student has the following publications before submission of the thesis for adjudication and has produced evidence for the same in the form of acceptance letter or the reprint in the relevant area of his research:

1. **Shailendra Sahu**, Pankaj Singh Dholaniya, & Rani, T.S. **Identifying the candidate genes using co-expression, GO, and machine learning techniques for Alzheimer's disease**. *Network Modeling Analysis in Health Informatics and Bioinformatics*, Volumne 11, Pages 10, 2022, ISSN: 2192-6670. DOI: 10.1007/s13721-021-00349-9. (Indexed in ESCI, SCOPUS, UGC-CARE. IF: 2.05). The work reported in this publication appears in Chapter 3.

2. **Shailendra Sahu**, T. Sobha Rani, **A neighbour-similarity based community discovery algorithm**, *Expert Systems with Applications*, Volume 206, Pages 117822, 2022, ISSN: 1873-6793. DOI: 10.1016/j.eswa.2022.117822. The work reported in this publication appears in Chapter 4.

Further the student has passed the following courses towards fulfilment of course work requirement for Ph.D.

Course Code	**Name of the Course**	**Credits**	**Pass/Fail**
CS-801	Data Structures and Algorithms	4	Pass
CS-802	Operating Systems and Programming	4	Pass
AI-840	Bioinformatics	4	Pass
AI-852	Learning & Reasoning	4	Pass

Dr. T. Sobha Rani
(Supervisor)
School of Computer &
Information Sciences
University of Hyderabad
Hyderabad - 500046, India

Dr. Atul Negi
(Dean)
School of Computer &
Information Sciences
University of Hyderabad
Hyderabad - 500046, India

DECLARATION

I, **Shailendra Sahu**, hereby declare that this thesis entitled **"Disease Network Construction and Analysis: A case study of Alzheimer's Disease"** submitted by me under the guidance and supervision of **Dr. T. Sobha Rani** is a bonafide research work and is free from any plagiarism. I also declare that it has not been submitted previously in part or in full to this University or any other University or Institution for the award of any degree or diploma. I hereby agree that my thesis can be submitted in Shodganga/INFLIBNET.

A report on plagiarism statistics from the University Librarian is enclosed.

Date :

Signature of the Student:

Name: Shailendra Sahu

Reg. No.: 16MCPC15

To my parents, ***Late Smt. Rukmani Devi Sahu and Shri Ram Babu Sahu****, and friends without whose support and encouragement, this would not have been possible.*

Acknowledgements

Completing this Ph.D. thesis would not have been possible without the guidance, support, and encouragement of several individuals and institutions. First and foremost, I would like to express my heartfelt gratitude to my research supervisor, Dr. T. Sobha Rani, for her invaluable guidance, motivation, and continuous support throughout my research journey. Her insightful feedback, constructive criticism, and unwavering encouragement have been instrumental in shaping my research work.

I am deeply grateful to my dissertation committee members, Dr. Durga Bhawani and Dr. Pankaj Singh, for their insightful feedback and valuable suggestions that have helped me improve my research work. I also thank the School of Computer & Information Sciences and its Dean, Dr. Atul Negi, for providing me with the necessary resources and a conducive research environment to pursue my Ph.D. studies.

I would also like to acknowledge the University of Hyderabad for providing me with the opportunity to pursue my Ph.D. studies and the financial support provided in the form of a fellowship.

Finally, I would like to extend my gratitude to my friends Anil, Ishan, ND Patel, Baby, Sebanti, and others for their unwavering support, encouragement, and motivation throughout my Ph.D. journey. Their support and friendship have made this journey a memorable one.

Thank you all for your invaluable support and encouragement!

Shailendra Sahu

Abstract

Each cell in a biological system requires thousands of proteins to perform specific tasks at specific times and locations to function correctly. Gene variants can occasionally prevent one or more proteins from functioning correctly by making a protein malfunction or not be produced at all by altering the gene's instructions which produces it. A variant can impair normal development or result in a disease when it changes a protein essential to the body. This information can be represented as a interaction network. Construction and analysis of Gene-Gene networks, constructed using gene expression data, are popular ways to understand the underlying mechanisms of complex diseases. The major challenge with gene expression data with a large number of variables(genes) and a comparatively very small number of samples is to extract disease-related information, as the gene expression data contains a vast amount of redundant data and noise.

This thesis focuses on constructing a statistically and biologically meaningful Alzheimer's disease gene networks from the gene expression data and then the problem-specific analysis of the constructed disease gene networks. As the first contribution to the thesis, we have introduced a novel framework to construct Alzheimer's disease gene networks. The framework uses t-test, correlation, Gene Ontology categories machine learning techniques to construct the disease gene network and to detect the potential biomarker genes. In the second contribution, we have used the proposed framework to construct the stage-wise Alzheimer's disease gene networks and carried out the community analysis. We have proposed a new stable community discovery algorithm (Neighbour-based community discovery algorithm) for community analysis. In the third contribution, we have analyzed the stage-wise Alzheimer's disease gene network to identify the genes that may be responsible for or play an important role in the disease progression. We

have introduced a new centrality measure to rank the genes according to their involvement in the network progression. Through all these methods, we could identify a large number of genes that are proven to be important in Alzheimer disease progression and onset. Further, this framework is generic enough for it to be used with any other disease.

Contents

List of Figures

List of Tables

List of Algorithms

Chapter 1

Introduction

Gene expression studies and gene-network analyses are vitally important for understanding complex diseases. This facilitates an understanding of the underlying process of these disorders. Gene activity is affected and regulated by other genes. One of the popular ways to represent this information is through gene regulatory networks, which are constructed using gene expression data where genes are represented as nodes and edges represent the dependencies and possibly functional relations among genes. However, constructing a disease gene network with equal statistical and biological significance remains a major challenge among researchers. The vast amount of data and noise in the gene expression data makes it a challenge to construct and analyse the networks and extract the disease-specific information.

1.1 Biological Networks

Genes are the basic physical and functional units of heredity which are parts of the DNA. Every gene comprises of a particular set of instructions for a particular function or protein. According to Human Genome Project, humans have approximately 20000 to 25000 genes. Gene-gene interaction networks refer to the relationships between genes and how they interact with each other to regulate various biological processes in the body. These networks consist of a group of genes that are interconnected and influence each other's expression and function. These interactions can be direct, where one gene directly regulates the expression or function of another gene, or they can be indirect, where one gene influences the expression or function of another gene through

intermediate pathways or regulatory factors.

Gene-gene interaction networks are important in understanding the intricate genetic basis of various diseases and disorders, as well as identifying potential therapeutic targets for these conditions. There are several different gene-gene interaction networks that are important in disease analysis:

- Genetic pathways: These are networks of genes that are involved in specific biological processes or pathways. For example, the signaling pathway that regulates cell growth and division is a genetic pathway. Figure 1.1 shows a gene-gene network based on their GO categories[See 2.2.2].

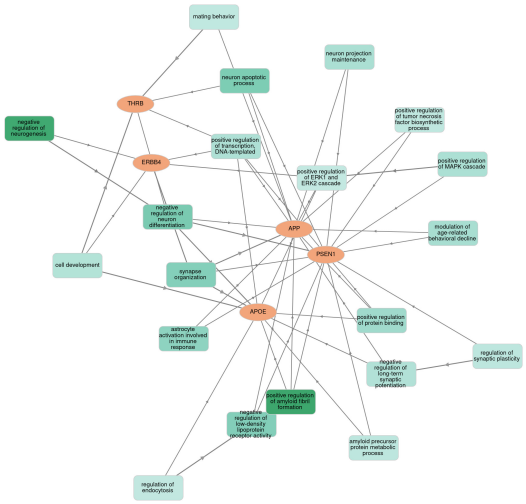


Figure 1.1: GO graph using GOnet tool.

- Protein-protein interaction networks (PPI): These are networks of proteins that interact with each other to perform specific functions in the cell. Fig. 1.2 shows an example of PPI network. These interactions can be disrupted in diseases, leading to abnormal function.

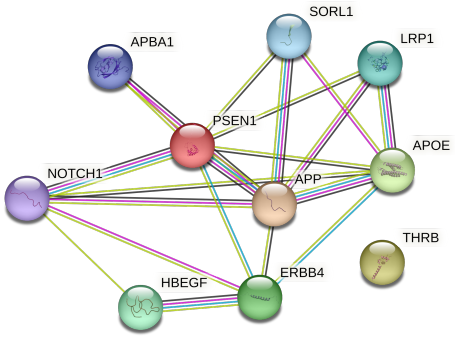


Figure 1.2: PPI network using STRING database.

- Genetic networks: These are networks of genes that interact with each other to regulate gene expression. Dysregulation of these networks can lead to abnormal gene expression and disease.
- Genetic regulation networks: These are networks of genes that regulate the expression of other genes. Dysregulation of these networks can lead to abnormal gene expression and disease manifestation.
- Genetic disease networks: These are networks of genes that are associated with a specific diseases. Analysis of these networks can help identify potential therapeutic targets and improve diagnosis and treatment of diseases.

Overall, the importance of gene-gene interaction networks in disease analysis lies in their ability to provide insights into the hidden genetic basis of the disease and identifying the potential therapeutic targets. Network-based methods in disease analysis have several advantages over other methods, including:

- Integrative analysis: Network-based methods allow for the integration of multiple data types and sources, providing a more comprehensive understanding of the underlying biological structure of a disease.
- Contextual information: Network-based methods provide additional context and relationships between genes, proteins, and biological pathways, helping to identify

potential biomarker genes.

- Prioritization: Network-based methods can prioritize candidate genes and pathways based on their centrality or importance in the network, making it easier to identify the most promising targets for further investigation.
- System-level understanding: Network-based methods provide a systems-level view of disease biology, allowing for the identification of common disease mechanisms and potential drug targets.
- Data visualization: Network-based methods can visually represent complex biological relationships, making it easier to communicate results and understand complex biological systems.

1.2 Alzheimer's Disease

Brain can be divided into different regions, each with its own distinct functions and characteristics. These regions include the frontal lobe, parietal lobe, occipital lobe, temporal lobe, cerebellum, brainstem, and limbic system. The frontal lobe is responsible for decision-making, problem-solving, and impulse control, while the parietal lobe plays a key role in processing sensory information such as touch and spatial awareness. The occipital lobe is primarily involved in processing visual information, while the temporal lobe is important for language, memory, and hearing. The cerebellum is responsible for coordination and movement, while the brainstem controls basic functions such as breathing and heart rate. The limbic system, which includes the hippocampus, amygdala, and hypothalamus, is involved in emotions, motivation, and memory formation. Figure 1.3 shows the different brain regions and functions associated with the respective regions. Figure 1.4 show the limbic system.

Any disease network can be viewed as pre and post disease. Pre, a normal network before the disease afflicts a person and post where the disease has manifested to a certain extent.

Alzheimer's disease is a prevalent form of dementia. It is an irreversible disease with a progressive loss of memory and worsening cognitive function. The leading cause of AD is said to be the abnormal deposits of protein forms amyloid plaques and tau tangles throughout the brain [1]. The hippocampus is a brain structure located deep

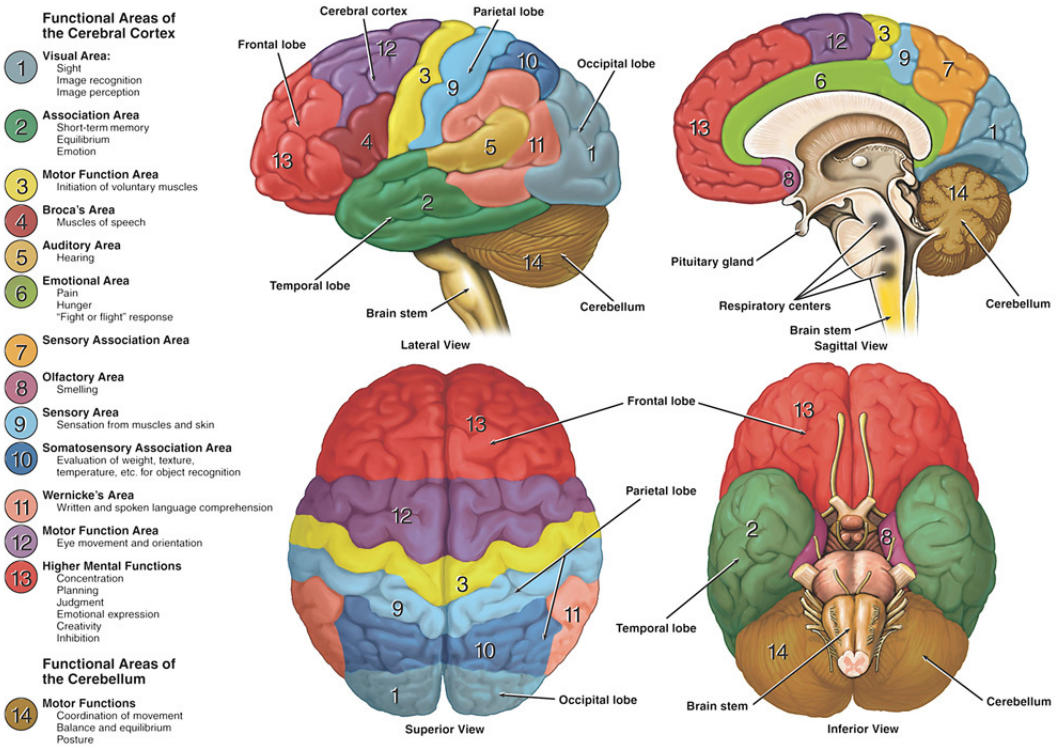


Figure 1.3: Regions of Brain. Source: Dana.org(Neuroanatomy: The Basics)

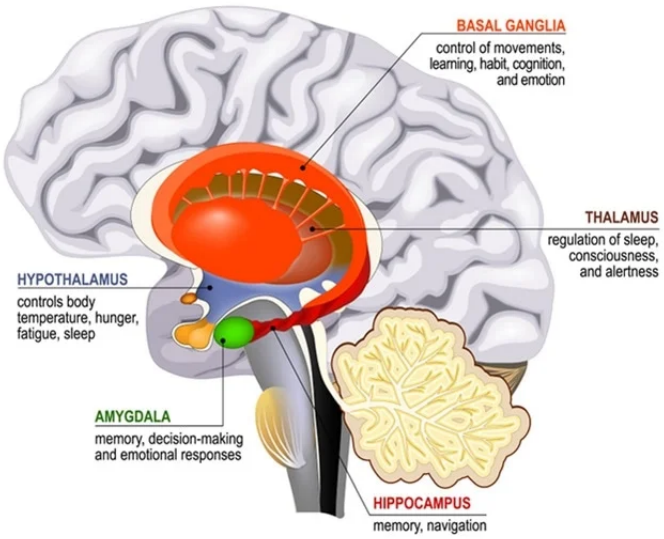


Figure 1.4: Limbic System. Source: Designua/Shutterstock

within the temporal lobe that plays a crucial role in all aspects of semantic memory. It is often reported to be the first area affected in Alzheimer's disease [2, 3]. Figure 1.5 shows the brain images during the Alzheimer's disease progression.

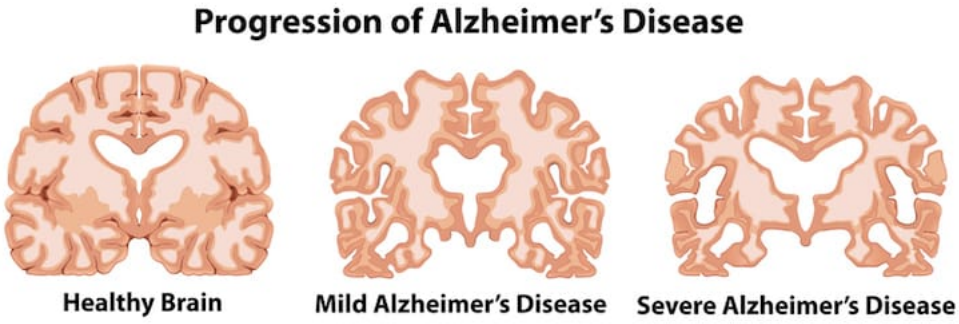


Figure 1.5: Progression of Alzheimer's Disease. Source: https://www.drugwatch.com/health/alzheimers-disease/

1.3 Motivation

Alzheimer's disease is a neurological disorder that affects an individual's memory, motor functions, behavior, and thought process. Its analysis poses several challenges:

- Early Diagnosis: It is difficult to diagnose Alzheimer's disease in its early stages, because the symptoms are similar to those of other age-related disorders such as mild cognitive impairment.
- Lack of a definitive diagnostic test: Currently, the diagnosis of AD is based on clinical evaluation and neuroimaging, but a definitive test is still not available.
- Heterogeneity of the disease: AD can manifest with different symptoms and progression patterns, making it hard to study and treat.
- Complexity of the disease mechanisms: Alzheimer's is a complex disease with multiple causes, making it difficult to pinpoint specific genetic or environmental factors that contribute to its development.

1.3.1 Current research

Current research on Alzheimer's disease is focused on:

- Early detection and diagnosis: Developing new biomarkers and diagnostic tools to detect AD in its early stages.
- Understanding the disease mechanism: Research is ongoing to uncover the molecular and cellular mechanisms involved in the development of AD.
- Development of new treatments: There is ongoing research to develop new drugs and therapies to treat AD, including the use of gene therapy, immuno-therapy, and stem cell therapy.

1.3.2 Key Research Methods

Some of the key research methods for analysing the Alzheimer's disease are as follows:

- Genome-wide association studies (GWAS): GWAS are used to identify genetic variants that are associated with Alzheimer's disease.
- Epigenetic changes: Researchers are exploring the role that epigenetic changes may play in the development of Alzheimer's disease.
- Biomarkers: Researchers are working to identify biomarkers that can be used to diagnose Alzheimer's disease and monitor its progression.
- Targeted therapy: Researchers are exploring the use of targeted therapy to treat Alzheimer's disease by targeting specific genes or biological pathways.

1.3.3 Challenges

Challenges associated with these methods:

- Sample size: Large sample sizes are required for GWAS to have enough power to detect genetic variants associated with Alzheimer's disease.
- Complexity of epigenetic changes: Understanding the complex interplay between genetic and environmental factors that contribute to epigenetic changes is challenging.
- Limited understanding of biomarkers: There is a limited understanding of the specific biomarkers that are indicative of Alzheimer's disease, making it difficult to develop effective diagnostic tests.

- Cost and complexity of targeted therapy: Targeted therapy is a complex and costly approach, requiring significant funding and resources.

1.4 Objectives

Objectives of this work are:

- Construction of more biologically, statistically meaningful Alzheimer's disease gene network.
- Analysis of Alzheimer's disease stage-wise gene networks:
  - Identifying the genes whose communities got changed/disturbed during the disease progression from one stage to the next stage.
  - Dynamic network analysis of Alzheimer's disease networks to identify the genes which may play an important role in the disease progression.

1.5 Contributions

1.5.1 Construction of Alzheimer's Disease Gene Network

Most often, t-test and correlation are used to identify significant genes at the initial level. As the genes are differentially expressed, their classification power is generally high. These genes might appear significant, but their degree of specificity towards the disease might be low, leading to misleading interpretations. Similarly, there may be many false correlations between the genes that can affect the identification of relevant genes. We introduced a new framework, tcGONet, to reduce the false correlations and find the potential bio-markers for the disease. The tcGONet framework concerned uses the t-test, correlation, Gene Ontology (GO) categories, and machine learning techniques to find bio-marker genes. The tcGONet framework detects Alzheimer-related genes in every dataset considered. Some of the identified genes which are directly involved in Alzheimer's are APP, GRIN2B, and APLP2. The proposed framework also identifies genes like ZNF621, RTF1, DCH1, and ERBB4, which may play an important role in Alzheimer's. Gene set enrichment analysis (GSEA) is also carried out to determine the major GO categories: down-regulated and up-regulated. The work in this contribution has been published in [4].

1.5.2 Community analysis

Detecting communities/subnetworks in disease conditions or drug treatments can provide valuable insight into disease etiology or therapeutic responses. However, choosing a suitable community discovery algorithm is important. Every algorithm has its own pros and cons.

As a contribution to the thesis, we have proposed a new community discovery algorithm, NBCD, which shows more stability in detecting better community structures according to popular measures than the other state-of-the-art and newly published algorithms. The NBCD algorithm is used to analyze the changes in the neighbourhood(community) of genes as the disease progresses. We have analyzed the top 20 genes according to the changes in their neighbourhood and interestingly, we identified genes related to AD.

1.5.3 Disease Progression

Temporal network analysis has become a powerful tool for unveiling the network evolution over time. In recent times, different centrality measures have been proposed to measure the importance of nodes in different scenarios. However, there is no centrality measure yet introduced to measure the importance of nodes in the network progression. This work introduces a new centrality measure, transition centrality, to measure the node's importance in network evolution between two given time stamps.

Transition centrality can play an important role in the analysis of disease progression. In the past, many studies have been done to identify the potential genes related to diseases. However, the stage-wise analysis of diseases is less explored. Identifying the role of the gene in disease progression or the gene's role in a particular stage of the disease is not studied extensively. Believing that different genes are responsible for different stages of disease progression, we evaluate the transition centrality measures on three different temporal disease datasets; Alzheimer's disease, Parkinson's disease and the Human breast cancer cell cycle. Using the transition centrality, we have identified the stage-specific genes which may play a crucial role in the disease progression. The identified genes' specificity to a particular disease stage validates our findings.

1.5.4 Publications

The list of papers published during the Ph.D.:

1. **Shailendra Sahu**, Pankaj Singh Dholaniya, & Rani, T.S. **Identifying the candidate genes using co-expression, GO, and machine learning techniques for Alzheimer's disease**. *Network Modeling Analysis in Health Informatics and Bioinformatics* 11, 10 (2022). DOI: 10.1007/s13721-021-00349-9.

2. **Shailendra Sahu**, T. Sobha Rani, **A neighbour-similarity based community discovery algorithm**, *Expert Systems with Applications*, Volume 206, 2022, 117822. DOI: 10.1016/j.eswa.2022.117822.

1.6 Outline

Chapter 2 gives an overview of Alzheimer's Disease. It explains the preliminary background required to understand the terms and techniques which are required to understand the further chapters. The third chapter will present our proposed framework to construct and analyse the disease networks. Chapter 4 will show our novel community discovery algorithm(NBCD) and our findings related to Alzheimer's disease using NBCD. The fifth chapter presents the stage-wise analysis of the Alzheimer's disease network and our proposed centrality measure to rank the genes according to their involvement in the network progression. Finally, Chapter 6 presents the conclusion of the thesis.

Chapter 2

Literature Survey

Neurological diseases are conditions that affect the central nervous system (CNS), which includes the brain and the spinal cord. These diseases can result in conditions from mild to severe and can affect a person's ability to move, speak, and think. There are several such diseases like Alzheimer's, Parkinson's, Multiple sclerosis and so on.

Most common neurological disease is Alzheimer's disease, which is a type of dementia that affects memory, thinking, and behavior. It is caused by the degeneration of brain cells and is typically diagnosed in older people.

Another neurological disease is multiple sclerosis (MS), which is an autoimmune disorder that affects the myelin sheath, a protective layer surrounding nerve fibers. This damage can disrupt communication between the brain and the rest of the body, leading to symptoms such as muscle weakness, numbness, and balance problems.

Parkinson's disease is a progressive neurological disorder that affects the brain's ability to control movement. It is caused by the loss of nerve cells that produce dopamine, a neurotransmitter that helps coordinate muscle movement. Symptoms include tremors, stiffness, and difficulty with balance and walking.

Stroke is a neurological disorder caused by a disruption of blood flow to the brain. This can be caused by a blockage or bleeding in the brain and can lead to serious damage or death. Symptoms include paralysis, loss of speech, and difficulty with memory and cognition.

Epilepsy is a neurological disorder characterized by seizures, which are sudden, uncontrolled electrical discharges in the brain. Seizures can range in severity and may cause changes in behavior, consciousness, or body movements.

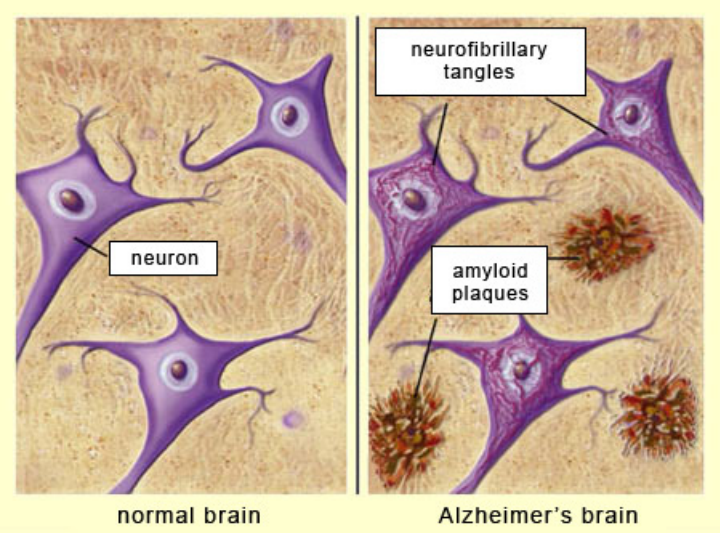


Figure 2.1: Amyloid Plaques and Neurofibrillary Tangles. Source: http://www.interet-general.info/article.php3?id_article=13241

There are many other neurological diseases, such as brain tumors, spinal cord injuries, and traumatic brain injuries, which can have significant impact on a person's physical and cognitive abilities. It is important to seek medical attention and follow a treatment plan to manage and potentially improve symptoms of these diseases.

2.1 Alzheimer's Disease

According to the National Institute of Aging, U.S., Alzheimer's disease (AD) is a brain disorder that gradually impairs thinking and memory abilities as well as the capacity to complete even the most basic tasks. AD is the most typical cause of dementia in older adults. The disease is named after Dr. Alois Alzheimer. Amyloid plaques and tau tangles are considered as the primary cause of Alzheimer's disease. During Alzheimer's disease, amyloid plaques, which are abnormal clumps of protein, and tau tangles, which are tangled bundles of fibers, develop throughout the brain. The amyloid plaques build between the neurons, and tau tangles build inside the neurons, which interrupts the communication between neurons. Figure 2.1, shows the comparison between a normal and Alzheimer's brain in context of amyloid plaques and tau tangles. APOE is said to be the most common gene associated with AD [1]. Apart from APOE, APP, PSEN1, and PSEN2 are also observed as the cause of AD [5].

Initial manifestations of these damages can be observed in the entorhinal cortex and hippocampus, two brain regions important for memory. Later, it has an impact on the parts of the cerebral cortex that controls language, thought, and social interactions. Eventually, brain's many other regions suffer a harm. As AD progresses, the brain starts shrinking. A person with Alzheimer's disease gradually loses the ability to live and work independently over time. However, the method by which the disease develops remains unclear; all forms of Alzheimer's appear to share overproduction and/or decreased clearance of a type of protein called amyloid beta peptides.

2.2 Background

The study of genes, Gene Ontology, and pathways is important for understanding the complex mechanisms behind genetic inheritance and the roles that genes play in various biological processes. This knowledge can help researchers identify potential genetic markers for diseases, develop new treatments, and improve our understanding of the genetic basis of traits and disorders.

2.2.1 Gene

Gene is the basic unit of heredity passed from the parent to the child. Genes are made up of sequences of DNA and are arranged, one after the other, at specific locations on chromosomes in the nucleus of cells. They contain information for making specific proteins that lead to the expression of a particular physical characteristic or trait, such as hair color or eye color, or to a particular function in a cell [6].

2.2.2 Gene Ontology and Pathway

The Gene Ontology (GO) is a standardized vocabulary used to describe the functions of genes and gene products in a consistent and systematic way. It consists of three main categories: molecular function, cellular component, and biological process [7].

- Molecular function: This category describes the specific chemical or physical activity of a gene product, such as enzyme activity or receptor binding.
- Cellular component: This category describes the location or structure within the cell where a gene product is found, such as the cytoplasm or mitochondria.

- Biological process: This category describes the broader physiological or developmental processes in which a gene product is involved, such as cell growth or immune response.

A pathway is a series of biochemical reactions that are connected and interdependent, leading to the production of a specific product or the achievement of a specific function. Pathways can be either metabolic (involved in the production and breakdown of molecules) or signaling (involved in the transmission of information within cells or between cells) [8].

2.2.3 Gene Expression Data

Gene expression data refers to the measurement of the levels of gene activity in a cell or tissue at a specific time point. These measurements can be used to understand the functional role of specific genes, how they are regulated, and how they respond to different environmental conditions or treatments. Gene expression data is in table form where rows represent the genes and columns represent the various samples such as experimental conditions or tissue, and every cell in the table has a number which characterizes the expression level of the particular gene in that particular sample. There are several methods for measuring the gene expression data, including microarray-based techniques, RNA-sequencing (RNA-seq), and quantitative polymerase chain reaction (qPCR).

Pros:

- Gene expression data can provide valuable insights into the mechanisms underlying biological processes and diseases.
- Gene expression data can be used to classify tissues or cell types based on their gene activity patterns.
- Gene expression data can help identify the functions of newly discovered genes.

Cons:

- Gene expression data may not accurately reflect the true levels of gene activity in a cell or tissue due to technical limitations of the methods used to measure gene expression.

- Gene expression data may be influenced by environmental factors such as diet and stress, which can confound the results.
- Gene expression data may be influenced by the genetic background of the organism being studied, which can limit the generalizability of the findings.

For any biological process, the microarray gene expression data offers a simultaneous gene expression profile of thousands of genes. Among the thousands of genes, a few important genes typically play a dominant role in the development of a disease. An important area of bioinformatics research involves a computational method to identify disease-related genes, as any classification scheme based on gene expression data faces a significant bottleneck because, despite the small sample size, the feature space is enormous, containing tens of thousands of genes[9, 10].

2.2.4 Differentially expressed genes

We can better understand the pathology of diseases and, eventually, treat them, by examining the difference between the diseased and healthy states o the genes. Differentially expressed genes (DEGs), which involve the identification of genes that are differentially expressed in disease, is a major area of investigation. A gene is said to be differentially expressed if there is a statistically significant difference or change in expression levels between two experimental conditions. DEGs can be helpful in identifying potential biomarkers, therapeutic targets, and gene signatures for diagnostics in pharmaceutical and clinical research. Even though specific gene expression changes may not always result in biological activity, such information can still be combined with other biological data in a high-throughput manner to produce integrated analyses, such as mapping the disease's target landscape[11, 12].

2.2.5 Gene Regulations

Gene up-regulation and down-regulation refer to the changes in the expression levels of a particular gene.

- Up-regulation refers to the process by which a gene's expression is increased, resulting in an increased amount of the protein that the gene encodes. This can occur naturally, or it can be induced by external factors such as stress, hormones, or exposure to certain drugs.

- Down-regulation, on the other hand, refers to the process by which a gene's expression is decreased, resulting in a reduced amount of the protein that the gene encodes. This can occur naturally as part of a normal developmental process, or it can be induced by external factors such as environmental toxins, drugs, or disease.

The regulation of gene expression is a critical process that allows cells to adapt to changing conditions and maintain proper function. Up-regulation and down-regulation of genes are important mechanisms for controlling cellular processes and maintaining homeostasis in the body.

In this work, statistically and biologically meaningful networks are constructed to identify the genes that are responsible for disease onset and progression. Community discovery discovery algorithm are used to detect the structure within the networks that may not be directly available. Progression of the disease in dynamic networks is studied through centrality measures. Sections 2.3, 2.4 and 2.5 provide the details about the background for each of these proposals.

2.3 Network Construction

Networks, or graphs, can be a useful tool for analyzing diseases for a number of reasons. One reason is that they can be used to represent the relationships between different components of a system, such as the relationships between different genes or proteins in a biological system, or the relationships between different individuals in a population. This can help researchers understand how different components of the system are connected and how they may influence each other.

Another reason why networks are useful for analyzing diseases is that they can be used to identify patterns and trends that may not be apparent when looking at data in other formats. For example, network analysis can be used to identify clusters of genes or proteins that are highly connected and may be playing a key role in the disease. It can also be used to identify important nodes or hubs in the network that may be driving the disease process.

Use of networks and graph theory in disease analysis can help researchers better understand the complex systems underlying diseases and identify potential targets for intervention or treatment. There are several methods that can be used to construct

networks for studying diseases:

- Literature-based networks: These networks are constructed by extracting information about disease-associated genes, proteins, or other molecular entities from the scientific literature. The nodes in the network represent the molecular entities, and edges are drawn between nodes that are mentioned in the same context in the literature. For example in [13], Mallory et al. used 100000 full-text PLOS articles to extract both protein–protein and transcription factor interactions. In [14], Garand et al. identified the list of potential signature genes for the multifaceted disease using Acumenta Literature Lab^{TM} (LitLab). LitLab is an online literature mining tool to extract the information about genes, pathways and other biological functions related to user's query. But there are several potential drawbacks in using literature-based gene-gene networks:
  - Limited coverage: Literature-based gene-gene networks are only as comprehensive as the published literature. If there is limited research on a particular gene or interaction, it may not be represented in the network.
  - Bias: The published literature is subject to various biases, such as publication bias, which means that certain types of studies or results are more likely to be published. This can lead to a biased view of the gene-gene interactions.
  - Incomplete information: Literature-based gene-gene networks are limited to the information that is available in the published literature. This means that they may not include all of the relevant information about a particular gene or interaction.
  - Inaccuracies: There may be errors or inconsistencies in the published literature, which can lead to inaccuracies in the gene-gene network.
  - Complexity: Gene-gene networks can be complex and difficult to interpret, especially for people without a strong background in biology or genetics.
- Data-driven networks: These networks are constructed from large datasets, such as gene expression data or protein-protein interaction data. The nodes in the network represent genes or proteins, and edges are drawn between nodes that show correlated expression or physical interaction. Data-based gene networks

are the widely used networks to identify the potential genes for disease, analyzing disease progression and underlying process of biological functions and the disease. Enormous research has been carried out based on the gene-expression and protein interaction data. For example in [15], Lemoine et al. developed a R package GWENA for constructing and analysing gene-expression networks. In [16], Lau et al. investigates the changes in gene expression patterns during Drosophila melanogaster embryogenesis. This study provides insight into the complex changes in gene expression that occur during Drosophila melanogaster embryogenesis and highlights the importance of studying gene expression over time in understanding developmental processes. Despite the popularity of gene-expression network, there are many challenges associated with the gene-expression data. In [17], Burns et al. performed experiments on 475 datasets and concludes that up to 97% of edges in the gene-expression network can be false or incorrect. Below are few challenges which make it difficult to construct and analyze the data-based gene networks:

  - Limited sample size: Data-driven approaches often rely on large amounts of data to identify patterns and relationships. However, this can be a problem when dealing with gene expression or protein interaction data, as the number of samples available for analysis may be limited. This can lead to unreliable or biased results, as the sample may not accurately represent the larger population.
  - Complexity of data: Gene expression and protein interaction data can be extremely complex and multi-dimensional, making it difficult to accurately analyze and interpret the results. This can lead to errors or misunderstandings of the data, which can have significant consequences for downstream applications such as drug development or disease diagnosis.
  - Dependence on data quality: The quality of the data collected is critical to the accuracy and reliability of data-driven approaches. If the data is contaminated or poorly collected, it can lead to misleading or incorrect results.

- Hybrid networks: These networks combine information from the literature with data-driven approaches. For example, a hybrid network might include edges between genes that are supported by both literature evidence and data-driven

evidence. Many researchers have integrated different biological networks into one gene-gene interaction network. For example, in [18, 19, 20], authors integrated the gene co-expression and protein-protein interaction networks in order to construct a more meaningful gene network. However, there are some challenges associated with the integration of different networks such as:

  - Complexity: Hybrid gene networks can be complex, with many different types of genetic elements interacting in intricate ways. This complexity can make it difficult to understand and predict the behavior of the network.
  - Robustness: Hybrid gene networks may be less robust to perturbations or changes in the environment compared to simpler regulatory systems. For example, a change in the expression level of a single transcription factor could have downstream effects on the expression of many genes in the network.
  - Dynamics: The dynamic behavior of hybrid gene networks can be difficult to predict, as the interactions between different genetic elements can produce nonlinear or non-intuitive outcomes.

- Clinical networks: These networks are constructed from clinical data, such as patient records or electronic health records. The nodes in the network represent patients, and edges are drawn between patients who share certain characteristics, such as a diagnosis or a treatment. Some major drawbacks of clinical networks are:

  - Cost: Participating in a clinical network often requires a financial investment, which may be a burden for smaller practices or organizations.
  - Time commitment: Participating in a clinical network requires a time commitment, as members must attend meetings, participate in conference calls, and complete required training.
  - Loss of autonomy: Joining a clinical network may require member organizations to cede some control and decision-making authority to the network.
  - Complexity: Clinical networks can be complex organizations, with multiple levels of governance and decision-making. This can make it difficult for members to navigate and understand the inner workings of the network.

  - Limited reach: Clinical networks may only cover a limited geographic area, which can limit their effectiveness for organizations or patients located outside of the network's coverage area.

2.4 Community Discovery

Community discovery, refers to the process of identifying groups or clusters of nodes that are densely interconnected within the group, but less connected to nodes outside of the group in a network. This process is often used to identify patterns or structures within the network that may not be immediately apparent.

In the context of disease gene networks, community discovery can be helpful in identifying groups of genes that are closely related to a particular disease. For example, if a group of genes is found to be highly interconnected within the network, and these genes are also known to be associated with a particular disease, this could suggest that these genes play a central role in the development or progression of the disease.

Community discovery can also be useful in identifying potential therapeutic targets for a disease. For example, if a particular group of genes is found to be important in the disease process, targeting these genes with drugs or other therapies may be a potential way to treat or prevent the disease from originating or progressing.

Overall, community discovery can be a valuable tool for understanding the underlying mechanisms of a disease and identifying potential therapeutic approaches. Some of the state-of-the-art community discovery algorithms are as follows:

- Walktrap: It is a community discovery algorithm based on the random walks[21]. One potential drawback of this algorithm is that it may not always produce high-quality communities as its performance strongly depends on the degree distribution of the network [22].

- Infomap: This algorithm uses the principle of information theory to partition a network into communities[23]. Infomap accurately uncovers the communities that are strongly connected internally, but fails to do so for loosely connected communities [24].

- Modularity maximization: Algorithms based on the modularity maximization like Greedy modularity [25] and Louvain [26] aims to partition a network into commu-

nities such that the modularity of identified community structure is maximum. One major drawback of this algorithm is that it is sensitive to the resolution limit, which means that it may not be able to identify smaller communities within larger ones [27].

- Label Propagation Algorithm (LPA): This algorithm is based on the idea of propagating labels through a network in order to identify communities[28]. It is a fast and efficient community detection algorithm. A major drawback of LPA is the randomness in grouping nodes that leads to instability and the formation of large communities.

In summary, every community discovery algorithms has its own strengths and weaknesses and is well-suited for certain types of networks and use cases.

2.5 Centrality Measures

Centrality measure is a statistical method used to determine the importance or influence of a particular node or vertex in a network. It helps to identify the most influential nodes within a network, which may be used to understand the structure and dynamics of the network.

There are several types of centrality measures, including:

- Degree centrality [29]: This measure calculates the number of connections a node has in the network. Nodes with a high degree centrality are considered highly connected and influential. The main drawback of degree-based centrality is that it only provides local information about a network vertex.

- Betweenness centrality [30]: This measure calculates the number of times a node acts as a bridge or connector between other nodes in the network. Nodes with high betweenness centrality are considered important for information flow in the network. One of the major limitation of this measure can be that it is computationally expensive to calculate and it may not be appropriate when there are a number of parallel edges between nodes.

- Closeness centrality [31]: This measure calculates the average distance between a node and all other nodes in the network. Nodes with high closeness centrality are

considered well-connected and able to reach other nodes quickly. This measure cannot be used for disconnected networks.

- Eigenvector centrality [32]: This measure calculates the influence of a node based on the influence of the nodes it is connected to. Nodes with high eigenvector centrality are considered influential due to the influence of their connections.

No centrality measure is best or worst. They all are application specific. The nodes which are important/central according to one centrality measure are often not that important according to another centrality measure.

Chapter 3

Identifying the Candidate Genes using tcGONet Framework for Alzheimer's Disease

Gene-Gene interaction networks can be used to understand the underlying process that are responsible for the these interactions. Construction of such networks itself is not a trivial task, since specious things could lead to different interpretations. In this chapter, a framework tcGONet, is proposed to construct a statically and biologically meaningful disease networks.

3.1 Literature Survey

Alzheimer's disease is a neurological disorder that affects an individual's memory, motor functions, behaviour, and thought process. It has been observed that the hippocampus is the first region that gets affected by Alzheimer's. Hence a study of the hippocampus region may identify genes responsible for the occurrence of the disease. This can be the early stage of the disease.

Various studies have been carried out to identify the genes which are differentially expressed in the AD affected brains [5, 33]. T-test, gene correlation networks are the most common statistical techniques used to identify the significant genes. The t-test is used to test the significant difference in gene expression levels [34]. For example, in [35], Zhu and Yang et al. used the rejection region of the t-test to identify the candidate

genes for AD. In [33], Sumanta Ray et al. analyzed the preservation patterns of gene co-expression networks during Alzheimer's disease progression. However, the t-test only gives the significant difference in the mean expression values of genes between control and disease sets, which is not enough to determine the significant influence of genes on the disease. There could be many other reasons apart from the disease, which can result in a change in the expression value of a particular gene.

A gene correlation network, also known as a gene co-expression network, is a computational method used to analyze the relationships between genes based on their expression patterns. The network is constructed by measuring the levels of gene expression across a large number of samples or tissues and then calculating the correlation between the expression patterns of pairs of genes. Genes that have highly correlated expression patterns are considered to be functionally related and are often involved in the same biological processes or pathways.

The gene correlation network can be visualized as a graph, where each gene is represented by a node and the edges between the nodes represent the correlation between the expression patterns of the corresponding genes. By analyzing the structure of the gene correlation network, researchers can identify groups or modules of genes that are co-expressed and may have related functions. This approach has been widely used to study the genetic basis of various diseases and traits, as well as to identify potential drug targets and biomarkers.

Rui-ting et al. [36] constructed a co-expression network using WGNCA and analyzed their clinical features. As a result, they identified four genes(ENO2, ELAVL4, SNAP91, and NEFM) said to be associated with AD. In [37], Xia J et al. constructed the co-expression network using the method proposed by Ruan and Zhang [38]. Then they ranked the genes based on a new topological overlap formula, a modified version of the formula described in [39, 40]. The main concern with constructing a co-expression network using this method is that it depends on the user-defined value α. Different values of α result in a different number of edges. This means every gene in the co-expression network is connected to its top α co-expressed genes. It may impact the removal of positive edges. Like the t-test, the correlation between two genes is not enough to tell that two correlated genes interact with each other. There may be many false correlations.

As the gene expression datasets are vast, various machine learning techniques are

used along with the other statistical methods. Takahiro Koiwa et al. and K. Nishiwaki et al. [41, 42] used the random forest to identify the AD-related genes. In [43], AL-Dlaeen et al. used a decision tree classifier to predict the AD. There are many other algorithms, such as the K-means clustering algorithm, Principal component analysis(PCA), ant colony algorithm (ACO), independent component analysis algorithm (ICA), the angle cosine distance algorithm and Chebyshev inequality algorithm (ACD), which produce less efficient and unstable results [35]. In [44], Sharma et al. combined two feature selection techniques, LASSO and Random forest, for gene selection and achieved a high classification accuracy. In [45], Ramya et al. used the t-test, Signal to noise ratio and f-test for the initial selection of genes and then selected genes were used in a modified particle swarm optimization algorithm to obtain further refined genes.

Cheng et al. [46] observed that the machine learning model's average classification accuracy is higher than that of conventional methods. Apart from this, the authors also observed that machine learning approaches could also recognize oxidative phosphorylation genes in the Alzheimer's pathway. In [47], Saputra et al. compared different decision trees with particle swarm optimization as feature selection methods and observed that the random forest gives the best accuracy. Kuang et al.([48]) compared the performance of three machine learning algorithms, artificial neural network(ANN), decision tree and logistic regression models, to predict the AD. They found that ANN worked better than the other two models and observed that the age, daily routine, urine neuronal thread protein associated with AD, smoking, alcohol intake and sex are the crucial factors.

Almost every feature selection technique is applied on differentially expressed genes, i.e. genes obtained after the t-test. As the genes are differentially expressed, their classification power is generally high. These genes might appear significant, but their degree of specificity towards the disease might be low, leading to misleading interpretations. Some genes are expressed in basic cellular pathways and possess a higher probability of being differentially expressed across several biological conditions [49]. Nevertheless, as AD's causes probably include genetic, environmental, and lifestyle factors, different genes are identified as important in different AD datasets. Due to these various factors involved in AD, statistical methods and machine learning techniques alone are inadequate.

3.2 Dataset

The gene expression datasets GSE48350, GSE5281 and GSE28146, are downloaded from Gene Expression Omnibus (GEO), NCBI. The datasets GSE48350[1](Dataset 1) and GSE5281[2](Dataset 2) contain gene expression data of control and Alzheimer's disease patients. The dataset GSE28146[3](Dataset 3) contains microarray data of the hippocampal gray matter. The GSE48350 and GSE5281 datasets contain samples from different brain regions. We took only Hippocampus data for analysis as it is said to be affected first in Alzheimer's disease [3]. Table 3.1 describes the data.

Table 3.1: Dataset description

Datasets	Control	AD
GSE48350(Dataset 1)	25	19
GSE5281(Dataset 2)	13	10
GSE28146 (Dataset 3)	8	22

3.3 tcGONet

In this chapter, a new framework tcGONet, in addition to t-test and correlation network, GO-similarity matrix, and feature selection for filtering genes of less interest is proposed. Figure 3.1 shows the tcGONet, in particular for Alzheimer's disease.

Initially, differentially expressed genes are identified using the t-test. Then the identified genes are used to create two separate correlation networks for a disease and control sets using Pearson's correlation. There may be many false correlations, so a GO similarity matrix is introduced to reduce the false correlations. GO matrix consists of the number of similar GO terms between every pair of genes. Then the GO similarity matrix is used to eliminate edges in the correlation networks that do not fall under the pre-defined criteria. The resultant correlation networks are then used for further analysis. Genes present in the control correlation network but not in the disease correlation network and vice-versa are selected as the genes of interest. A separate Gene

[1]https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE48350
[2]https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE5281
[3]https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE28146

Set Enrichment Analysis(GSEA) has been carried out for selected genes to identify the affected GO categories. The feature selection algorithm is now applied to the selected genes to determine the most important genes from the important ones. This framework is generic and can be used for the construction and identification of important genes responsible for disease onset and progression. As a case study, Alzheimer's disease is chosen to verify the usability of this framework in identifying the important genes. All the components of the tcGONet are explained in detail in the following sections.

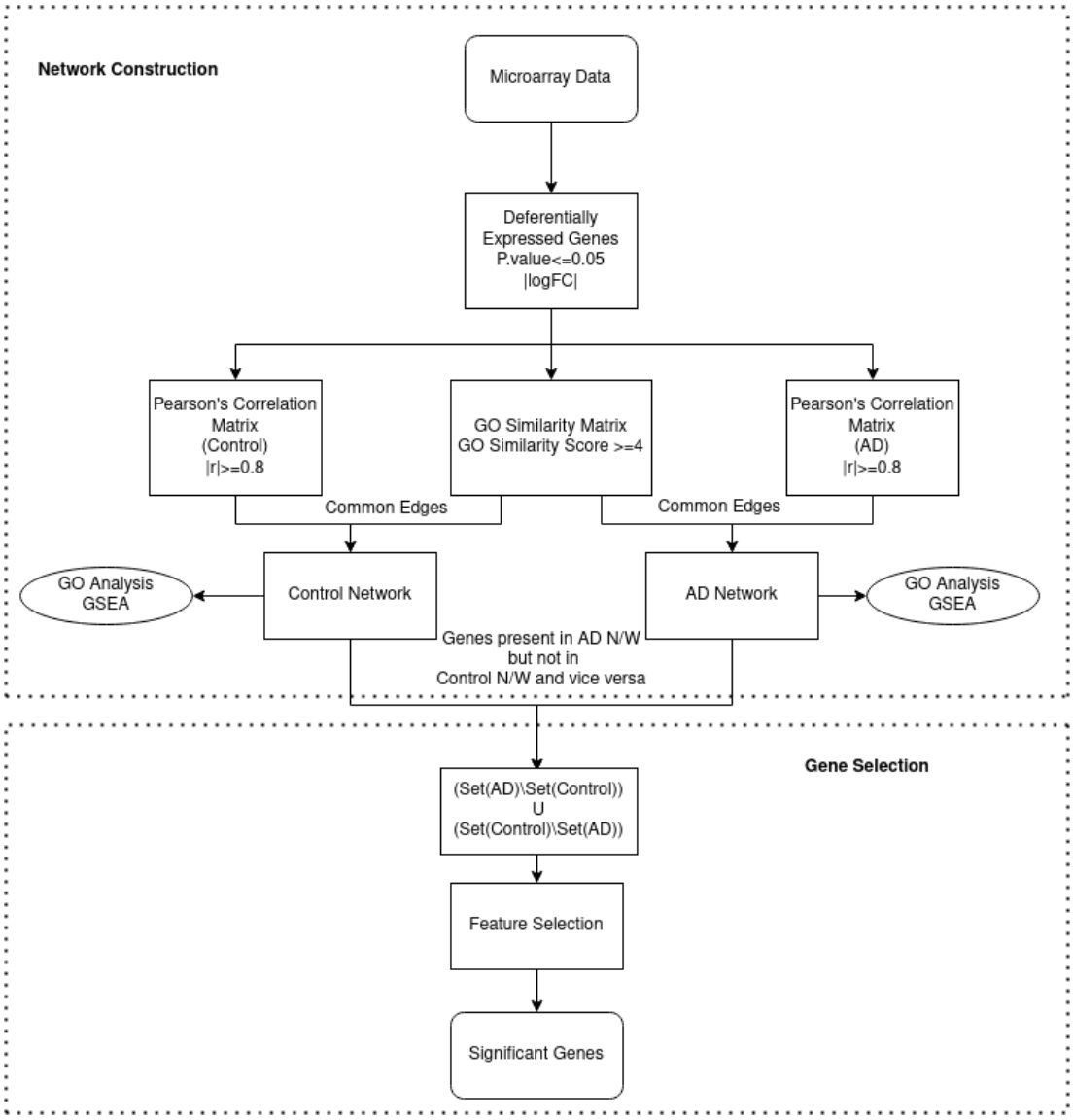


Figure 3.1: tpGONet framework for network construction and gene selection.

3.3.1 T-test

T-test is a statistical test that is used to determine whether there is a significant difference between the means of two groups. It is commonly used to compare the means of two independent samples, or to compare the means of two related samples (e.g. a pretest and a posttest).

The mathematical formula for a t-test is:

$$t = \frac{\bar{x_1} - \bar{x_2}}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

where $\bar{x_1}$ is the mean of group 1, $\bar{x_2}$ is the mean of group 2, s1 is the standard deviation of group 1, s2 is the standard deviation of group 2, and n1 and n2 are the sample sizes of group 1 and group 2, respectively.

A t-test was performed on all the datasets, i.e., GSE48350, GSE5281 and GSE28146, to find the significant difference in the expression values of genes in control and AD patients using GEO2R analysis tool[NCBI]. $p.value \leq 0.05$ and fold count, $|logFC| \geq 0.8$ are used as the threshold values to filter out the edges. These are standard values used in the literature [50]. As many genes have different probe ids, we took the average expression and fold count values. 696, 7222 and 1893 differentially expressed genes(DEGs) are obtained from dataset 1, dataset 2 and dataset 3, respectively.

3.3.2 Gene Co-Expression Network

Correlation is a measure of the relationship between two variables. It tells us how closely two variables are related, and the strength and direction of that relationship. There are several types of correlation, but Pearson's correlation is one of the most commonly used.

Pearson's correlation coefficient (also known as Pearson's r) is a measure of the strength and direction of a linear relationship between two variables. It is calculated using the following formula:

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}}$$

where x_i and y_i are the values of x and y variables in ith sample, $\bar{x}$ and $\bar{y}$ are the means of x and y variables.

The value of Pearson's r ranges from -1 to 1. A value of -1 indicates a strong negative relationship, a value of 1 indicates a strong positive relationship, and a value of 0 indicates no relationship.

Pearson's correlation is used to calculate the correlation between each pair of genes after performing the t-test. ± 0.8 is taken as the threshold value as it is interpreted as strong/high correlation [51, 52]. Akoglu and Mukaka have pointed out that a correlation value of 0.7 to 0.9 indicates a high positive correlation and 0.9 as a very high positive correlation. Hence a value of 0.8 is chosen as the threshold. All the correlation values which are greater than or equal to $|0.8|$ are considered as 1, and the rest of the values are considered as 0. The resultant adjacency matrix is used to create the gene co-expression matrix. Two separate networks for control and AD are constructed using the binarized Pearson correlation values as edges.

3.3.3 GO Similarity Matrix

Gene ontology (GO) [53] has become an accepted norm to evaluate the practical connections among gene products. GO is a scientific classification of biological terms identified using the properties of genes or their products. There are three GO categories: biological process, cellular component and molecular function. Two proteins engaged in the same biological process are bound to interact than proteins engaged with various biological processes [54]. Besides, two proteins need to come into close contact (essentially momentarily) to communicate; subsequently, co-localization can likewise be utilized to anticipate protein-protein interactions. Hence, the tcGONet uses GO categories for measuring the strength of the connection between genes in the correlation network.

GO similarity matrix consists of the GO similarity score between a pair of genes. Go similarity score is calculated as the number of common GO terms between two genes. For example, if Gene1 has 5 GO terms GO1, GO2, GO3, GO4 and GO5, and Gene2 has 4 GO terms GO1, GO3, GO5, and GO6. There are three common GO terms between the genes Gene1 and Gene2, which are GO1, GO2, and GO5. Hence the GO similarity score($GO_{(Gene1,Gene2)}$) between Gene1 and Gene2 is 3. GO categories of the differentially expressed genes (DEGs) identified by the t-test are used to construct the GO similarity matrix. The GO categories of all the DEGs are downloaded from DAVID (The Database for Annotation, Visualization, and Integrated Discovery) [55]. In the first dataset (GSE48350), out of 696 DEGs, 646 DEGs have known GO terms,

and in the second dataset (GSE5281), out of 7222 DEGs, 6377 DEGs have known GO terms. In dataset 3 (GSE28146), out of 1893 DEGs, 1210 DEGs have known GO terms. All the three GO categories, i.e., Biological Process (BP), Molecular Function (MF) and Cellular Component (CC) are considered for the construction of the GO similarity matrix. Gene similarity matrix consists of the GO similarity score between all pairs of genes, shown in Fig. 3.2.

$$\begin{bmatrix} & Gene1 & Gene2 & ..... & GeneN \\ Gene1 & 0 & GO_{(1,2)} & ..... & GO_{(1,N)} \\ Gene2 & GO_{(2,1)} & 0 & ..... & GO_{(2,N)} \\ . & . & . & ..... & . \\ . & . & . & ..... & . \\ . & . & . & ..... & . \\ GeneN & GO_{(N,1)} & GO_{(N,2)} & ..... & 0 \end{bmatrix}$$

Figure 3.2: GO Similarity Matrix

This GO similarity matrix is used to create the GO network. In order to determine the cut-off score for the GO similarity score, 4000 genes(except the genes considered in the experiment) having nearly 11000 edges that are experimentally proven are taken[DAVID]. The GO similarities between the genes having experimentally proven interactions are analyzed. Average number of similar GO terms between two genes(having experimentally proven edges (interactions)) is 3.14. Hence the ceiling value 4 is taken as the threshold value. All the edges whose weight (GO similarity score) is less than four are deleted. An edge between two genes is to be considered if they have at least four common GO terms.

3.3.4 Common Genes and Edges Between GO and Correlation Networks

A combined network is constructed to take care of the false correlations by mapping gene correlation networks (Control and AD) to the GO network. As genes sharing more GO terms will tend to have a high biological association, combining the correlation and GO network helps to eliminate the edges with less biological significance [54, 56]. A combined AD network is constructed using the common edges between the AD correlation network and the GO network. A similar combined network is constructed for the control network using the control correlation network and GO network. Table 3.2

shows the count of edges in the correlation network and GO network. Table 3.3 shows the network description of combine networks constructed using GO and correlation networks.

Table 3.2: Edge description of correlation and GO networks.

Dataset	Edges		
	Correlation N/W		GO N/W
	Control	AD	
GSE48350	16441	22814	6740
GSE5281	403894	364752	767620
GSE28146	52128	508	43803

Table 3.3: Network description of combined networks.

Dataset	Combined Control Network		Combined AD Network	
	Nodes	Edges	Nodes	Edges
GSE48350	240	673	219	774
GSE5281	2487	20486	3138	15499
GSE28146	589	989	12	7

3.3.5 Gene Set Enrichment Analysis(GSEA)

For the biological validation of constructed combined networks that are constructed, we performed the gene set enrichment analysis. The gene set enrichment analysis of AD and control networks is performed using GSEA 4.0 application, which can be downloaded from `http://software.broadinstitute.org/gsea` [57]. The all_GENE_ONTOLOGY database is used for this analysis. The GSEA analysis provides us the information about the biological functions which may got affected in AD.

Tables 3.4 and 3.5 list the common GO terms of dataset 1 and dataset 2 which got down-regulated and up-regulated in the control and AD networks, respectively. All GO terms related to dataset 3 are provided in supplementary data.

Table 3.4: Down-regulated and Up-regulated GO Terms in Control Network.

Down-regulated			Up-regulated		
GO Term	Number of Genes In Dataset 1	Number of Genes In Dataset 2	GO Term	Number of Genes In Dataset 1	Number of Genes In Dataset 2
GO AXON	48	150	GO BIOLOGICAL ADHESION	32	260
GO AXON PART	31	93	GO CELL MOTILITY	40	307
GO CYTOPLASMIC VESICLE PART	39	306	GO DNA BINDING TRANSCRIPTION FACTOR ACTIVITY	32	392
GO DISTAL AXON	25	67	GO DOUBLE STRANDED DNA BINDING	19	250
GO EXOCYTIC VESICLE	23	52	GO NEGATIVE REGULATION OF TRANSCRIPTION BY RNA POLYMERASE II	18	256
GO EXOCYTOSIS	26	190	GO POSITIVE REGULATION OF RNA BIOSYNTHETIC PROCESS	41	437
GO INTRACELLULAR TRANSPORT	49	384	GO POSITIVE REGULATION OF TRANSCRIPTION BY RNA POLYMERASE II	26	335
GO MEMBRANE PROTEIN COMPLEX	33	217	GO REGULATORY REGION NUCLEIC ACID BINDING	22	275
GO NEURON PROJECTION TERMINUS	15	35	GO RESPONSE TO WOUNDING	17	136
GO ORGANELLE LOCALIZATION	29	161	GO SEQUENCE SPECIFIC DNA BINDING	21	296
GO POSTSYNAPSE	38	164	GO SEQUENCE SPECIFIC DOUBLE STRANDED DNA BINDING	18	236
GO PRESYNAPSE	42	121	GO SKELETAL SYSTEM DEVELOPMENT	16	105
GO SECRETORY VESICLE	31	183	GO TRANSCRIPTION FACTOR BINDING	17	225
GO SYNAPSE	72	290			
GO SYNAPSE PART	67	235	GO TRANSCRIPTION FACTOR BINDING	17	225
GO TRANSMEMBRANE TRANSPORT	49	246			
GO TRANSPORT VESICLE	26	27	GO TRANSITION METAL ION BINDING	15	171
GO TRANSPORT VESICLE MEMBRANE	21	46			
GO VESICLE LOCALIZATION	17	83	GO TUBE DEVELOPMENT	24	239
GO VESICLE MEDIATED TRANSPORT IN SYNAPSE	16	54			
GO WHOLE MEMBRANE	49	28	GO ZINC ION BINDING	15	144

GO terms which got down-regulated and up-regulated in control network with p.value ≤ 0.05

Table 3.5: Down-regulated and Up-regulated GO terms in AD Network.

Down-regulated			Up-regulated		
GO Term	Number of Genes In Dataset 1	Number of Genes In Dataset 2	GO Term	Number of Genes In Dataset 1	Number of Genes In Dataset 2
GO CYTOPLASMIC VESICLE PART	46	385	GO CELL MOTILITY	33	401
GO DISTAL AXON	29	91	GO ENZYME LINKED RECEPTOR PROTEIN SIGNALING PATHWAY	21	305
GO EXOCYTIC VESICLE	30	64			
GO EXOCYTOSIS	25	234	GO LOCOMOTION	38	461
GO NEURON PROJECTION TERMINUS	17	46	GO NEGATIVE REGULATION OF RNA BIOSYNTHETIC PROCESS	16	416
GO PRESYNAPSE	51	149			
GO SECRETORY VESICLE	3 5	233	GO POSITIVE REGULATION OF LOCOMOTION	16	152
GO SYNAPSE PART	83	289			
GO SYNAPTIC VESICLE MEMBRANE	22	33	GO POSITIVE REGULATION OF TRANSCRIPTION BY RNA POLYMERASE II	19	413
GO TRANSPORT VESICLE	33	112	GO REGULATION OF CELL POPULATION PROLIFERATION	24	427
GO WHOLE MEMBRANE	49	430	GO TUBE DEVELOPMENT	20	277

GO terms which got down-regulated and up-regulated in AD network with p.value ≤ 0.05

3.3.6 Analysis of Networks

All the genes in the combined networks may be important regarding Alzheimer's disease, but, generally, a gene of interest functions differently in normal and AD affected persons. Hence, for further analysis, we have not considered the genes common to both AD and control networks. We only choose those genes that are only present in either AD or control network, i.e. genes that are present in AD network but not in control network and vice-versa. Hence, both AD and control common networks are analyzed and culled the genes present in the AD network but not in the control network (AD-CTRL) and vice-versa. As a result, 79 such genes are identified in dataset 1, 1107 genes in dataset 2 and 1 gene in dataset 3. Similarly, genes which are present in the control network but not in the AD network are identified. They are 100, 456 and 584 genes in dataset 1, dataset 2 and dataset 3 are respectively. Genes common to both networks are 140, 2031 and 11 in each of the datasets. Fig. 3.3 shows the Venn diagram of different pools in every dataset.

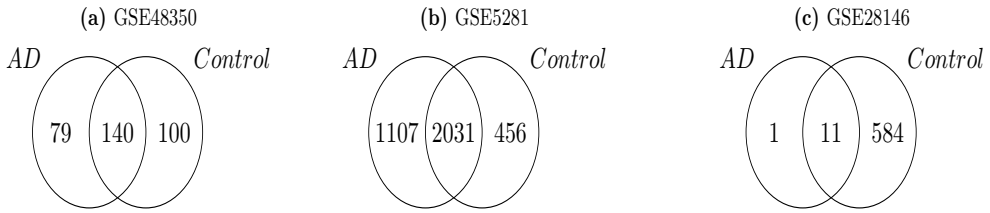


Figure 3.3: Venn diagram analysis of gene in control and AD networks

To further shortlist the genes, feature selection using correlation-based feature subset selection for machine learning algorithms [58] is performed. After performing feature selection, we have obtained 13 genes out of 179 genes (79 + 100, Fig.3.3a) in dataset 1 (GSE48350), 101 genes out of 1563 (1107 + 456, Fig.3.3b) genes, and 54 genes out of 585 (1 + 584, Fig.3.3c) genes, in dataset 2 (GSE5281) as the top ranking genes.

3.4 Comparison with other works

For the comparison purpose, we have considered two recently published frameworks: the first is based on Lasso and random forest (LASSO & RF)[44], and the second is based on t-test, genetic algorithm, and a modified particle swarm optimization algo-

rithm (MPSO)[45]. For a fair comparison, if the number of genes obtained by the frameworks is more than 20, we chose only the top 20 genes for the comparison using *CorrelationAttributeEval* feature selection algorithm in WEKA which ranks the attributes based on their correlation with the class. Table 3.6 shows the top genes obtained from the different frameworks for different datasets. Table 3.7, 3.8 and 3.9 list all the identified genes in dataset 1, dataset 2 and dataset 3 respectively. Genes selected for dataset 3 are provided in supplementary data.

Table 3.6: Different genes selected by the proposed algorithm, Lasso and MPSO.

Framework	Dataset1	Dataset2	Dataset3
tpGONet	ATP2B3, FGF12, MDFIC, NSG1, TAC1, ZNF621, BTK, CD44, CD5, DACH1, ERBB4, RTF1, TAB3	ABI2, ELAVL3, AP2A2, CEP97, ADGRB3, SRRM2, AGFG1, SEC22C, EAPP, AKAP13, TNRC6B, ARHGAP21, CHMP2A, BICD1, FAM120A, COPG1, YTHDC1, INTS3, ERC1, BRD9	ARL8B, PMAIP1, THRB, BHLHE40, ZNF711, BNIP3, DIS3, ZMYM2, HNRNPA0, MAPKAPK2, KPNA6, KBTBD7, MAP2K6, AHNAK, CD44, IL1R1, LRP8, NCOA2, CDH5, ZBTB17
Lasso and Random Forest	ANKIB1, FBRSL1, LOC101927151, RAE1, RTF1, SLC25A46, ZNF621	ARHGAP5, CDK5RAP2, CKMT1A, CKMT1B, DUSP8, FAM120A, FAM168A, FAM63A, KTN1, LOC101927562, OSBPL1A, PEBP1, RHOB, TESK1, ZNF532	BNIP3, CD44, HPS3, MCCC1, NSUN6, ST6GALNAC5
MPSO	ZNF621, LOC101927151, SLC25A46, ANKIB1, RAE1, RTF1	ANKRD12, ELAVL3, ERC1, GPR155, KTN1, NAV1	IL13RA1, DEFB125, RFX4, CXorf38, JAM3, ZFP41, TGFB1I1, TTL

Table 3.7: Genes identified in Dataset 1 (GSE48350)

Gene.symbol	P.Value	logFC
ATP2B3	0.000131	-0.9364480
BTK*	1.05e-05	-0.8254824
CD44*	0.000118	1.04783081
CD5	0.00123	0.87985092
DACH1	2.8e-05	0.94417236
ERBB4*	0.00074	0.82142565
FGF12	0.00134	-1.1104277
MDFIC	0.000417	0.8739675
NSG1*	0.001075	-1.05723075
TAB3	0.00173	1.00097339
RTF1	1.46e-17	-1.7616841
TAC1*	0.0139	1.32196006
ZNF621	2.32e-24	2.96528119

*Identified in literature.

Table 3.8: Genes identified in Dataset 2 (GSE5281)

Gene.symbol	P.Value	logFC	Gene.symbol	P.Value	logFC
ABI2	0.0067832	0.961026636667	INTS3	2.79E-06	-1.37560425
ACBD5	1.25E-06	2.3668954	IPO7	0.000274	0.85463051
ACO1	0.000295	-0.98042327	JPH3	0.000359	2.11788507
AGFG1	9.36E-06	2.19387208	KDM5A	0.00524225	1.9725445
ACTR1B	0.000606	-0.97274028	KIF1A	2.31E-05	1.27250433
ACVR1B	0.000422	1.39831289	KTN1	4E-10	2.34921503
ADAM22	0.0216053333333	0.53768401	L1CAM	4.46E-07	1.83246955
ADAM23	4.08E-06	-1.73163989	LAMP1	0.0038554085	-1.73519342
ADCY2	2.95E-05	1.31786542	MAGI2	3.35E-08	3.020489
AKAP13	0.00647204185714	1.04407532714	MAP6	6.54E-07	1.89564215
AKAP8L	0.000989	1.87231609	MARK3	0.0001442	-0.04287374
ANK3	0.000124	2.44887394	MIB1	3.95E-09	2.14433497
AP2A2	0.000291433333333	-1.11659461	MORF4L2	0.001740079	2.404579215
AP3D1	0.00770605	1.39260274	MRPS5	0.000319606666667	2.40202342333
ADGRB3	6.93E-07	2.46317656	NAP1L4	2.12E-05	-1.09098646
ANKRD11	0.00421	1.24675	NEUROD1	4.28E-05	1.56594988
ARHGAP21	4.55E-08	2.64329836	CBX3	0.00117	2.43672
BBX	0.003446645924	1.888399814	NR2F2///NR2F1	5.13E-05	1.37623146
BICD1	7.65E-07	1.88214093	NSL1	0.014850297	1.450132955
BNIP3L	0.000345	0.94418923	NUCKS1	2.05365E-06	0.11134135
BRD9	4.94E-06	-1.80179381	PNISR	0.000345876666667	1.60721201
C12orf10	0.00169	-0.82914204	PRKAB2	0.0002555	1.431380315
			PTBP3	4.61E-05	1.91441374
CAMSAP2	0.00028575	0.050696345	PTP4A1	0.005201145	0.24193515
CAPRIN2	4.8E-06	-1.48271188	PTPRJ	4.73E-06	-1.34847657
CBL	1.29E-06	1.32748593	REV3L	1.08E-06	1.33991121
CEP97	0.00550010933333	1.99005593	RFK	4.39E-05	1.1322082
CHMP2A	5.25E-05	-1.91874975	KDELR2	3.08E-05	-1.07980691
CLN8	6.19E-06	1.24748557	SEC22C	9.851395E-05	1.871307925
COPG1	3.09E-08	-2.10125131	SLC25A36	0.010800795	1.20768185
CORO1C	1.01E-05	-1.35341065	SLC8A1	8.59E-08	1.66146541
CTSC	0.005502715	1.421647035	SRRM2	0.0028989782	2.03190842
DGKG	1.35E-05	2.67416378	STOML2	6.49E-08	-2.13214491
EAPP	1.16E-08	-1.71677758	SUZ12P1///SUZ12	0.000103	0.8732712
EIF5B	0.00467275	1.32003578667	TBL1XR1	1.155237E-07	-0.45937598
ELAVL3	1.33E-10	2.89606404	TNPO2	9.1E-07	1.7570126
ELMO1	3.08E-06	-2.49796439	TNRC6B	2.17015266667E-05	1.56078730333
ERC1	4.31E-10	2.47720499	TRIM23	1.72E-05	1.39388483
ERCC3	2.21E-06	-1.46854842	MNT	5.07E-05	1.43479172
MICAL1	0.0234	1.22246841	PABPC3	3.48E-06	1.14393254
ESF1	0.009050321	1.38541814	UNKL	3.21E-05	1.08877527
FAM120A	0.000426554166667	1.51672673333	USP10	0.000656245	1.216217585
UHMK1	1.07E-06	1.8235725	WDR82	6.58E-06	3.907264
ZNF532	1.66E-09	-2.46526681	YTHDC1	0.00066	1.7781
GALNT1	0.00010224	1.28699208	ZBTB1	0.00018795	2.059890885
GLG1	9.93E-09	1.55542645	ZMAT3	9.7E-06	1.33505671
GOLGA2	0.00525002385	1.82953746	ZNF148	0.000365265	1.22672275
GRIN2B	0.0004368415	1.678765335	ZNF264	4.16E-06	1.46456226
GRK3	1.63E-09	2.31565161	ZNF652	0.0006725	1.62317446
HSPH1	5.07E-06	1.55584984	ZNF770	7.36E-05	1.06839715
INO80D	0.0056500398	0.291794275	MPRIP	2.1775E-05	-2.412150

Table 3.9: Genes identified in Dataset 3 (GSE28146)

Gene	P.Value	logFC	Gene	P.Value	logFC
AHNAK	0.00262204	0.98198608	MAP2K6	0.01595241	0.87513261
ARID2	0.00729477	1.10568321	MAPKAPK2	0.03803999	0.99704209
ARL8B	0.00339015	-1.06449196	MMP25	0.00227936	1.63779073
BABAM1	0.01135208	-1.23711096	NCOA2	0.03326324	1.18257032
BHLHE40	0.00038624	1.28253737	NR2F2	0.01002185	-1.17016344
BNIP1	0.00758357	-1.43055366	PHLDB2	0.01342753	1.61673762
BNIP3	0.00434216	-0.9784713	PMAIP1	0.00032392	2.12952376
BOK	0.00031164	1.62960415	PRDM9	0.01930253	1.59091908
CD44	0.00192801	1.22502707	PRMT2	0.01439603	-1.2391834
CD44	0.0060737	1.48449061	PTBP1	0.0091615	1.04351217
CDH5	0.00999381	0.94147819	RBL2	0.01267011	1.13216523
CHRNA3	0.01426888	-1.10012937	RBM15B	0.00410289	1.01976731
CITED2	0.0007325	-1.17412862	RC3H2	0.00068678	-0.81815358
CLN6	0.01049238	-1.48885166	REV1	0.02505561	-0.98487729
CYBB	0.00043958	2.09599291	RHOT1	0.00983653	-2.06345313
DIO2	0.01657926	-1.08859571	SELPLG	0.00102717	1.75407251
DIS3	0.03408111	0.86868341	SIRPB1	0.01981797	-1.41708899
FGD6	0.01946995	-1.41193122	SORBS1	0.00527204	0.83808703
GFM1	0.01715968	-1.08458122	STEAP4	0.03592597	-0.8679121
GLI3	0.00561154	-1.13185341	SYNGAP1	0.0178071	-1.05676322
GRIA4	0.03489461	-1.07582032	THRB	0.00547597	-0.98552353
HNRNPA0	0.01948719	-0.84039497	TMEM88	0.00293001	-1.7443467
IL1R1	0.00519789	1.35390824	TYMS	0.00662579	1.86512571
IRAK3	0.00155661	2.15419575	ZBTB17	0.00436392	-1.10062518
KBTBD7	0.00198938	-1.36386637	ZMYM2	0.003935	-1.66437016
KPNA6	0.00782112	-1.44454199	ZNF174	0.01271933	-1.28963757
			ZNF711	0.00172779	-1.16541981
LRP8	0.03749307	-1.11916343	ZNF91	0.01719465	-1.19189899

As observed from Table 3.6, the significant genes obtained by all frameworks are almost different for all the datasets. This does not provide us with any inference. Therefore, we compared the degree of specificity of genes obtained by the tcGONet, LAASO & RF and MPSO, towards Alzheimer's disease. We checked the direct interactions of

the genes obtained with the AD pathway genes using the STRING database. We did not find any common pattern in the number of interactions, making it difficult to draw any conclusion. We further used DAVID[1] to obtain the diseases in which the genes are involved after feature selection which did not yield significant results as the number of genes is less, and some are not characterized. It is well known that interacting proteins regulate the function of a protein [59]. So retrieving the interacting partners and the associated diseases can give us a deeper insight into the genes obtained from our framework. HIPPIE[2] is used to fetch the high confidence primary interacting proteins of the genes obtained from our analysis. The primary interacting genes are then subjected to DAVID analysis to obtain the corresponding diseases.

It is observed that in dataset 1 and dataset 2, primary interactions of the genes obtained by the tcGONet are directly associated with Alzheimer's disease with high significance. In contrast, the interacting partners of genes obtained from other algorithms are not at all related to any neurological disorders. Although in dataset 3, genes obtained from the tcGONet, LASSO & RF and MPSO framework have interacting partners associated with Alzheimer's disease. However, it is interesting to note that the significance and count of genes associated with AD in the tcGONet are quite high compared to the LASSO & RF and MPSO framework. The supplementary data provides the table of all the diseases related to the genes, the gene count, and their corresponding p-values.

3.5 Results

Using the framework introduced, we are able to identify genes in all datasets that are directly or indirectly related to AD with a high classification power. Table 3.7 and 3.8 list the genes identified. The link between the identified genes and the AD pathway genes is analyzed to find out the importance of the identified genes in this work. As a result, it is found that most genes have either direct or one-hop interaction with the AD pathway genes. Table 3.10 shows some direct interactions between top genes of dataset 1 and AD pathway genes. As the top genes in both datasets are different, we tried to determine the relationship between both datasets' top genes. STRING database[1] is

[1]https://david.ncifcrf.gov/
[2]http://cbdm-01.zdv.uni-mainz.de/ mschaefer/hippie/
[1]https://string-db.org/

used to find interactions between the genes, only interactions that are experimentally proven are from the curated database with at least medium confidence value 0.4(as mentioned in STRING database) are considered. All the top genes of dataset 1 have either direct or one-hop connections with at least one top gene of dataset 2 (a few of the interactions are shown in Table 3.11). We also checked the GO similarity between the top genes of both datasets and the GO similarity of top genes with the AD pathway genes to find the similarity between them. Also checked the primary interactions of the identified genes and found them related to AD with high significance compared to the genes identified by other considered frameworks. All the interactions, GO similarity, disease-associated and primary interactions files can be found in the **"Supplementary Data"**.

Table 3.10: STRING interactions between top genes of dataset 1 and AD pathway genes.

Top Genes of Dataset 1	AD Pathway Genes	STRING Interaction Score
ATP2B3	CALM1	0.69
BTK	FAS	0.935
ERBB4	PSEN1	0.9
FGF12	CALM1	0.96
TAB3	TNF	0.902
TAC1	APP	0.9

In dataset 1, out of 13 genes identified, 5 (BTK, CD44, ERBB4, NSG1, and TAC1) are found to be related to AD in the recent literature. Similarly, many genes (ADAM22, AGFG1, GRIN2B, MPRIP, ZNF532 etc.), identified in dataset 2 are listed in the literature.

- Gene ATP2B3[1] has human phenotype ontology of ataxia, cerebellar atrophy, cerebellar hypoplasia, clumsiness.
- Gene FGF12 has a human phenotype ontology of abnormal myelination, abnormality of vision, absence of speech.
- In [60], Keaney et al. observed that the activation of phospholipase gamma 2, a genetic risk factor in AD, is decreased due to the blockade of **BTK**.

[1] www.genecards.org

Table 3.11: STRING interactions between top genes of dataset 1 and data set 2.

Top Genes of Dataset 1	Top Genes of Dataset2	STRING Interaction Score
BTK	CBL	0.95
CD44	ANK3	0.8
ERBB4	GRIN2B	0.9
ATP2B3	CALM1	0.69
CALM1	ADCY2	0.64
CD5	CD4	0.861
CD4	AGFG1	0.9
DACH1	NCOR1	0.426
NCOR1	TBL1XR1	0.98
FGF12	CALM1	0.96
CALM1	ADCY2	0.64
MDFIC	CTNNB1	0.9
CTNNB1	TBL1XR1	0.935
FGF12	SRPK2	0.442
SRPK2	SRRM2	0.442
TAB3	MAP3K7	0.986
MAP3K7	PRKAB2	0.817
RTF1	SUPT16H	0.995
SUPT16H	ERCC3	0.9
TAC1	TACR1	0.965
TACR1	AGFG1	0.9
ZNF621	TRIM28	0.922
TRIM28	ZNF770	0.902

- In [61], Elhanan et al. investigated the expression values of **CD44** splice variants in the hippocampus region of AD patients and compared it with the control patients and observed that the expression values of splice variants of CD44 are significantly higher in AD patients when compared to the normal person. The research suggested that some splice variants of CD44 contribute to AD pathology.

- Ran-Sook Woo et al. [62] found that up-regulation of the immunoreactivity of **ERBB4** may involve in Alzheimer's disease progression.

- In [63], Abhik Ray Chaudhury et al. observed that Neuregulin-1 and **ERBB4** immunoreactivity is associated with plaques formation in the AD brain.

- Eric M. Norstrom et al. [64], report that NEEP21 protein (gene name: **NSG1**), affects the processing of APP and Aβ production.
- Marco Magistri et al. [65] analyzed that in the hippocampus region of the brain in AD patients, TAC1 is downregulated compared to controls hippocampus.

The GSEA analysis shows that out of 22, 12 GO terms that got down-regulated in the control network are not being regulated in the AD network and vice-versa, which indicates that there may be a disturbance in the regulation of those 12 GO terms. Similarly, out of 21, 18 GO terms that got up-regulated in the control network is not being regulated in the AD network vice-versa.

In the GO terms that are identified, we find that some are found to be disturbed in the Alzheimer's disease, like, GO SYNAPTIC VESICLE MEMBRANE, GO AXON, GO TRANSPORT VESICLE MEMBRANE, GO VESICLE MEDIATED TRANSPORT IN SYNAPSE, GO NEGATIVE REGULATION OF RNA BIOSYNTHETIC PROCESS, GO REGULATION OF CELL POPULATION PROLIFERATION, GO NEGATIVE REGULATION OF TRANSCRIPTION BY RNA POLYMERASE II, GO RESPONSE TO WOUNDING, GO SKELETAL SYSTEM DEVELOPMENT, GO POSITIVE REGULATION OF RNA BIOSYNTHETIC PROCESS and GO ZINC ION BINDING.

K. Blennow et al. [66], found that the level of **synaptic vesicle membrane** protein rab3a was reduced in Alzheimer's disease in the hippocampus. In the literature, it is found that in Alzheimer's disease, the amyloid-beta disturbed the **vesicle transport in synapse** in the hippocampus [67, 68]. In [69], Wo Y. et al. observed that the **cell proliferation** gets slowdown when the APP is over expressed. Nicole T. Watt et al. [70] discussed the role of **Zinc** in Alzheimer's disease. Zinc binds to amyloid-beta, advancing its conglomeration into neurotoxic species, and disturbance of zinc homeostasis in the brain results in synaptic and memory deficiencies. Kiecolt et al. [71], observed that **wound healing** took a long time significantly in AD patients than in controls. Chen et al. [72] conclude that AD increase the risk of osteoporosis (**Skelton disorder**). The overexpression of amyloid-beta might happen in both cerebrum and bone, meddling with the RANKL signalling cascade, improving osteoclast activities, and prompting osteoporosis.

3.6 Conclusions

Most often, t-test and correlation are used to identify significant genes at the initial level. As the genes are differentially expressed, their classification power is generally high. These genes might appear significant, but their degree of specificity towards the disease might be low, leading to misleading interpretations. Similarly, there may be many false correlations between the genes that can affect the identification of relevant genes. This work introduces a new framework to reduce the false correlations and find the potential biomarkers for the disease. The framework concerned uses the t-test, correlation, Gene Ontology(GO) categories, and machine learning techniques to find potential genes. The tcGONet detects Alzheimer related genes in every dataset considered. Some of the genes identified which are directly involved in Alzheimer are APP, GRIN2B and APLP2. The tcGONet also identifies genes like ZNF621, RTF1, DCH1, ERBB4, which may play an important role in Alzheimer's. Gene set enrichment analysis (GSEA) is also carried out to determine the major GO categories: down-regulated and up-regulated.

3.7 Summary

In summary, in this chapter, a framework that includes t-test, correlation, GO categories, feature selection methods is developed to identify the potential biomarkers for Alzheimer's disease. The GO categories are analyzed and used to create a more biologically significant network, which helps in eliminating false correlations. Feature selection is used to list out the top genes.

Biological interactions between the top genes of all datasets are studied in which the top genes either have direct or one-hop experimentally proven interactions with one another. Biological interactions between top genes and AD pathway genes are also studied. As a result, many of the genes were found to have direct experimentally proven interactions with the AD pathway genes. Primary interactions of selected genes show that the genes selected by the tcGONet are associated with Alzheimer's disease.

Gene set enrichment analysis of AD and control networks is also carried out and found that GO terms which got up-regulated/down-regulated in AD network but not in control network and vice-versa, may get disturbed in Alzheimer's disease. The literature shows that the genes identified by the decision tree classifier whose logFC values indicate

that these genes that need to be up-regulated are down-regulated and vice versa. The results consist of the genes and GO terms that are related to Alzheimer's disease in the literature, which adds more credibility to the results. The results show that the genes identified by the tcGONet have a high degree of association with AD in comparison to the genes identified by the other frameworks considered. In future, the tcGONet can be applied to other diseases too, and an automated tool based on the tcGONet can be developed.

Chapter 4

Stage-wise Commmunity Analysis of Alzheimer's Disease Networks

Network representation has emerged as the popular method of real-world complex systems representation in the past few years. Community structures play a significant role in analysing complex networks like social networks, biological networks, computer networks and various other kinds of networks. Community detection enables us to discover same set of nodes on the basis of an area of interest. However, there are still some issues with the community discovery algorithms which remain unaddressed such as tightness relationship between the nodes and some cases of conflicts which creates ambiguity in determining he node's best-fit community. In this chapter have introduced a novel neighbour-based community discovery algorithm, NBCD is proposed to address these issues. NBCD then is used to identify the potential genes of Alzheimer's disease.

4.1 Background

4.1.1 Community

A community in a network is taken to be a subset of nodes within the graph such that connections between the nodes within the community are denser than connections with the rest of the network [73].

There are many community discovery algorithms proposed in the literature.

Broadly, there are two approaches for community detection; the first is the optimization based approach, which optimizes a defined criterion. For example, Greedy Modularity, looks for Modularity optimization. The second is the non-optimization-based community detection approaches like LPA, Walktrap, neighbour-based similarity algorithms, etc.

Tightness relationship among nodes within the communities is an important issue. Within communities, tightness refers to the degree of connectivity between nodes that belong to the same community. Nodes within a community are typically more tightly connected to each other than they are to nodes outside the community. This high level of internal connectivity is what distinguishes communities from the rest of the network. Modularity is a popular measure to check the tightness of nodes within the communities and sparsity of the communities and hence many popular community discovery algorithms look for Modularity optimization. Although, Modularity suffers from resolution limit [27], that is, unable to detect the small communities resulting in low F-score. In addition, this also suffers from high degeneracy [74], that is, more than one community structure with equally high Modularity which leads to a conflict/confusion.

Furthermore, in community discovery there are other issues like handling the conflicts, that is, selecting the best-fit community of a node when the node belongs equally to more than one community. This can lead to inaccurate communities at the initial stage and makes it difficult to reach a high-quality community structure eventually. For instance, consider a case where there is a bridge node with degree two, and both of it's neighbours belong to two different communities. Then there arises confusion about the bridge node's community. Similarly, confusion occurs when a node has the same similarity score with two or more nodes of different communities. A single wrong prediction may lead to a wrong prediction for other nodes too which may result in detecting a poor community structure.

4.1.2 Community Discovery in Biological Systems

In the field of biology, community detection has been applied to various problems, such as identifying functional modules in protein-protein interaction networks, studying the organization of food webs, and analyzing the spread of diseases in contact networks.

One example of its application is in protein-protein interaction networks, where community detection algorithms can be used to identify groups of proteins that are likely to

function together in a common biological process. This can help researchers understand the organization of the cell and predict new interactions between the proteins. Another example is in the spread of diseases in contact networks, where community detection can be used to identify subgroups of individuals that are more likely to spread an infection to one another. This can help public health officials target interventions to the most at-risk populations and slow the spread of the disease. Overall, community detection is a powerful tool that can help researchers and practitioners better understand the structure and organization of complex networks in biology.

In [75], Cantini et al. explore the use of gene communities as a means of identifying key players in the development of cancer. The study utilizes a multi-network approach, which involves combining multiple types of genomic data to better understand the underlying mechanisms of cancer. They applied the community detection method to three different types of genomic data: protein-protein interaction networks, co-expression networks, and somatic mutation networks. By combining these three networks, the authors were able to identify gene communities that were highly enriched for cancer-related genes.

Calderer wt al. in [76] discuss the use of community detection algorithms to identify groups of interacting elements within large-scale bipartite biological networks. The authors tested several community detection algorithms on large-scale bipartite networks of protein-protein interactions, gene-gene interactions, and gene-disease associations. They found that these algorithms were able to identify biologically meaningful communities, such as groups of proteins that function in similar pathways or groups of genes that are associated with specific diseases within the networks. The study also found that the performance of the community detection algorithms varied depending on the type of bipartite network and the specific algorithm used. For example, some algorithms performed better on protein-protein interaction networks, while others performed better on gene-disease association networks.

In [77], Wilson et al. used the community detection techniques to identify functional and disease pathways in protein-protein interaction networks. The authors used a community detection algorithm called "modularity optimization" to identify groups of proteins that are highly interconnected and likely to play a role in specific biological processes or diseases.

In [78], M'barek et al. present a novel approach for identifying communities or

groups of highly interconnected nodes in biological networks. The authors propose the use of a genetic algorithm (GA) for community detection, which is a computational method that mimics the process of natural selection to find optimal solutions to a problem.

In [79], HU et al. propose a new method for detecting communities in biological networks, specifically focusing on the identification of disease-causing genes. The authors propose a multi-scale approach that uses a significance-based algorithm to identify communities of genes that are likely to be involved in a specific disease.

In [80], Singhal et al. introduced a new method for community detection in networks, called multiscale community detection (MCODE). The authors developed the MCODE algorithm as a plugin for the Cytoscape software, which is a popular tool for visualizing and analyzing networks. The main finding of the study is that MCODE is able to detect communities in networks at multiple scales, which means that it ca n identify both big and small groups of highly interconnected nodes. This is in contrast to traditional community detection methods, which often only identify large groups of nodes. The authors demonstrate the effectiveness of MCODE by applying it on several different types of networks, including protein-protein interaction networks, metabolic networks, and social networks.

Allen et al., in [81] proposed a statistical network model BANYAN (Bayesian ANalysis of communitY connectivity in spAtial single-cell Networks) which is capable of discerning community connectivity structure in high throughput spatial transcriptomics data. In [82], Dilmaghani et al. used deferential network analysis on RNA-sequencing (RNA-seq) time series datasets. Then they applied community detection algorithms on deferential networks to understand the temporal behaviour of genes. In [83], Pathak et. al proposed a new local community detection algorithm, (lcda-go), which detects the communities based on GO functions and network topology. However, there is not much study on community detection in temporal disease networks.

We in this work proposed the NBCD algorithm to analyze the stage wise networks of Alzheimer's disease for Hippocampus region. A thorough study of genes whose neighbourhood (community) is changed drastically in the next stage of AD is carried out.

4.2 Neighbour-Based Community Discovery Algorithm(NBCD)

The proposed algorithm finds the similarity between the node and its neighbours. The similarity measure and similarity parameter α are introduced to find the similarity between two nodes which is discussed in Section 4.2.1 of the paper. In order to further increase the quality of communities, the NBCD algorithm in phase 2 shifts the nodes to the communities where most of the node's neighbours are present. The NBCD algorithm can handle cases where there may arise a confusion in selecting the node's community. The "friends of friend" concept and similarity parameter α take care of the tightness relation between the nodes. At present, the NBCD algorithm detects only non-overlapping communities, i.e. disjoint communities.

4.2.1 Similarity Measure

Two novel similarity measures are introduced here. The first phase of NBCD works using these similarity measures. The concepts from social network analysis are used in forming these similarity measures. The first similarity measure, $sim_{nn}(x,\ y)$, checks the similarity between two nodes. This works on a "Friends of Friend" concept, i.e. friends who share a certain number of mutual friends are most probably are alike in some sense. For example, they may have the same set of interests or may belong to the same school/college.

In the same way, the second similarity measure between the node and community, $sim_{nc}(x,\ c)$, points to the fact that if one's majority of friends belong to a particular community, there are more chances that the person also belongs to the same community. According to first similarity measure, $sim_{nn}(x,\ y)$, if node y is a neighbour of the node x and have $(100/\alpha)\%$ number of same neighbours, then node y is similar to node x and belongs to the same community as x, where α is the similarity parameter. The similarity function, $sim_{nn}(x,\ y)$ returns 1, if node y is similar to x or else returns 0.

$$sim_{n,n}(x,y) = \begin{cases} 1, & if \quad Count[Nbr(x) \cap Nbr(y)] + 1 > \frac{Deg(y)}{\alpha}, Deg(y) > 2 \\ 1, & if \quad Deg(y) \leq 2 \\ 0, & otherwise \end{cases} \tag{1}$$

The second similarity measure, $sim_{nc}(x,\ c)$, tells whether the node x belongs to community c or not. It returns 1 if x belongs to community c, -1, to take action according to defined ground rules (see subsection 4.2.2) and 0, if x does not belongs to community c.

$$sim_{n,c}(x,c) = \begin{cases} 1, & if \quad Count[Nbr(x) \cap mem(c)] > \frac{Deg(x)}{\alpha}, Deg(x) > 2 \\ -1, & if \quad Deg(x) \leq 2 \\ 0, & otherwise \end{cases} \tag{2}$$

In both $sim_{nn}(x,y)$ and $sim_{nc}(x,c)$, $Deg(x)$ and $Nbr(x)$ represent the degree and neighbours of x respectively. $mem(c)$ represents the members of community c. Through the similarity parameter α, the user can choose the tightness of the community i.e. how well the nodes are connected in a community. To detect the best value of α, different values of α are taken, and the performance measures are compared for all the datasets considered. The values of α are taken in the range of 1.5 to 3 with an increment of 0.5. $\alpha = 1.5$ means 66.6% similarity and $\alpha = 3$ means 33.33% similarity.

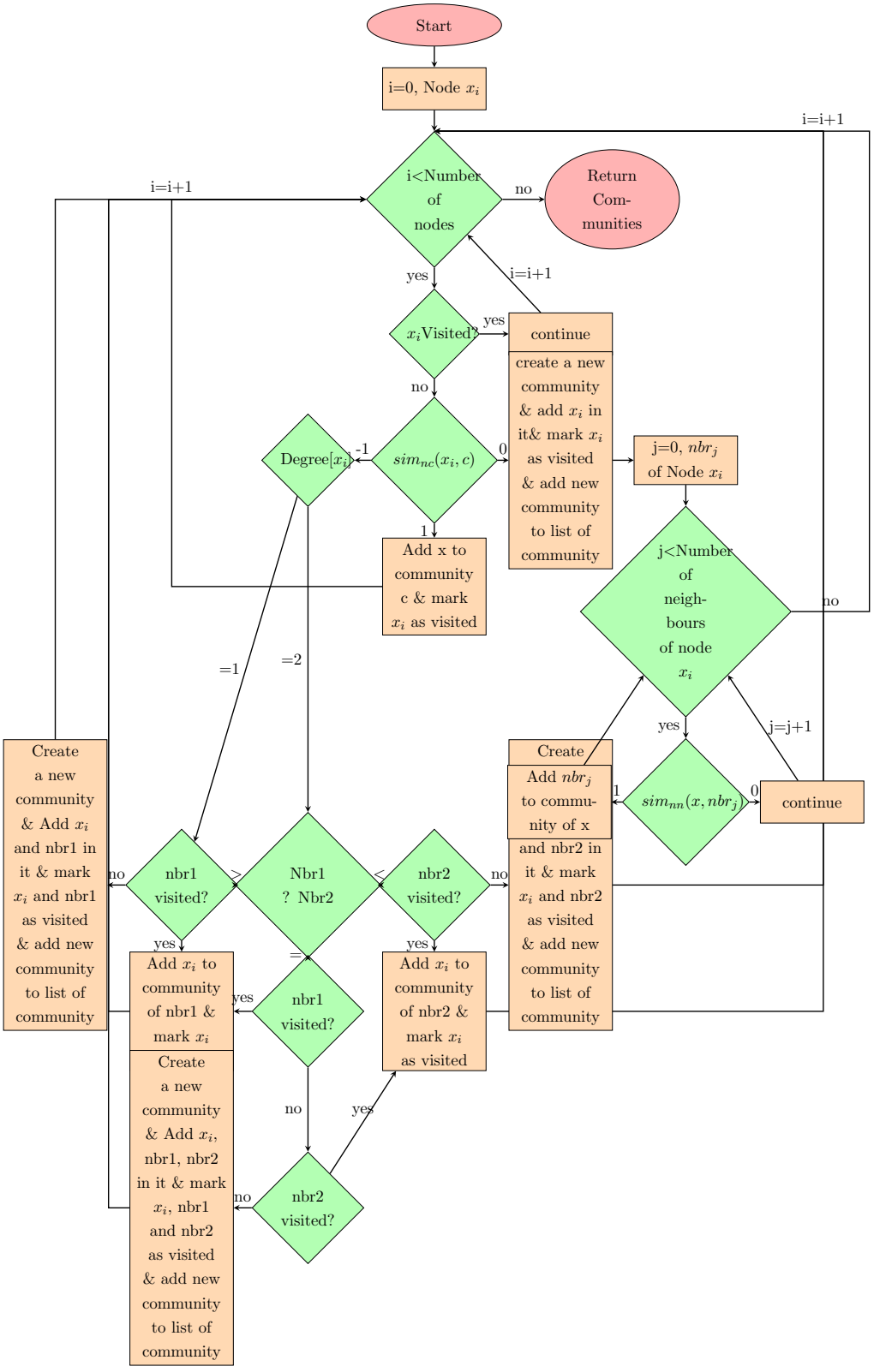
4.2.2 Basic Steps of the algorithm

The proposed algorithm(NBCD) works in two phases; Clustering of nodes based on the neighbour's similarity and reshuffling of the nodes among the detected communities. The NBCD[1] algorithm is implemented in Python(Python 3.8.10) language. The proposed method is described in Algorithm 1 and 2. The flow chart of NBCD is shown in Fig. 4.1, where *Nbr(x)* represents list of neighbours of node *x, nbr1* and *nbr2* represents the neighbour one and two of node x (degree=2) and *nodes* is the list of nodes in descending order according to the degree. Some ground rules are constructed to handle the conflicts. The rules are based on a straightforward concept, "Majority wins"; whenever a conflict happens between nodes or communities, the conflict is resolved in favour of the node and community, which has more connections(degree) and members, respectively. For example, according to the ground rule, if a node has a degree of two, i.e. only two neighbours, the node will go with the neighbour with more connections, i.e. high degree. Similarly, if a node has an equal number of neighbours in two or more communities, then the node belongs to the community with more members. The ground rules defined for the proposed algorithm are as follows:

[1] https://drive.google.com/drive/folders/1UD7Lat3I4M5L187KYGO8srDhT9fxTUsJ?usp=sharing

- Nodes in the descending order of the degree are given as input to the algorithm.
- Node y, a neighbour of node x, belongs to the community of node x if $sim_{nn}(x,y) = 1$.
- If node y is the neighbour of node x and the degree of node y is one, y belongs to the community of node x.
- Nodes with degree two belong to the community of its neighbour with the highest degree.
- A node belongs to a community which have the highest number of its neighbours.
- A node having an equal number of neighbours present in different communities belongs to the community where the sum of the degrees of the node's neighbours is the highest.

Figure 4.1: Flow Digram of NBCD's Phase 1 (community allocation).



Algorithm 1: Stage 1: Community Allocation

Input: Graph(G), α (Similarity parameter)
Output: Detected Communities

```
list_of_community[]          // Initially empty
nodes[ ]          //list of nodes in descending order according to degree
for  x in nodes[ ] do
    if  x not visited then
        for community in list_of_communities do
            if sim_nc(x, community)==1 then
                Add x to the community
                Mark x as visisted
                community_found=True
                break
        if community_found=True then
            continue
        else
            if  sim_nc(x, community)==0 then
                Create a new community
                Add x to the new community
                Mark x as visited
                for  n in Neighbours[x] do
                    if  sim_nn(x, n)==1 then
                        Add n to the new communty
                        Mark n as visited
                Add new community to list_of_community[]
            else if sim_nc(x, community)==-1 and Deg[x]==2 then
                neighbour_1=Neighbour 1 of x
                neighbour_2=Neighbour 2 of x
                if  Degree[neighbour_1]>Degree[neighbour_2] then
                    if  neighbour_1 is visited then
                        Add x to the community of neighbour_1
                        Mark x as visited
                    else
                        Create a new community
                        Add x and neighbour_1 to the new community
                        Mark x and neighbour_1 as visited
                        Add new community to list_of_community[]
                else if  Degree[neighbour_2]> Degree[neighbour_1] then
                    if  neighbour_2 is visited then
                        Add x to the community of neighbour_2
                        Mark x as visited
                    else
                        Create a new community
                        Add x and neighbour_1 to the new community
                        Mark x and neighbour_1 as visited
                        Add new community to list_of_community[]
                else
                    if  neighbour_1 is visited then
                        Add x to the community of neighbour_1
                        Mark x as visited
                    else if  neighbour_2 is visited then
                        Add x to the community of neighbour_2
                        Mark x as visited
                    else
                        Create a new community
                        Add x, neighbour_1 and neighbour_2 to the new
                          community
                        Mark x, neighbour_1 and neighbour_2 as visited
                        Add new community to list_of_community[]
            else
                n=Neighbours[x]
                if  n is visited then
                    Add x to n's community
                    Mark x as visited
                else
                    Create a new community
                    Add x and n to the new community
                    Mark x and n as visited
                    Add new community to list_of_community[]
```

Algorithm 2: Stage 2: Node Shifting

Input: Detected communities and it's Modularity(Old_modularity) in Phase 1
Output: Final Communities

```
do
    for x in Nodes do
        Shift x to the community where the maximum number of it's neighbours are
          present. If two or more communities contain an equal number of neighbours,
          then shift x to the community whose neighbours have the highest degree (Sum of
          the degree of all neighbours).
    New_modularity= Modularity(community)
    If New_modularity > Old_modularity then
        Old_modularity=New_modularity
        shift=1
    Else
        shift=0
while shift ==1
Delete empty communities
```

4.2.3 Example

To demonstrate the technique intuitively, we take the Karate dataset as an example. For demonstration, we are considering $\alpha = 2$ as the similarity parameter for NBCD.

The Karate network consists of 34 nodes and 78 edges[84]. In the community detection phase, initially, the NBCD starts with the node having the highest degree, which is node '34'. Since the list of communities is empty; NBCD creates a community, *c[0]*, and adds node '34' to it. All the neighbours of node '34' are then checked for similarity using the similarity measure $sim_{nn}(x, y)$. Where x is node '34' and y is the neighbour of node '34'. For example, Nodes '10', '15', '16', '19', '21', '23' and '27' are neighbours of node '34' and their degree is two. So according to similarity measure these nodes will be directly added to the community *c[0]*. Another case is node '9', which is a neighbour of node '34' whose degree is greater than two. So the similarity measure $sim_{nn}(x, y)$ checks for the common neighbours between nodes '34' and '9'. Both nodes have two common neighbours, nodes '31' and '33'. So the conditions are satisfied:

$$Count[Nbr(34) \cap Nbr(9)] + 1 > \frac{Deg(9)}{\alpha}$$

$$2 + 1 > \frac{5}{2}$$

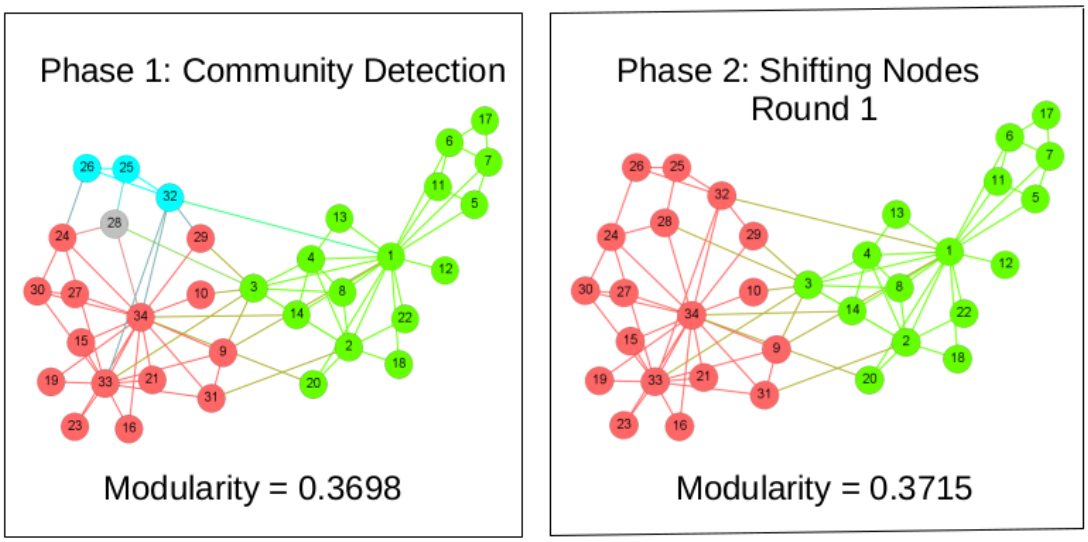
$$\therefore sim_{nn}(34, 9) = 1$$

That is node '9' is similar to node '34' and hence is added to the community *c[0]* and marked as visited. Similarly, the similarity between node '34' and its remaining neighbours is checked, and similar neighbours are added into the community *c[0]* and marked as visited. At the end the community *c[0]* is added to the list of communities *C*. Now nodes '34', '9', '24', '29', '30', '31', '33', '10', '15', '16', '19', '21', '23' and '27' are in the community *c[0]*.

The node with the highest degree in the remaining nodes is then selected, which is node '1'. Now, node '1' is checked to see whether it belongs to any community present in the list of communities or not, through the similarity function $sim_{nc}(x,\ y)$ for all communities in the list of communities, C, where x is node '1' and c is the community in the list of communities, C. As node '1' do not belong to any community present in C, a new community, *c[1]*, is created and node '1' is added in *c[1]*. Again all the neighbours of node '1' are checked for the similarity using similarity function $sim_{nn}(x,\ y)$. The process continues till all nodes are not visited. As the result of community detection phase, NBCD detects 4 communities $C\{$*c[0], c[1], c[2], c[3]*$\}$, which are shown in Fig 4.2, where different colours represent different communities. The Modularity of community structure detected in phase 1 is 0.3698.

All nodes are checked for their best-fit community in the shifting phase, i.e., where their maximum number of neighbours are present. In round one, the current community of node '32' is *c[2]* and a maximum number of its neighbours are present in *c[0]*, so node '32' is shifted from community *c[2]* to *c[0]*. Similarly, node '28' is shifted from *c[3]* to *c[0]*, and nodes '25' and '26' are shifted from *c[2]* to *c[0]*. The degree of node '10' is two, and its current community is *c[0]*. One neighbour '34' is present in community *c[0]*, and another neighbour '3' is present in *c[1]*. No shifting is done because the degree of node '34' is greater than that of node '3'. After round one is finished, the Modularity of community structure is 0.3715. As the new Modularity is approximately equal to the previous Modularity, NBCD won't go for the next shifting round. NBCD only go for the next round of shifting if the Modularity increases by 0.01 i.e.1%. Hence, NBCD terminates and gives the final set of communities.

Figure 4.2: Steps of NBCD on the Karate Network.



4.2.4 Experiments

Proposed NBCD algorithm is compared with six other popular algorithms: Walktrap, Greedy Modularity, Label propagation algorithm(LPA) [85], Louvain and Eigenvector, and also with five recently published algorithms: Synwalk, FPPM, DSLPA, SimCMR and NSA, to determine the efficiency of NBCD. Walktrap, Louvain and Eigenvector from the cdlib package(python) [86], and LPA and Greedy Modularity algorithms are taken from the Networkx package(python) [87]. For DSLPA, we used the parameter values mentioned in [88]. Experiments are performed on HP-g6-Notebook with Intel Core i5-3210, 4GB RAM and Ubuntu 20.04.3 LTS (Focal Fossa).

Table 4.1: Worst-case time-complexities of considered community discovery algorithms.

Algorithm	**Time-complexity**	**Remarks**	**Source**
Walktrap	$\mathcal{O}(m.n^2)$	n = number of nodes, m = number of edges.	[89]
Greedy Modularity	$\mathcal{O}(m.d \log n)$	n = number of nodes, m = number of edges, d = depth of the dendrogram.	[25]
LPA	$\mathcal{O}(m)$	m = number of edges.	[90]
Louvain	$\mathcal{O}(n \log n)$	n = number of nodes	[26]
Leading-Eigenvector	$\mathcal{O}(m(m + n))$	n= number of nodes	[86]
DSLPA	$\mathcal{O}(m^2)$	m = number of edges	[87]
NSA	$\mathcal{O}(n \log n)$	n - number of nodes	[91]
SimCMR	$\mathcal{O}(n.k.F + k^2.G + k.H)$	n= number of nodes, k= number of candidates to be shifted, F= computation time of stage 2, G= computation time of stage 3, H= computation time to calculate Modularity in stage 4.	[92]
Synwalk	$\mathcal{O}(n^2)$	n= number of nodes.	[84]
FPPM	$\mathcal{O}(m)$	m= number of edges	[93]
NBCD	$\mathcal{O}((k.n \log n))$	n= number of nodes, k= number of rounds in the shifting phase	[94]

Comparison is made on eight datasets with ground-truth communities. Also made a comparison on artificial networks, i.e. LFR networks created using the Networkx package(python) [87]. Table 4.2 describes the real-world networks.

- *The Zachary's Karate Club Network* shows the friendship among the karate club's members, where a node represents a member, and an edge between two nodes represents the friendship between those members. Zachary observed the club for three years and witnessed the split in the club into two groups due to the dispute between instructor and administrator; these two groups represent the communities in the network [84].
- *The Lusseau's Dolphin Social Network* represents the network of the doubtful sound bottlenose dolphins, where a node represents a dolphin, and an edge represents the co-occurrence of dolphins. The network consists of four ground-truth communities [93].
- *The Risk Network* is the network of a political strategical game where 42 territories(nodes) are divided into 6 continents(ground-truth communities) [92].
- *American College Football Network;* represents the games played between different college teams, where college teams represent the node, and an edge represents the game played between two teams. All the teams are divided into conferences(ground-truth communities) where each conference can consist of 8 to 12 teams [94].
- *The Yeast Network* represents the protein-protein interaction in budding yeast, where nodes represent the proteins and edges represents the interactions between proteins [95].
- *The Amazon Network* is an e-commerce network, where nodes are products and the edges between two product represents that the two products are bought frequently [96].
- *DBLP Network* represents a bibliography network, where authors are considered as nodes, and an edge between two authors represents the co-authorship [96].
- *Youtube Network* represents a social network where each user of Youtube is considered as a node, and an edge represents the friendship between two users [97].

For the datasets; DBLP, Youtube and Amazon, the top 5000 ground-truth communities are initially taken, and then the bottom one-fourth communities are removed according to their internal density. For the remaining communities, maximum independent disjoint sets are found for the ground-truth communities in order to get the disjoint communities. The communities are first arranged in descending order according to the number of nodes present in them. Then the disjoint sets are taken, starting from the first community (a community consisting of the highest number of nodes) [98, 99].

Table 4.2: Real-world networks with ground-truth communities.

DataSet	Nodes	Edges	Avg_Degree	Ground_Communities
Amazon	6428	16223	5.05	893
DBLP	23654	75618	6.3937	2974
Dolphin	62	159	5.129	4
Football	115	613	10.66	12
Karate	34	78	4.529	2
Risk	42	83	3.952	6
Yeast	2284	6646	5.8196	13
Youtube	10364	18717	3.6119	2798

4.2.4.1 Lancichinetti–Fortunato–Radicchi(LFR) Networks

Table 4.3 shows the parameters used for creating the LFR networks [100]. Other than parameters shown in Table 4.2 there is one more important parameter, the mixing parameter (μ), the fraction of intra-community edges added to each node. The smaller the value of μ is, the clearer the community structure will be. Hence, $\mu = 0.5$ can be taken as a transition point [101]. The μ value above 0.5 can lead to overlapping communities in the network. For the experiments, the μ value is varied from 0.1 to 0.5 with an increment of 0.1 for each group of LFR networks. Total 100 networks are generated for each value of μ while keeping the other parameters the same, and the results are the average of 100 networks.

Table 4.3: The parameters for LFR network construction where K_{avg} = Average degree, K_{max} = Maximum degree, C_{min} = Minimum community size, C_{max} = Maximum community size and tau1 and tau2 are the parameters for power law distribution.

Name	Nodes	K_{avg}	K_{max}	C_{min}	C_{max}	tau1	tau2
LFR1000.a	1000	4	15	5	20	2	1.1
LFR5000.a	5000	4	15	5	20	2	1.1
LFR1000.b	1000	4	20	10	40	2	1.1
LFR5000.b	5000	4	20	10	40	2	1.1

4.2.4.2 Evaluation Measures

Four popular measures, NMI (Normalized Mutual Information) [102], F-score, ARI (Adjusted Rand Index), AMI(Adjusted Mutual Information) and Modularity are used to evaluate the performance of the proposed method and the other state-of-the-art community discovery algorithms and also the recent algorithms available in the literature.

- *Normalized Mutual Information(NMI)*: Let C and $C^{'}$ be the ground-truth communities and predicted communities, respectively, $|C|$ and $|C^{'}|$ be the number of communities in C and $C^{'}$ respectively, then the NMI between C and $C^{'}$ can be calculated as follows:

$$I(C, C^{'}) = \sum_{k=1}^{|C|} \sum_{l=1}^{|C^{'}|} \frac{n_{kl}}{N} \log \left(\frac{N n_{kl}}{n_k m_l} \right)$$

$$E(C) = -\sum_{k=1}^{|C|} \frac{n_k}{N} \log \left(\frac{n_k}{N} \right)$$

$$NMI(C, C^{'}) = \frac{I(C, C^{'})}{\sqrt{E(C).E(C^{'})}}$$

  where $I(C, C^{'})$ is the mutual information between C and $C^{'}$, $E(C)$ and $E(C^{'})$ are the entropies of C and $C^{'}$, N is the total number of nodes in the network, n_k is the number of nodes belonging to the k^{th} community (C_k) of C, m_l is the number of nodes belonging to the l^{th} community ($C^{'}_l$) of $C^{'}$ and n_{kl} is the number of nodes belonging to both ground-truth (C_k) and predicted communities ($C^{'}_l$), for all $k = 1$ to $|C|$ and $l = 1$ to $|C^{'}|$ [98]. The desirable NMI score ranges from 0 to 1 [103].

- *F-Score*: The F-score of the system is defined as the weighted harmonic mean of its Precision(P) and Recall(R), that is

$$F = \frac{1}{\beta \frac{1}{P} + (1-\beta)\frac{1}{R}} \quad or \quad F = \frac{2 \times R \times P}{R + P}$$

where the weight $\beta \in [0, 1]$ [104].

- *Adjusted Rand Index(ARI)*: In 1985, Hubert and Arabie proposed the ARI for the comparison of partitions [105]. If $C = \{C_1, C_2 ... C_{\mathrm{i}}\}$ and $C' = \{C'_1, C'_2 ... C'_{\mathrm{i}}\}$ are the ground-truth and predicted communities respectively then the ARI between C and C' is calculated as:

$$ARI(C, C') = \frac{2\left(N_{00}N_{11} - N_{01}N_{10}\right)}{\left(N_{00} + N_{01}\right)\left(N_{01} + N_{11}\right) + \left(N_{00} + N_{10}\right)\left(N_{10} + N_{11}\right)}$$

where N_{00} indicates the number of node-pairs in the same community in C and C'. N_{11} shows the number of node-pairs that are in the same community in both C and C'. N_{01} represents the number of node-pairs that are in the same community in C but are not in the same community in C'. N_{10} indicates the number of node-pairs that are not in the same community in C but are in the same community in C' [88].

- *Adjusted Mutual Information(AMI)*: The AMI score was proposed by Vinh et al. in 2010 [106]. If $C = \{C_1, C_2 ... C_{\mathrm{i}}\}$ and $C' = \{C'_1, C'_2 ... C'_{\mathrm{i}}\}$ are the ground-truth and predicted communities respectively then the AMI between C and C' is calculated as:

$$AMI(C, C') = \frac{I\left(C, C'\right) - E\left(\left(C, C'\right)\right)}{1/2H\left(C\right) + H\left(C'\right) - E\left(I\left(C, C'\right)\right)}$$

where $I\left(C, C'\right)$ is the mutual information between C and C'. $E\left(I\left(C, C'\right)\right)$ is expected mutual information. $H\left(C\right)$ and $H\left(C'\right)$ are Entropy of C and C' [107].

- *Modularity*: Modularity(Q) is a popularly used performance measure to check the quality of community structure. The Modularity, Q, of a community structure can be calculated as:

$$Q = \frac{1}{2m} \sum_{c \in C} \left(2m_c - \frac{K_c^2}{2m} \right)$$

where m_c is the number of edges inside the community c, K_c is the sum of degrees of nodes in community c, and m is the total number of edges in the network [27].

Python implementation of NMI, F-score(weighted average), ARI, AMI and Modularity are used for the evaluation using the *sklearn package* [108].

4.2.5 Results

Table 4.4, 4.5, 4.6, 4.7 and 4.8 show the NMI, F-score, ARI score, AMI and Modularity score obtained by the NBCD algorithm and the other competitive algorithms for all the datasets. The scores in Table 4.4, 4.5, 4.6, 4.7 and 4.8, marked in bold font, indicate the best scores in the respective rows. The NBCD algorithm gives a better NMI, ARI and AMI score for 5 out of 8 datasets, and F-score in 4 out of 8 datasets. Although NBCD doesn't give top score for 3 datasets but still gives the scores in top 4 slots for rest of the three datasets. Table 4.9 shows the time taken by different algorithms for different datasets. Table 4.10 shows the parameters considered for NBCD, NSA and DSLPA algorithms. Figure 4.4 and 4.5 show the plots of performance of NBCD and other competitive algorithm on LFR networks. Figure 4.7 shows the plot of performance of NBCD for different values of similarity parameter, α, on considered datasets.

Table 4.4: The NMI score of different algorithms on real-world networks. The largest NMI scores are in bold. ($*$: *JAVA:Out of Memory Error*)

DataSet	Walktrap	Greedy	LPA	Louvain	Eiginvector	DSLPA	NSA	SimCMR	Synwalk	FPPM	NBCD
Amazon	0.9975	0.9975	0.9964	0.9975	0.9975	0.9975	0.9975	0.9975	0.9975	0.9927	**0.9981**
DBLP	0.9790	0.9210	0.9930	0.9200	0.7510	$*$	0.9314	0.9224	0.9200	0.9389	**0.9957**
Dolphin	0.6320	0.7030	0.7460	0.8370	0.6350	0.6297	0.8185	0.8312	0.6530	0.8495	**0.9053**
Football	0.8874	0.7436	0.8547	0.8561	0.6987	0.8873	0.8882	0.8967	**0.9242**	0.8561	0.9095
Karate	0.5042	0.6925	0.7210	0.5866	0.6771	**1.0000**	0.6995	0.6955	0.4361	**1.0000**	**1.0000**
Risk	0.8480	0.8940	0.9030	**0.9450**	0.7230	0.6542	0.8483	0.8406	0.8708	0.8483	0.8971
Yeast	0.2550	0.1270	0.2260	0.1300	0.0440	0.2017	0.0749	0.1110	**0.2656**	0.1890	0.2176
Youtube	**0.9440**	0.8580	0.9320	0.8480	0.6820	0.9280	0.7739	0.8360	0.8294	0.8905	0.9210

Table 4.5: The F-score of different algorithms on real-world networks. The largest F-scores are in bold.($*$: *JAVA:Out of Memory Error*)

DataSet	Walktrap	Greedy	LPA	Louvain	Eigenvector	DSLPA	NSA	SimCMR	Synwalk	FPPM	NBCD
Amazon	0.9609	0.9609	0.9806	0.9609	0.9609	0.9609	0.9609	0.9609	0.9609	0.9637	**0.9897**
DBLP	0.8470	0.4660	0.9420	0.4590	0.4540	$*$	0.4738	0.4690	0.4589	0.5676	**0.9748**
Dolphin	0.7400	0.7700	0.7460	0.8370	0.7630	0.7211	0.8396	0.9191	0.5851	0.9353	**0.9673**
Football	0.7791	0.3915	0.8950	0.5551	0.4710	0.7773	0.6846	0.7392	**0.9116**	0.5551	0.7878
Karate	0.6918	0.8155	0.9185	0.7714	0.7624	**1.0000**	0.8840	0.8789	0.6345	**1.0000**	**1.0000**
Risk	0.8470	0.9310	**0.9740**	0.9050	0.7600	0.4456	0.8468	0.8316	0.8560	0.8468	0.9058
Yeast	0.1170	0.1290	0.0020	**0.1630**	0.1130	0.0882	0.0855	0.1010	0.1431	0.0842	0.1146
Youtube	**0.7910**	0.4820	0.7680	0.4340	0.3910	0.4970	0.3915	0.3920	0.4075	0.5445	0.6715

Table 4.6: Adjusted Rand Index(ARI) of different algorithms on real-world networks. The largest ARI are in bold.(∗ : *JAVA:Out of Memory Error*)

DataSet	Walktrap	Greedy	LPA	Louvain	Eigenvector	DSLPA	NSA	SimCMR	Synwalk	FPPM	NBCD
Amazon	0.9727	0.9727	0.9756	0.9727	0.9727	0.9727	0.9727	0.9727	0.9727	0.9466	**0.9877**
DBLP	0.3170	0.2480	0.9360	0.2620	0.0030	∗	0.3322	0.2656	0.2567	0.3367	**0.9656**
Dolphin	0.5750	0.6550	0.6400	0.7910	0.4960	0.5892	0.7427	0.8034	0.4747	0.8207	**0.9090**
Football	0.8154	0.4737	0.6205	0.7071	0.4641	0.8132	0.7975	0.8272	**0.8967**	0.7071	0.8465
Karate	0.3331	0.6803	0.7531	0.4619	0.5121	**1.0000**	0.7022	0.6656	0.2668	**1.0000**	**1.0000**
Risk	0.6880	0.8340	**0.8680**	0.8390	0.5500	0.5017	0.6880	0.6386	0.7377	0.6880	0.7552
Yeast	0.0170	0.0330	0.0100	**0.3290**	-0.0090	0.0120	0.0276	0.0290	0.0214	0.0362	0.0207
Youtube	0.0650	0.0290	0.0370	0.0360	0.0030	0.3170	0.1120	**0.0660**	0.0260	0.0265	0.0431

Table 4.7: Adjusted Mutual Index(AMI) of different algorithms on real-world networks. The largest AMI are in bold.(∗ : *JAVA:Out of Memory Error*)

DataSet	Walktrap	Greedy	LPA	Louvain	Eigenvector	DSLPA	NSA	SimCMR	Synwalk	FPPM	NBCD
Amazon	0.9922	0.9922	0.9882	0.9922	0.9922	0.9922	0.9922	0.9922	0.9922	0.9759	**0.9940**
DBLP	0.9270	0.7830	0.9730	0.7810	0.4910	∗	0.8051	0.7859	0.7812	0.8165	**0.9836**
Dolphin	0.6070	0.6820	0.7190	0.8220	0.6040	0.5895	0.8030	0.8200	0.6464	0.8394	**0.8988**
Football	0.8561	0.7028	0.8191	0.8205	0.6332	0.8507	0.8574	0.8681	**0.8992**	0.8205	0.8820
Karate	0.4727	0.6808	0.7083	0.5653	0.6610	**1.0000**	0.6874	0.6840	0.3264	**1.0000**	**1.0000**
Risk	0.8000	0.8660	0.8730	**0.9280**	0.6510	0.5951	0.8003	0.7976	0.8299	0.8003	0.8696
Yeast	**0.1330**	0.0860	0.1200	0.0940	0.0170	0.0994	0.0510	0.0780	0.1338	0.1167	0.1123
Youtube	**0.7560**	0.5840	0.7180	0.5730	0.3380	0.7060	0.4749	0.5500	0.5442	0.6325	0.6931

Table 4.8: Modularity of different algorithms on benchmark datasets. The largest Modularity are in bold.(∗ : *JAVA:Out of Memory Error*).

DataSet	Walktrap	Greedy	LPA	Louvain	Eigenvector	DSLPA	NSA	SimCMR	Synwalk	FPPM	NBCD
Amazon	**0.998**	0.986	**0.998**	**0.998**	**0.998**	**0.998**	**0.998**	**0.998**	**0.998**	0.959	0.991
DBLP	0.970	0.959	0.992	**0.993**	0.768	0.992	∗	0.992	**0.993**	0.975	0.962
DOLPHIN	0.489	0.499	0.495	0.519	0.491	0.502	0.439	**0.526**	0.373	0.526	0.513
FOOTBALL	0.603	0.552	0.568	**0.604**	0.493	0.603	0.565	0.603	0.601	**0.604**	0.601
KARATE	0.353	0.355	0.381	**0.419**	0.393	0.402	0.371	0.395	0.284	0.371	0.371
RISK	0.624	0.606	0.625	**0.634**	0.547	0.624	0.502	0.609	0.631	0.624	0.591
YEAST	0.524	0.335	0.570	**0.587**	0.250	0.528	0.430	0.567	0.522	0.541	0.503
YOUTUBE	0.696	0.672	0.756	**0.779**	0.598	0.700	0.592	0.735	0.752	0.710	0.696

Table 4.9: Time(in Seconds) taken by the algorithms considerd for different datasets.(∗ : *JAVA:Out of Memory Error*)

DataSet	Walktrap	Greedy	LPA	Louvain	Eiginvector	DSLPA	NSA	SimCMR	Synwalk	FPPM	NBCD, α=2
Amazon	0.421	1.880	1.823	0.727	0.678	6.270	4.123	2.697	**0.085**	3.53	1.095
DBLP	4.573	9.820	20.091	3.661	1.689	∗	120.237	20.144	**0.357**	1588.636	5.512
Dolphin	**0.002**	0.020	0.011	0.007	0.021	0.034	**0.002**	0.006	0.039	0.14	0.114
Football	**0.007**	0.083	0.015	0.014	0.030	0.058	0.009	0.010	0.0122	0.312	0.108
Karate	0.002	0.010	0.007	0.004	0.014	0.009	**0.001**	0.004	0.035	0.156	0.029
Risk	0.115	0.009	0.005	0.035	0.153	0.018	**0.001**	0.002	0.0317	0.173	0.083
Yeast	0.510	8.241	0.968	0.576	0.216	13.340	2.402	0.226	**0.157**	26.404	8.236
Youtube	1.873	30.107	7.164	2.341	1.608	95.929	43.109	11.295	**0.356**	267.968	10.369

Table 4.10: Parameters used for different Algorithms.($*$: *JAVA:Out of Memory Error*)

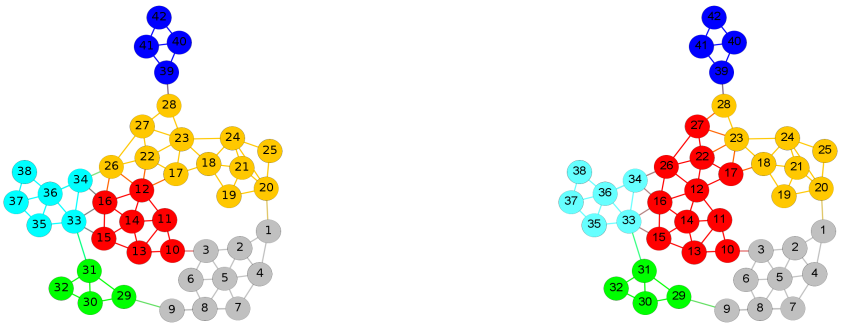
DataSet	NSA	DSLPA	NBCD
Amazon	0.10	6.00	2.5
DBLP	0.05	$*$	1.5
Dolphin	0.10	2.00	2.0
Football	0.05	1.50	3.0
Karate	0.10	2.00	2.0
Risk	0.05	1.50	1.5
Yeast	0.05	1.50	3.0
Youtube	0.05	0.83	3.0

Figure 4.3: Karate Network; (a) Ground-truth communities, (b) Communities detected by NBCD. Dolphin network; (c) Ground-truth communities, (d) Communities detected by NBCD. Risk network; (e) Ground-truth communities, (f) Communities detected by NBCD. Football network; (g) Ground-truth communities, (h) Communities detected by NBCD.

(a) **(b)**

(c) **(d)**

(e) **(f)**



(g) **(h)**

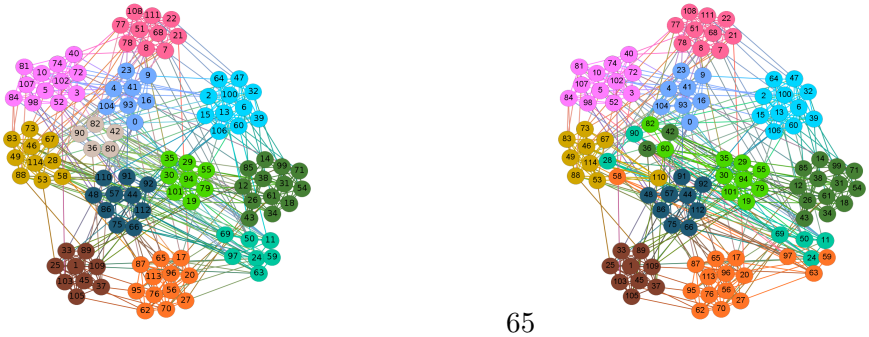
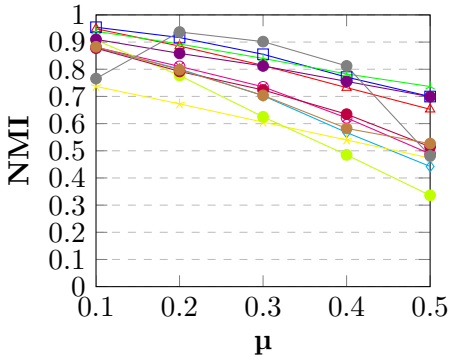
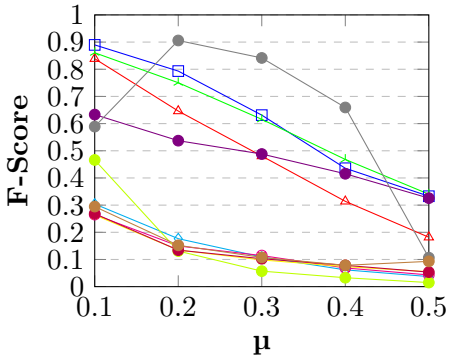


Figure 4.4: Comparison of different evaluation measure of considered community discovery algorithm on LFR-1000 networks. NBCD, $\alpha = 0.2$ (), Walktrap (), Greedy Modularity (), LPA (), Louvain (), Eigenvector (), SimCmr (), NSA (), DSLPA (), Synwalk (), FPPM ().

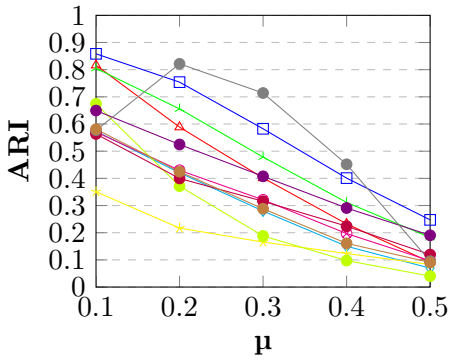
LFR 1000.a [K_min=4, K_max=15, C_min=5, C_max=20]



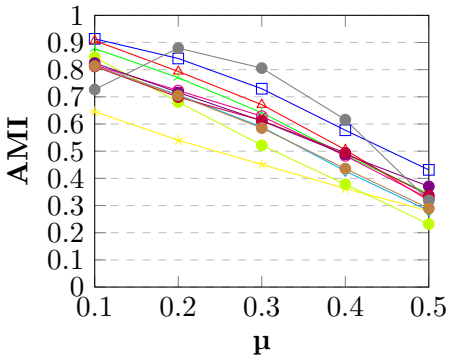
LFR 1000.a [K_min=4, K_max=15, C_min=5, C_max=20]



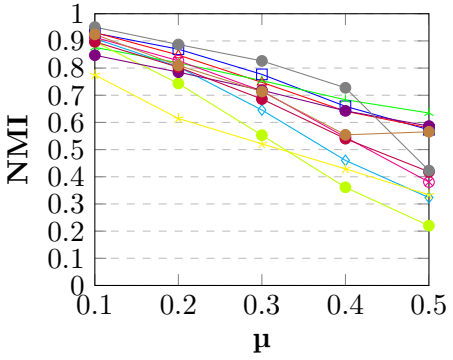
LFR 1000.a [K_min=4, K_max=15, C_min=5, C_max=20]



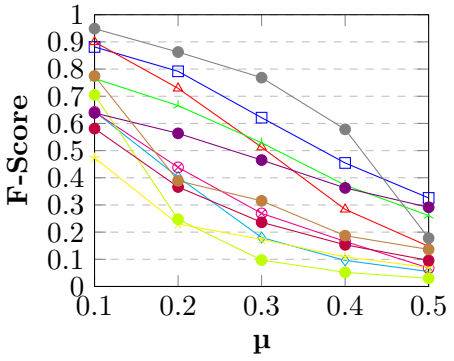
LFR 1000.a [K_min=4, K_max=15, C_min=5, C_max=20]



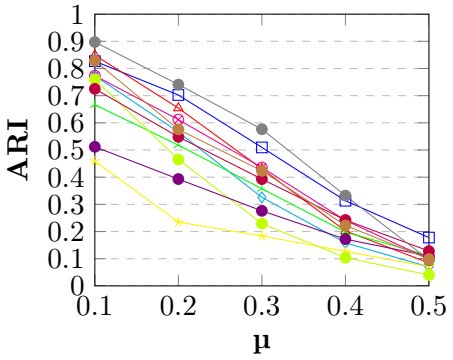
LFR 1000.b [K_min=4, K_max=20, C_min=10, C_max=40]



LFR 1000.b [K_min=4, K_max=20, C_min=10, C_max=40]



LFR 1000.b [K_min=4, K_max=20, C_min=10, C_max=40]



LFR 1000.b [K_min=4, K_max=20, C_min=10, C_max=40]

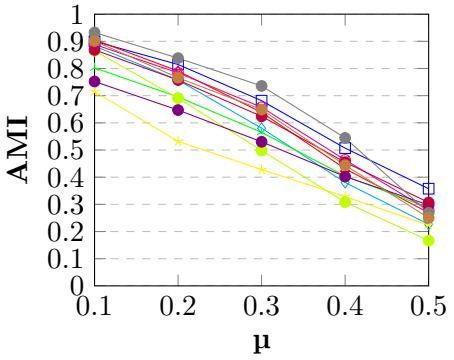
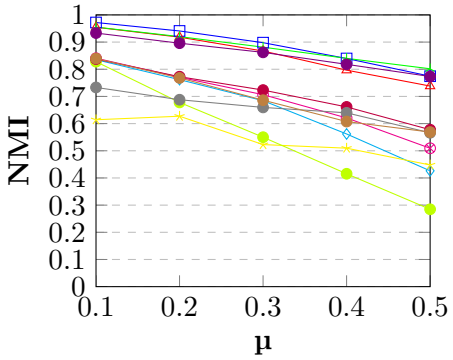
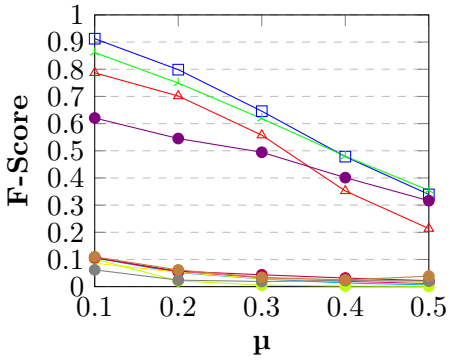


Figure 4.5: Comparison of different evaluation measures of considered community discovery algorithm on LFR-5000 networks. NBCD, $\alpha = 0.2$ (), Walktrap (), Greedy Modularity (), LPA (), Louvain (), Eigenvector (), SimCmr (), NSA (), DSLPA (), Synwalk (), FPPM ().

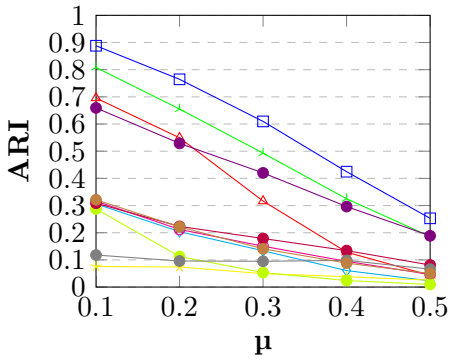
LFR 5000.a [K_min=4, K_max=15, C_min=5, C_max=20]



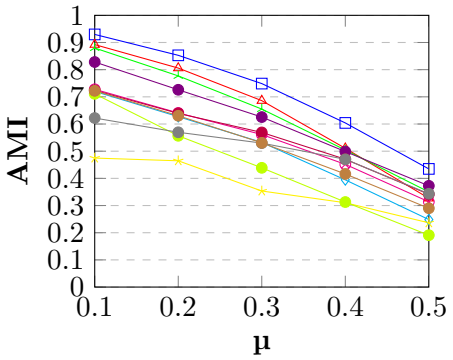
LFR 5000.a [K_min=4, K_max=15, C_min=5, C_max=20]



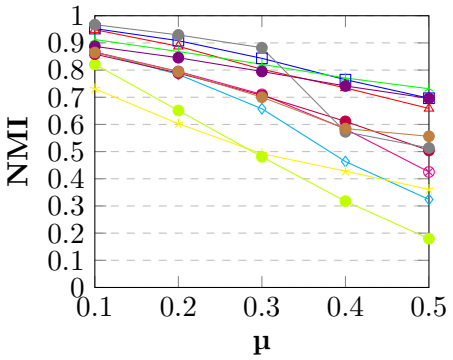
LFR 5000.a [K_min=4, K_max=15, C_min=5, C_max=20]



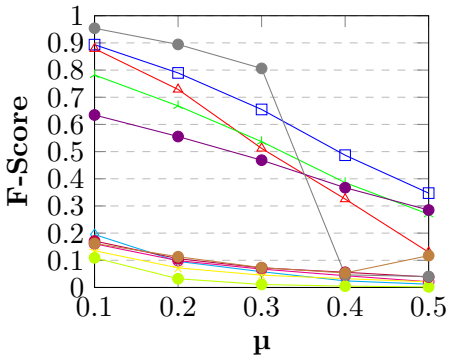
LFR 5000.a [K_min=4, K_max=15, C_min=5, C_max=20]



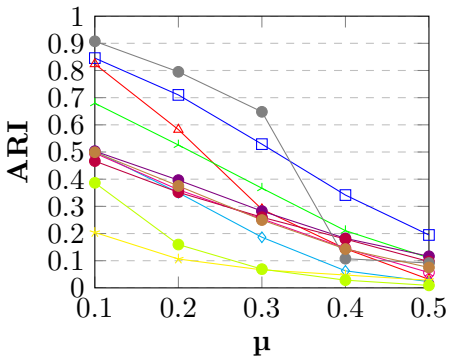
LFR 5000.b [K_min=4, K_max=20, C_min=10, C_max=40]



LFR 5000.b [K_min=4, K_max=20, C_min=10, C_max=40]



LFR 5000.b [K_min=4, K_max=20, C_min=10, C_max=40]



LFR 5000.b [K_min=4, K_max=20, C_min=10, C_max=40]

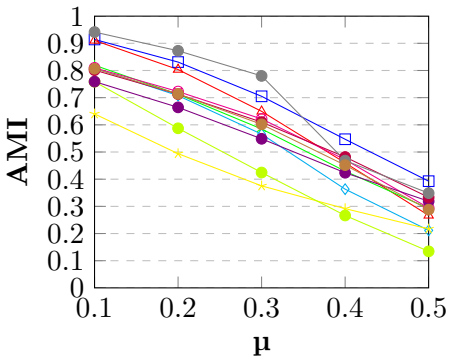
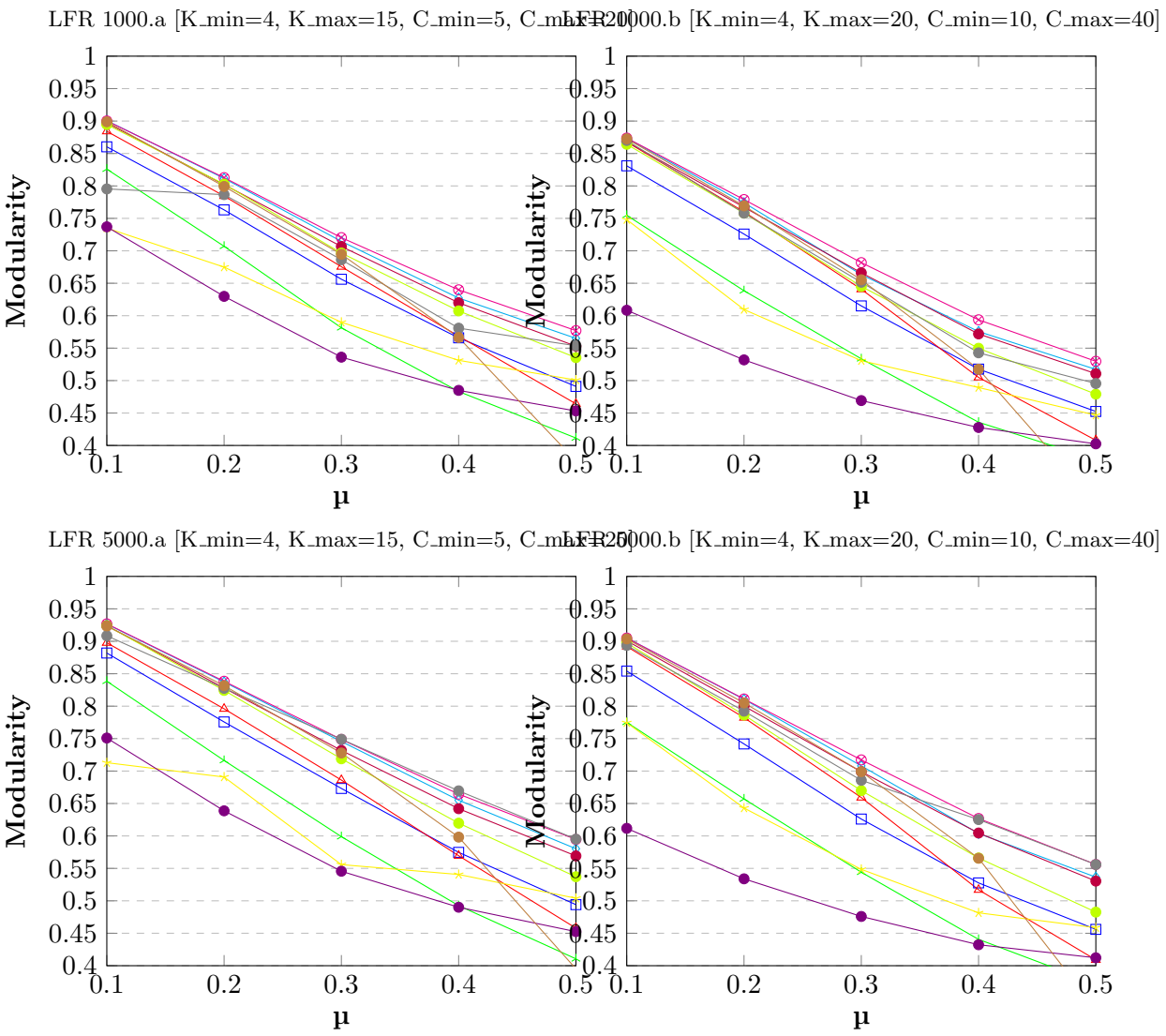


Figure 4.6: Comparison of Modularity measure of considered community discovery algorithm on LFR-1000 networks. NBCD, $\alpha = 0.2$ (), Walktrap (), Greedy Modularity (), LPA (), Louvain (), Eigenvector (), SimCmr (), NSA (), DSLPA (), Synwalk (), FPPM ().



4.3 Discussion

4.3.1 LFR Networks

Figures 4.4 and 4.5 show the plots for NMI, F-score, ARI and AMI scores and Figure 4.6 shows the Modularity score for NBCD and other algorithms considered. For in-

stance, if we don't consider Synwalk for comparison, NBCD outperforms all the other algorithms considered. NBCD gave a better score for all performance measures except Modularity. However, even though the Modularity score for NBCD is not the best, it is still very close to the best one. Furthermore, unlike Louvain, Greedy Modularity and SimCMR, which achieve the best Modularity, NBCD gives the best scores for all the other performance measures. Louvain, Greedy Modularity and SimCMR look for the Modularity optimization and hence achieve the best Modularity score compared to other algorithms considered, but fail to detect small communities, which results in the low F-score compared to the best F-score. On the other hand, even though the Modularity score of NBCD is not the best, it is still very close to the best one and yet achieves the best scores in all other performance measures. Now, if we compare NBCD and Synwalk, Synwalk shows inconsistent performance for different LFR networks. For some LFR networks, Synwalk performs very poorly (LFR1000.a (μ=0.1,0.5), LFR5000.a (for all μ values), LFR5000.b (μ=0.4, 0.5)). On the other hand, NBCD either gave the best score or else a close one to the best score; NBCD didn't perform comparatively poorly for any LFR networks for all the measures considered. Only for the F-score that too for LFR1000.a networks with μ values 0.3 and 0.4, there is a noticeable difference in the F-scores of NBCD and Synwalk. From the plots, it can be clearly seen that Synwalk shows inconsistency in its performance with regard to performance measures considered, while NBCD performs consistently for different LFR networks

4.3.1.1 Real-time Datasets

In the Karate network (Fig.4.3(a & b)), it can be seen that the NBCD algorithm detected ground-truth communities correctly. In the Dolphin network (Fig. 4.3(c & d)), the NBCD algorithm predicted the correct community for each node except for the nodes '28' and '39'. While in the Football network (Fig. 4.3(e & f)), the NBCD algorithm predicted most of the communities same as the ground-truth communities. In the Risk network (Fig. 4.3(g & h)), the proposed algorithm NBCD predicted the correct community for each node except for the nodes '17', '22', '26' and '27'. Although the NBCD doesn't give the best scores for the Youtube and Yeast networks, it still gives a better score than the other 6 algorithms. As there are only 13 ground-truth communities in the Yeast network and communities detected by NBCD are 181, the F-score is very low.

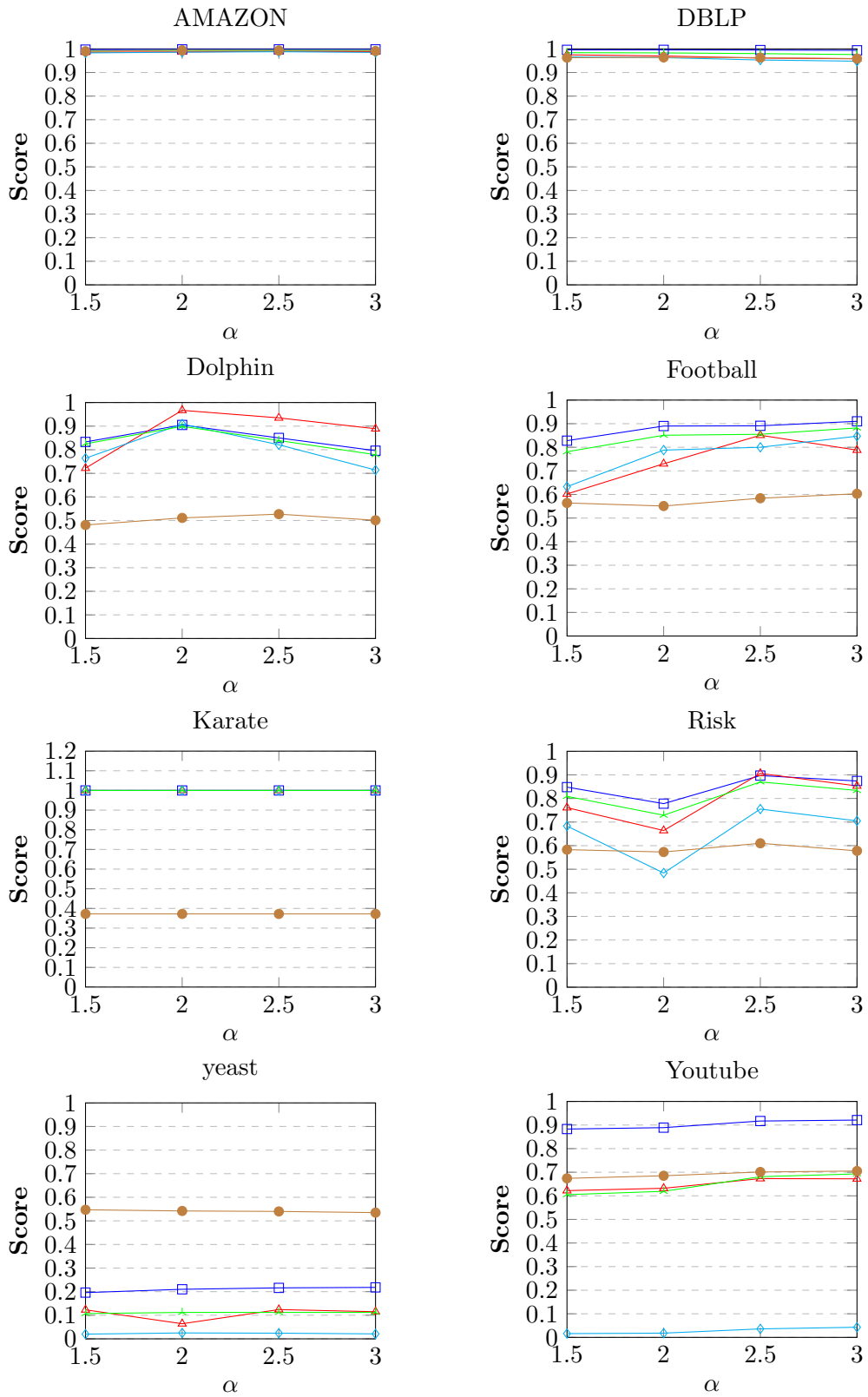
4.3.2 Effect of α

Figure 4.7 shows the plot between performance measures and α values. Table 4.11 shows the time taken by NBCD for different value of α. It can be concluded from Table 4.11 that for a higher value of alpha, NBCD takes less time, i.e. for the communities with tightly connected nodes, NBCD takes more time comparatively than the communities with loosely connected nodes. α=1.5 represents the communities with tightly connected nodes and α=3 represents those with loosely connected nodes. Through the experiments(Fig. 4.7), 2.5 is the recommended value of α in case where ground truth communities are not available.

Table 4.11: Time taken by NBCD for different α values.

DataSet	**α=1.5**	**α=2**	**α=2.5**	**α=3**
Amazon	1.829	1.095	0.996	0.976
DBLP	8.524	5.512	4.503	4.285
Dolphin	0.1976	0.0903	0.0609	0.0481
Football	0.4775	0.2032	0.0642	0.0634
Karate	0.039	0.029	0.011	0.011
Risk	0.1009	0.0595	0.0341	0.0225
Yeast	11.913	8.236	5.646	4.844
Youtube	14.452	10.369	5.49	4.57

Figure 4.7: Performance of NBCD for different α values on datasets considered. NMI ( ), F-score ( ), ARI ( ), AMI ( ), Modularity ( )



4.3.3 Comparison with the state-of-the-art-algorithms

From Table 4.4-4.7, it can be concluded that the proposed algorithm NBCD performs better than the state-of-the-art community discovery algorithms. NBCD gives better scores for 5 out of 8 real-life datasets. NBCD also provides a consistent and the best results for LFR networks. From Figure 4.4 and 4.5, it can be observed that the Louvain and Greedy Modularity perform best according to Modularity but if other performance measures are considered they performed badly. Eigenvector algorithms performed worst according to all the performance measures. In LFR networks, the number of communities detected by Louvain, Greedy Modularity, and Eigenvector algorithms is less compared to the ground-truth communities, resulting in low scores. In contrast, the number of communities detected by NBCD is close to the number of ground-truth communities, the F-score plot for LFR networks is a proof of that. The time taken by Walktrap, LPA, Louvain and LPA are less than NBCD. The Greedy Modularity algorithm takes less time for some datasets and more time for some datasets than NBCD. NBCD gives the best score or gives a score close to the best one with respect to all the performance measures. No other algorithm has shown this consistency.

4.3.4 Comparison with the recently published algorithms

The NBCD algorithm completely outperforms the recently published algorithms, i.e. Synwalk, FPPM, DSLPA, NSA and SimCMR. It can be seen in Table 4.4-4.7 that NBCD gives the best score(NMI, F-score, ARI and AMI) for almost every dataset. However, SimCMR an FPPM achieves the best scores when Modularity is considered, but there is a negligible difference between the Modularity score of NBCD and the best Modularity score. In Football dataset Synwalk gives the highest scores but NBCD also achieves the a scores which is very close to the highest ones. For the yeast dataset, NSA, FPPM and SimCMR give a better ARI score than NBCD. In Youtube, dataset DSLPA gives better NMI, ARI and AMI scores than NBCD. In the LFR networks, also NBCD completely outperforms the other algorithms except Synwalk on every score. The reason behind the poor performance of all three algorithms is that they are unable to detect the small communities. The number of communities detected by them is very less compared to ground-truth communities. The other important factor is parameters used by DSLPA and NSA. For the DSLPA, the author gives no specific range for

the parameter, ϵ. So the user needs to try different values of ϵ to achieve the best communities structure, which will take multiple runs to achieve the best community structure. Although in NSA, authors gave the range of threshold value(δ) of merging parameter(λ), which is 0.05 to 2, problem here is that one needs to increment/decrement the threshold value by 0.01 within the range to achieve the best community structure, which will take multiple attempts again. It can be observed from Table 4.9 that for the Youtube dataset, NSA takes 43 seconds and DSLPA takes around 96 seconds which is comparatively high. Similarly, for the DBLP dataset, NSA took about 2 minutes, and DSLPA doesn't run due to JAVA: out of memory error. For the yeast dataset, NBCD takes higher time in comparison to the other three algorithms.

4.4 Discovery of Communities Using NBCD in Alzheimer's Disease Dataset

The ground truth sets of communities are not known apriori in biological datasets. Hence, the NBCD provides an approximate method, using which communities can be discovered in these sets. For the community detection we used the NBCD with the suggested value of α i.e. $\alpha = 2.5$ 4.3.2.

4.4.1 Dataset

For the temporal analysis of Alzheimer's disease we used the gene-expression data, GSE28146[1]. The dataset is freely available at Gene Expression Omnibus(GEO). This dataset contains the gene-expressions from gray matter of hippocampus tissue. The dataset consists of 8 samples of control, 7 samples of incipient, 8 samples of moderate and 7 samples of severe stage of AD.

4.4.2 Network Construction and Community Detection

For the network construction we have used the tcGONet framework, proposed in the Chapter 3 [4]. Using the tcGONet framework , networks are constructed for each stage of AD i.e. control, incipient, moderate and severe stage. The community detection algorithm NBCD is applied on each network to detect the communities. Table 4.12

[1]https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE28146

shows the description of networks constructed and number of communities detected in each network.

Table 4.12: Network Description.

Stage	Nodes(Genes)	Edges	Number of communities using NBCD
Control	2435	11842	59
Early	2580	15076	49
Moderate	2343	8528	89
Severe	2594	15332	34

4.5 Analysis

For analysis, common genes between two consecutive stages are taken and then their neighbourhoods i.e communities are analyzed. Table 4.13 shows the details of common genes between two consecutive stages.

Table 4.13: Common Genes between two consecutive stages.

Stages	Common Genes
Control and Early	2251
Early and Moderate	2213
Moderate and Severe	2207

Changes in the communities of all common genes between in two consecutive stages are analyzed. For example, analyzing size of communities of gene X in stage 1 and stage 2 and common genes in communities of gene x in stages 1 and 2. For the analysis we picked top 20 genes according to Jaccard distance between communities of genes between the consecutive stages. Let, $C_{X_{ctrl}}$ and $C_{X_{early}}$ be the communities of gene X in control and early network respectively. Then, the Jaccard distance between communities of Gene X in control and early network can be calculated as:

$$JD(X_{ctrl\_vs\_early}) = \frac{|C_{X_{ctrl}} \cap C_{X_{early}}|}{|C_{X_{ctrl}} \cup C_{X_{early}}|}$$

Table 4.14: Community change bwtween the networs.

Gene	Size of community in control network	Size of community in incipient network	Number of Common genes	Jaccard Distance
OR2C3	847	3	0	0.9988
ADGRG1	25	825	0	0.9988
RALGAPA2	847	3	0	0.9988
SPIRE1	847	2	0	0.9988
LRG1	7	825	0	0.9988

Table 4.15: Top 20 genes whose communities are disturbed. Genes marked in red colour are involved in the AD and genes in orange colour are involved in some other neurological disorder.

Control to early	Early to Moderate	Moderate to Severe
OR2C3	ASAP3	ATAD3A
ADGRG1	HHEX	RABEPK
RALGAPA2	NUDCD1	SLC27A1
SPIRE1	OR10D3	CACNB4
LRG1	ITGAM	TM4SF5
TFCP2L1	JAM3	HIBADH
MED23	THBS1	GP2
TPM3	BTRC	ABCB5
MMP9	COG5	KCNJ11
IPO11	ZNF33B	ADGRE3
VPS53	GABRD	ADAM12
DDX52	WDR62	FBN3
COL8A1	ZNF449	HAUS8
CACNG8	S1PR5	AXIN2
ABCC8	FAM83A	GABRD
LAMA3	ADAMTS2	OR2B2
ITGA2	OPN4	SHISA9
MRPL41	RHOT2	ZNF177
PCDHB1	GP2	ZNF559-ZNF177
PPP2R1B	ELOVL6	SAR1B

4.5.1 Control to Early

In the top 20 genes whose communities got disturbed during control to early stage of AD, many genes like **MMP9, VPS53, ABCC8, LAMA3, ITGA2** are found to be associated with AD or some other neurological disorders. Other than these genes many other genes are recently shows some association with the AD.

In [109], Pathak et al. found that the **OR2C3** gene is one of the genes that is methylated in individuals with mild cognitive impairment (MCI) in Mexican American population. Methylation of this gene is associated with an increased risk for MCI, which is a precursor to Alzheimer's disease. The OR2C3 gene is involved in the transport of molecules in synapses, which are the connections between nerve cells in the brain. This suggests that changes in the OR2C3 gene may contribute to the development of cognitive impairments and may be a potential target for therapeutic intervention.

In [110], Folts et al. found that the **ADGRG1** gene encodes for a specific type of G protein-coupled receptor (GPCR) called an adhesion GPCR, which plays a critical role in the development and function of the nervous system. The study found that the ADGRG1 gene is highly expressed in the brain and spinal cord, and that mutations in this gene are associated with several neurological disorders, including Parkinson's disease, Alzheimer's disease, and Schizophrenia. In [111], Camilli et al. investigated the role of a protein called leucine-rich alpha-2-glycoprotein 1 (**LRG1**) in various disease conditions. The study found that LRG1 plays a role in the development and progression of several diseases, including cancer, inflammation, and cardiovascular disease.

4.5.2 Early to Moderate

During the early to moderate stage, we found that **HHEX, NUDCD1, JAM3, BTRC, ZNF33B, GABRD, ADAMTS2, OPN4, and ELOVL6** were among the top 20 genes that experienced community disturbance and are associated with either Alzheimer's Disease or some other neurological disorder. Other than this many other genes have shown the association with Alzheimer's disease. In [112] Su et al. found that Alzheimer's disease is deferentially associated with the **ASAP3** low expression phenotype. Salih et al. in [113] found that variations in the ITGAM gene, which codes for the protein CD11b, can influence a person's risk for developing Alzheimer's disease. The study suggests that individuals with certain variations in the **ITGAM** gene may

be more susceptible to the toxic effects of amyloid beta, a protein that forms plaques in the brains of Alzheimer's patients. This increased susceptibility may increase the risk of developing Alzheimer's disease.

The study [114] aimed at identifying the potential biomarkers for Alzheimer's disease using a proteomic analysis of platelet membrane proteins. One of the genes identified in the study as a potential biomarker for Alzheimer's disease is **THBS1** (thrombospondin 1). The study found that THBS1 levels were significantly increased in the platelets of Alzheimer's disease patients compared to healthy controls. Additionally, the study found that THBS1 levels were positively correlated with the severity of Alzheimer's disease symptoms, as measured by cognitive testing.

4.5.3 Moderate to Severe

We found that **CACNB4, HIBADH, KCNJ11, ADAM12, FBN3, AXIN2, and SHISA9** were among the 20 genes that experienced community disruption and are associated with either Alzheimer's Disease or some other neurological disorder during the moderate to severe stage. Apart from these genes many other genes among the top 20 genes found to be associated with the AD. In [115], Zhao et al. found that the protein ATAD3A plays a significant role in the development of neuropathology and cognitive deficits in Alzheimer's disease. The study found that when ATAD3A forms oligomers (clusters of multiple protein molecules), it leads to an increase in the formation of amyloid plaques in the brain, which is a hallmark of Alzheimer's disease. Additionally, the study found that these oligomers also cause cognitive deficits in the brain, such as memory loss and a decline in cognitive function. Overall, the findings of this study suggest that ATAD3A plays a significant role in the development of Alzheimer's disease and that targeting ATAD3A may be a potential therapeutic approach for treating the disease.

4.6 Conclusions

In this chapter, a new community discovery algorithm, NBCD, which works using two novel similarity functions based on a novel similarity parameter and a set of ground rules that can handle the conflicts is proposed. For a fair comparison, NBCD is compared with community discovery algorithms with different approaches, i.e. algorithms

based on random walks, label propagation, neighbours-based similarity, optimization of performance measures and hybrid approaches. From the results, unlike other community discovery algorithms used for comparison, NBCD performs equally well on all the performance measures considered.

The NBCD algorithm is then used to identify the gene's communities in different stages of AD. Genes whose communities are being changed or disturbed drastically during the disease progression from one stage to the next are being observed. Such genes are found to be associated with AD. Interestingly, some identified genes are recently found to be associated with AD.

Our findings suggest that the community disturbance of specific genes may play a role in the progression of Alzheimer's disease. Further studies are needed to fully understand the mechanisms underlying the changes in gene expression and the impact on disease progression.

4.7 Summary

In this chapter, a novel neighbour-based community discovery algorithm, NBCD, is introduced. The NBCD is then compared with different state-of-the-art and recently published community discovery algorithms. To evaluate the performance of the NBCD different popular performance measures was used. The extensive experiments shows that the NBCD perform better and consistent than the other considered algorithms. Later in the chapter NBCD was used to analyze the community structure of genes in Alzheimer's stage-wise disease networks.

Chapter 5

Temporal Analysis of Disease Networks

Temporal graphs, also known as dynamic graphs, are a type of graph data structure that represents the evolution of relationships over time. Temporal graphs can be classified into two main categories: static temporal graphs and dynamic temporal graphs. Static temporal graphs represent a snapshot of the graph at a specific time point, while dynamic temporal graphs represent the evolution of the graph over time. Temporal graphs can be represented as $G(V, E, T)$ where V is the set of vertices, T is the set of time stamps and E, a set of temporal edges, where each temporal edge is a triplet (u, v, t), with $u, v \in V$ and $t \in T$ [116]. Temporal graphs are particularly useful for modeling dynamic systems, such as social networks, transportation systems, and biological networks.

In the context of biological networks, temporal graphs can be used to track the evolution of gene expression or protein-protein interactions over time, providing insights into the underlying mechanisms of disease progression. In [116], Hosseinzadeh et. al discussed about the temporal graphs and their applications in biology and medicine.

5.1 Literature Survey

In [117], Thompson et al. highlight the potential for temporal network theory to advance the understanding of brain disorders, such as Alzheimer's disease, schizophrenia, and depression which are characterized by abnormal functional connectivity patterns.

Authors also used the temporal degree and closeness centrality to find out the central nodes in the brain network dynamics.

In [118], Li et al. constructed a temporal network using spatial information and gene expression data and then they applied the clustering algorithm on the temporal network constructed to identify protein complexes. Similarly in [119], Meng et al. constructed temporal-spatial dynamic PPI networks by integrating protein-protein interaction networks with gene expression data and subcellular localization information. They introduced the maximum degree centrality (MDC) method and constructed temporal network to evaluate the essentiality of hub proteins.

In [120], Hiram et al. discussed the importance of temporal graph ranging from epidemic modelling and predicting the epidemic propagation, to evaluation of measures for epidemic controlling. In [121], Humphries et al. introduced a framework for modeling the spread of epidemics on temporal networks. The framework consists of three main components: a model for the dynamics of the epidemic on the network, a model for the temporal evolution of the network, and a method for inferring the parameters of the model from data.

In [122], Gao et al. developed a method called temporal network flow entropy (TNFE) to detect the critical state during the disease. The authors defined the critical state as a pre-disease state which act as a tripping point which can be helpful to prevent the disease deterioration.

In [123], Wang et al. developed a tool, MitoTNT, to analyse the dynamics of Mitochondria network in the cell that rapidly changes through fission, fusion, and motility.

5.2 Centrality Measures for Temporal Graphs

The concept of centrality measures, originally applied in static graphs, has been expanded to include temporal graphs as well. Centrality measures are a crucial aspect of graph analysis and are frequently used to determine the most significant nodes within a graph. Various metrics have been developed to define centrality, each providing a distinct perspective on the importance of a node. Below are the three popular centrality measures which are extended to temporal graph:

- Degree Centrality: In a temporal graph, degree centrality is calculated in a similar

way to how it is done in a static graph, but with an additional step of summing across different time stamps. Specifically, the temporal degree centrality of a node v in a temporal graph is determined by adding up the degree of node v in each individual snapshot [117]. The temporal degree centrality of a node v, $TD(v)$ can be calculated as follow:

$$TD(v) = \sum_{t=1}^{t=n} deg_v(t)$$

where, $deg_v(t)$ is the degree of node v in graph G at timestamp t.

- Temporal Closeness Centrality: Closeness centrality measure takes into account the distance of a node from all other nodes in a graph. It is often used to identify nodes that have quick and easy access to information or resources within a network. Temporal closeness centrality of a node v is obtained by averaging the sum of the inverses of the optimal distances between v and other nodes of the temporal graph. The temporal closeness centrality TC(v) of node v can be calculated as follow:

$$TC(v) = \frac{1}{n-1} \sum_{u \in V \setminus u} \frac{1}{d_{u,v}},$$

where n is number of vertices and $d_{u,v}$ is the optimal distance between node u and v such that $u \neq v$.

- Temporal Betweenness Centrality: It is widely used centrality measure to identify the important nodes. Betweenness centrality of a node is calculated as number of times a node appears in an optimal path between any pair of nodes. The temporal closeness centrality TB(v) of node v can be calculated as follow:

$$TB(v) = \sum_{s \neq v \neq u \in V} \frac{\delta_{s,u}(v)}{\delta_{s,u}},$$

where $\delta_{s,u}$ is the number of optimal paths between node s to u $\delta_{s,u}(v)$, is the number of times v appears in the optimal paths between s and u.

5.2.1 Issues in implementing centrality measures for temporal graph

In section 5.2, for centrality measures we have used the word "optimal path" instead of shortest path which is considered as optimal path in static networks. The reason behind using the word "optimal path" is that there could be many optimal paths in

temporal networks. In a temporal network, path with the least number of edges(shortest path) can be considered as a optimal path or may be a path with earliest arrival time(foremost) or may be a path with the minimal travel time(fastest) or may be a combination of two like shortest foremost. Hence, deciding the optimal path is a important task.

Xuan et al. in [124], presented the algorithms to calculate shortest, foremost and fastest paths. In [125], Kim et al. gave an efficient algorithm with a polynomial time complexity to calculate betweenness centrality based on the shortest path. However, for a large network calculating betweenness centrality in a temporal graph is computationally very expensive. On the other hand researchers have found that finding the foremost and fastest walk is a NP-hard problem [116, 126, 127].

5.3 Motivation

Recently, many studies have been carried out comparing different neuro-degenerative diseases. However, the stage-wise analysis of diseases has not been initiated so far. Believing that different genes can be responsible for different stages of disease progression, we, in this chapter, have introduced a modified version of betweenness centrality named *transitioncentrality* for temporal graphs.

5.4 Experiments

In this chapter we introduce a new centrality measure for temporal networks called *transitioncentrality* measure which ranks the nodes on the basis of their importance in network progression. Proposed Transition centrality measure is used on three different gene expression datasets(Alzheimer's Disease, Parkinson's Disease and Human breast cancer cell cycle) to identify the genes that are responsible for disease progression.

Here we formally describe the definition and mathematical representation of temporal graph, temporal path and shortest temporal path.

Definition (Temporal Graph). An undirected Temporal graph G can be represented as $G(V, E, T)$ where V is the set of vertices, E is the set of edges, $E \subseteq \{(\{u,\ v\},\ t) \mid u,\ v \in V,\ u \neq v,\ t \in T\}$ and T is the time stamp, $T\ \in \{1, 2, ...N\}$. For a temporal graph G, V_t is the set of vertices and E_t is the set of the edges present in

the graph at time stamp t, $t \in T,\ V_t \subseteq V$ *and* $E_t \subseteq E$.

Definition (Temporal Path). A temporal path is a path such that the time stamp of every edge in the path should be equal or greater than it's previous edge.

Definition (Shortest Temporal Path). Shortest temporal path is a shortest path such that the path should be a temporal path.

5.4.1 Transition Centrality

Transition centrality(C_T) is a measure for temporal graphs that calculates the importance of nodes during a network evolution, that is, graph evolution from a particular time stamp to the next time stamp. Transition centrality identifies the nodes through which the nodes in graph G_{t_i} at time instant t_i are connected to the newly added nodes in graph $G_{t_{i+1}}$ at next time instant t_{i+1}. In other words, it identifies the nodes through which most communication happens between the old and new nodes. The mathematical representation of transition centrality for a temporal graph $G(V, E, T),\ T = \{t_i,\ t_{i+1}\}$ is as follow:

$$C_T(v) = \sum_{a \neq b} \frac{\sigma_{a,b}(v)}{\sigma_{a,b}},\ a\ \in V_{t_i},\ b\ \in \{V_{t_{i+1}} \setminus V_{t_i}\}$$

where $\sigma_{a,b}$ is the number of temporal shortest path from a to b and $\sigma_{a,b}(v)$ is the number of temporal shortest path between a to b that passes through v.

5.4.2 Datasets

Gene expression data of Alzheimer's disease, Parkinson's disease and human breast cancer cell cycle are considered for identification of crucial nodes responsible in disease progression.

- *Alzheimer's Disease(AD)*: Alzheimer's disease a neurological disorder. The AD dataset(GSE28146[1]) can be downloaded form Gene Expression Omnibus(GEO). This dataset contains the gene-expressions from gray matter of hippocampus tissue. The dataset consists samples from control, early, moderate and severe stages of AD.

- *Parkinson's data*: Parkinson's disease is another neurological disease. The data consists patients information at different time points including demographic

[1]https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE28146

data(gender and age in year 4) and clinical data(General PD Severity, Disability, Cognition, Autonomic Function, Sleep, and Mental Health). However, for our experiments we only consider the gene-expression data of patients at different PD severity level. After excluding the patients with incomplete data we got data from 4 stages stage 0 to stage 3 where stage 0 represents the control state and stage 3 represent the severity of PD. Parkinson's disease dataset can be downloaded from Parkinson's Progression Markers Initiative[1].

- *Human Breast Cancer Cell Cycle*: The cell cycle is a highly regulated cyclic process that leads to cell division. It comprises of distinct phases through which cells proceed in a pre-defined order leading to their duplication and transmission of genetic information from one generation to the next. The phases of the cell cycle are G1, S, G2 and H.

  G1 phase (which is the gap between two divisions and the cells prepare themselves for division), S phase (cells undergo DNA synthesis), G2 phase (between S phase and M phase) and M phase (a set of ordered processes which involves the generation of mitotic spindles leading to cell division). A quiescent G0 phase also exists before G1 and after the M phase, which includes cells that have temporarily or permanently halted cell division [128]. The cell cycle is regulated through a series of checkpoints (G1-S and G2-M interface), ensuring that each stage is completed fully before proceeding to the next. This is controlled through the interactions of large groups of genes that are dynamic and ensure temporal control [128]. To understand the evolution of interactions between genes in different cell cycle phases, Human breast cancer cell cycle data were obtained from the GEO database (GEO Accession ID: GSE94479[2]).

5.4.3 Graph Construction

For the initial graph construction for every stage of AD and PD dataset, we have constructed the networks using tcGONet framework, proposed in chapter 3 [4].

To compute the transition centrality, temporal graph for each consecutive stage is constructed. For example in AD datasets, temporal graphs between the con-

[1]https://www.ppmi-info.org/

[2]https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE94479

secutive AD stages i.e from control to incipient, incipient to moderate and moderate to severe are built. Similarly, temporal graphs for PD stages are also constructed. The algorithm for the temporal graph construction is as follows:

Algorithm 3: Temporal Graph Construction

Input: Two graph instances, $G(G_{t_i}(V_{t_i},\ E_{t_i}),\ G_{t_{i+1}}(V_{t_{i+1}},\ E_{t_{i+1}}))$

Output: Temporal Graph
$G_T(V, E, T),\ V = \{V_{t_i} \cup V_{t_{i+1}}\},\ E = \{E_{t_i} \cup E_{t_{i+1}}\},\ T = \{1,\ 2\}$

1 **for** *edge in* E_{t_i} **do**
2 Add edge in G_T with T=1
3 **for** *edge in* $\{E_{t_{i+1}} \setminus E_{t_i}\}$ **do**
4 Add edge in G_T with T=2

Using the framework [4] we constructed the network for all stages of each dataset. Table 5.1 shows the description of static networks of the disease constructed for all stages. Then using the proposed algorithm we constructed the temporal network for each consecutive stage. Table 5.2 describes the temporal networks constructed.

Table 5.1: Disease stage wise network description

Disease	Network	Nodes	Edges	Avg. Degree
AD	Control	2435	11842	9.73
	Incipient	2580	15076	11.69
	Moderate	2343	8528	7.28
	Severe	2594	15332	11.82
PD	Stage 0	1548	14736	19.04
	Stage 1	1985	45890	46.24
	Stage 2	2049	71447	69.74
	Stage 3	2045	62962	61.58
Cancer Cell-cycle	G0G1-S(R1)	3510	38490	21.93
	S-G2M (R2)	1617	9065	11.21
	G2M-G0G1 (R3)	2979	28156	18.90

Table 5.2: Disease temporal network description

Disease	Network	Nodes	Edges(t=1)	Edges(t=2)
AD	Control and Incipient	2764	11842	14518
	Incipient and Moderate	2710	15076	8093
	Moderate and Severe	2730	8528	14901
PD	Stage 0 and Stage 1	2038	14736	36044
	Stage 1 and Stage 2	2106	45890	29272
	Stage 2 and Stage 3	2114	71447	6989
Cell-cycle	R1 and R2	4298	38490	6191
	R2 and R3	3872	9065	26192
	R2 and R3	4648	28156	27455

5.5 Results

For every temporal graph constructed we calculated the transition centrality of all genes to rank the genes according to their importance in disease progression between two consecutive stages. For the analysis purpose we have examined the top 20 genes in each temporal network. Table 5.3, 5.4, 5.5 show the top 20 genes according to transition centrality in all diseases(stage-wise).

5.6 Analysis

For the validation of the findings, the role or involvement of identified genes in the particular stage are examined through literature. In addition, direct interaction of identified genes with the disease related genes(AD:PSEN1 , APP , APOE , PSEN2 , MAPT , NOS3 , HFE , ABCA7 , PLAU and MPO; PD: GBA, LRRK2, PRKN, SNCA, PINK1, PARK7 and VPS35) are also obtained.

5.6.1 Alzheimer's Disease

- Genes of interest during Control to Early(incipient) stages of AD:
  - RARA: Suping Cai et. al. in [129] found that the left superior temporal sulcus chortical thickness and RARA genetic expression are highly correlated

Table 5.3: Top 20 genes according to transition centrality for every temporal network in AD.

Control to Incipient	Incipient to Moderate	Moderate to Severe
ABCB4	JAK2	GPER1
RARA	TGFB1	CASR
KIF20B	DNM2	JAK2
RAB29	BCL2L1	AMFR
PDGFC	FGFR2	COL1A1
BCL2L1	CFTR	P2RX1
CCR2	CCR2	PLAUR
STK24	SLC8A1	CLIC4
F9	FBXW7	SERPINB13
LRP1	HLA-DRB1	CDK5
ERAP1	MAS1	FASLG
LRP6	SFN	ARHGDIB
DHRS2	PRKCD	TERT
CFTR	RARA	TRAF4
PPIA	GNAS	VCL
MARK2	PDLIM4	DBT
APC	NOS1	CCR2
HLA-DRB1	JAK1	DICER1
HMOX1	PLAUR	ABCB4
ATP7B	OAS3	PRKDC

Table 5.4: Top 20 genes according to transition centrality for every temporal network in PD

Stage 0 to Stage 1	Stage 1 to Stage 2	Stage 2 to Stage 3
RPS3	ACTB	LRRK2
BCL2	EGFR	EGFR
ACTB	BCL2	ACTB
ZBTB40	LRRK2	KRTAP6-1
ZNF629	KRAS	KRT73
HMX1	BCL2L1	ZIK1
SOHLH2	KCNC2	DLST
VAMP3	FTCD	MEF2C
EGFR	HSPA8	HSPA8
LRRK2	AK8	EXT2
AR	ZFP90	BCL2
BCL2L1	GTF2H3	FLYWCH1
SLC2A6	MAN2B2	THAP8
SLC45A1	HMGB1	KCNC2
HMGB1	ANXA4	COPG2
PCGF6	APOB	BCL2L1
ESS2	RSF1	DCTN3
FGG	MEF2C	GRIN2C
KDM6B	SLC25A5	LRRTM4
EEF1A1	EXT2	ELFN1

Table 5.5: Top 20 genes according to transition centrality for every temporal network in Cancer Cell-cycle.

R1 to R2	R2 to R3	R3 to R1
AKT1	ACTB	HSP90AA1
GAPDH	EGF	RPS27A
BRCA1	HIF1A	AKT1
MDM2	MDM2	RHOA
HSP90AA1	HSP90AA1	ACTB
RPS27A	ITGB1	GAPDH
ITGB1	RHOA	FN1
RHOA	RPS27A	VEGFA
CDK1	BRCA1	HDAC1
VEGFA	JUN	LRRK2
KDR	CDK1	CD4
FN1	FN1	DLG4
CACNA1C	CDKN1A	MDM2
HIF1A	PTPRC	CCNA2
HIST2H2BE	HIST1H2BD	JUN
PLK1	VEGFA	CDK2
TYMS	HDAC1	FOS
EGF	HIST1H2BK	GRB2
BMP4	CCND1	EGF
CD4	CDK6	AR

and associated with conversion from normal cognition to mild cognitive impairment. In [130], Deepanshi et al. also identified RARA as top gene associated with the AD progression.

- BCL2L1: Kitamura er al. in [131] investigate the levels of some of the proteins involved in the regulation of apoptosis (programmed cell death) in the brains of Alzheimer's disease patients compared to healthy controls. The proteins studied include Bcl-2, Bcl-x, Bax, Bak, Bad, ICH-1, and CPP32. Findings of the study suggest that there are changes in the levels of these proteins in the brains of Alzheimer's disease patients, particularly in regions involved in memory and learning. Specifically, the study found that the levels of Bax, Bak and Bad are increased, while the levels of Bcl-2 and Bcl-x have decreased in the brains of Alzheimer's disease patients.
- CCR2: In [132], El Khoury et al. investigated the role of a protein called CCR2 in the development and progression of Alzheimer's disease. The study used mice that were genetically modified to lack CCR2 and found that these mice showed a significant reduction in the number of microglia, a type of immune cell in the brain that is responsible for clearing away damaged cells and debris. The researchers found that the lack of CCR2 led to a significant increase in the accumulation of amyloid beta (Aβ), a protein that is known to play a key role in the development of Alzheimer's disease. This accumulation of Aβ was associated with a significant increase in the number of neurofibrillary tangles, another hallmark of Alzheimer's disease. The study also found that the mice lacking CCR2 showed a significant decrease in the number of synapses, the connections between neurons that are crucial for cognitive function. This was accompanied by a significant decline in cognitive function, as measured by tests of memory and learning. Overall, the findings of the paper suggest that CCR2 plays an important role in the accumulation of amyloid beta and the progression of Alzheimer's disease, and that the loss of this protein impairs microglial accumulation, leading to an accelerated progression of Alzheimer-like disease.
- LRP1: In [133], Shinohara discuss the findings of several studies that have investigated the role of LRP1 in the development of Alzheimer's disease. The

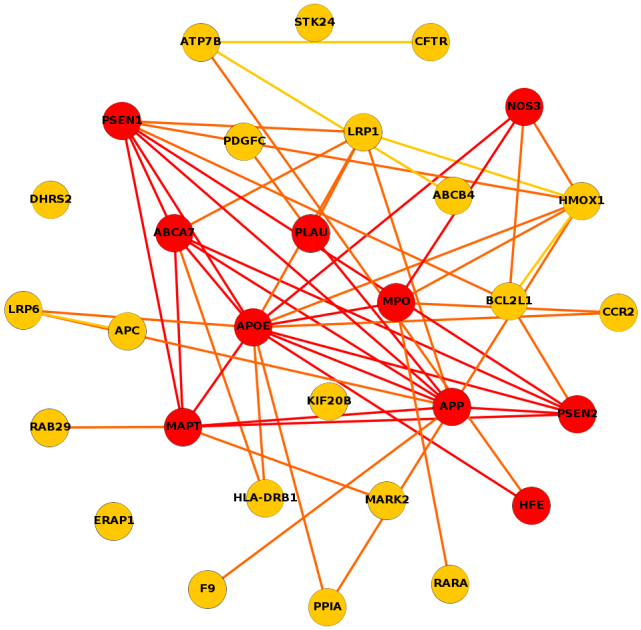


Figure 5.1: Protein-Protein interaction between identified genes(control to early stage) and top 10 AD related genes. AD ●, Identified genes ●.

studies suggest that LRP1 plays a critical role in the clearance of amyloid-β (Aβ) peptides, which are a key component of the plaques found in the brains of individuals with AD, and in the regulation of Apolipoprotein E (ApoE), a protein that is involved in the transport and clearance of Aβ. The authors conclude that the evidence from clinical and pre-clinical studies suggests that LRP1 plays a critical role in the pathogenesis of AD and targeting LRP1 may be a potential therapeutic strategy for the treatment of AD.

- ABCB4: It controls brain lipid transport and has been reported as a blood biomarker with APOE [134, 135].
- Apart from this, many of the genes have direct interaction with the AD related genes. Figure 5.1, shows the interaction between top 20 genes identified (control to early stage) and AD related genes.

- Genes of interest during Early(incipient) to moderate stage of AD:
  - JAK1/JAK2: The Nevado-Holgado et al. in [136] used a combination of

genetic data, clinical data, and empirical validation to identify potential targets for the development of therapeutic treatments for Alzheimer's disease. The study found that the Jak-Stat signaling pathway is a potential target for such treatments, as it appears to be strongly associated with the development of Alzheimer's disease.

- FGFR2 (important for brain development repair and maintenance) is one of the several FGF receptors. In [137], Klimaschewski et al. discuss about the role of fibroblast growth factor (FGF) signaling in various neurological disorders such as Alzheimer's disease, Parkinson's disease, and multiple sclerosis. The study found that FGF signaling is altered in these diseases, leading to changes in neural cell survival, differentiation, and migration. The study found that in Alzheimer's disease, FGF signaling is disrupted, leading to the formation of amyloid plaques and the death of neurons.
- PRKCD: The study in [138] found that the expression of PRKCD decreases in AD patients compared to healthy controls. This suggests that the decreased expression of PRKCD may contribute to the dysfunction of the Fc gamma receptor-mediated phagocytosis pathway in AD.
- TGFB1: The Von Bernhardi et al. in [139] suggest that abnormal TGFβ signaling may contribute to the formation of amyloid plaques and neurofibrillary tangles, which are characteristic features of Alzheimer's disease. They also suggest that abnormal TGFβ signaling may lead to inflammation and damage to the blood-brain barrier, which could further contribute to the development of Alzheimer's.
- DNM2: Finding in [140] suggests that the Dynamin 2 gene may play a role in the development of late-onset Alzheimer's disease, independent of the well established risk factor APOE-epsilon4.
- Figure 5.2 shows the interaction between top 20 identified genes(early to moderate stage) and AD related genes.

- Genes of interest during moderate to severe stage of AD:
  - GPER1: The findings of [141] indicate that GPER1/GPR30 is present in various regions of the brain, including the hypothalamus and hippocampus,

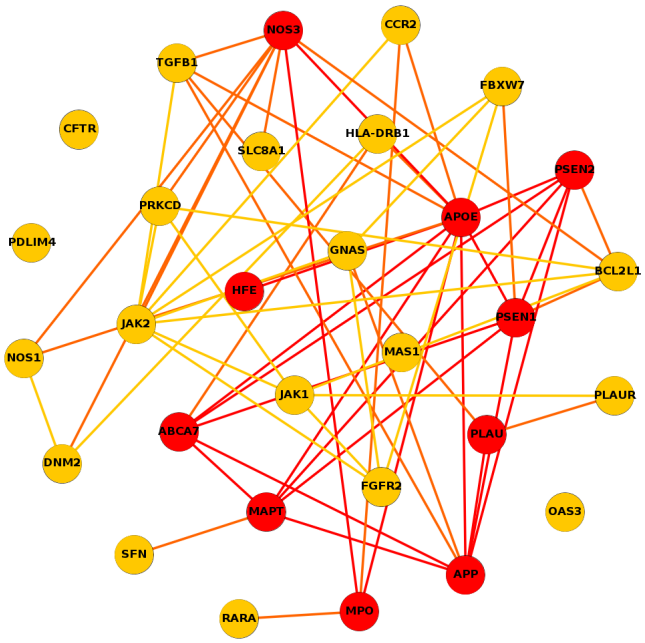


Figure 5.2: Protein-Protein interaction between identified genes(early to moderate stage) and top 10 AD related genes. AD ●, Identified genes ●.

and is involved in regulating various neural processes such as synaptic plasticity and neuroprotection.

– CASR: In [142], Gardenal et al. discussed the findings of a study in which researchers used a triple transgenic mouse model of Alzheimer's disease (AD) to investigate the expression of the calcium-sensing receptor (CaSR) in the hippocampus. The findings of this study suggest that increased expression of the CaSR in the hippocampus may play a role in the development of AD and that targeting the CaSR may be a potential therapeutic strategy for this disease. The study found that in the triple transgenic mouse model of AD, there was an increase in the expression of the CaSR in the hippocampus.

– AMFR: In [143], Yang et al. suggest that the autocrine motility factor (AMF) receptor plays a role in the processes of learning and memory in the brain. The study provides evidence that the AMF receptor is involved in synaptic plasticity, which is the ability of the connections between neurons (synapses) to change in strength. It is found that AMF receptor is essential

for spatial learning and memory in the hippocampus, a brain region involved in memory formation. The study also suggests that AMF receptor may be a potential target for the development of treatment for memory-related disorders such as Alzheimer's disease.

- CLIC4: The study [144] found that this protein is involved in the activation of the NLRP3 inflammasome, a complex of proteins that play a key role in the production of IL-1β, a pro-inflammatory cytokine. The dysregulation of the NLRP3 inflammasome is recognized as the common feature of chronic inflammatory and metabolic diseases including Alzheimer's disease.
- CDK5 (cyclin-dependent kinase 5): This is a protein that plays a role in the regulation of neural cell growth and survival. Fukasawa et al. in [145] found that the expression of the CDK5 gene was significantly higher in the brains of individuals with Alzheimer's disease as compared to those without the disease. This suggests that an over-activation of the CDK5 protein may contribute to the development of Alzheimer's.
- The interaction between top 20 identified genes(moderate to severe stage) and AD related genes are shown in Figure 5.3.

5.6.2 Parkinson's Disease

- **Control to stage 1:**
  - RPS3: In [146], De Graeve et al. found that the mammalian protein RPS3A can counteract the aggregation and toxicity of α-synuclein in a yeast model system. α-synuclein is a protein that is known to be involved in the development of Parkinson's disease and other neurodegenerative disorders. The researchers found that when RPS3A was added to yeast cells expressing α-synuclein, the protein was less likely to aggregate and caused less toxicity to the cells.
  - ACTB, cytoplasmic beta actin is associated with early-onset of severe deafness-dystonia syndrome, craniofacial dysmorphism [147].
  - EGFR, Epidermal Growth Factor Receptor signalling pathway including Cx26, might play an important role in dopaminergic neuronal cell death during the process of neuro-apoptosis [148].

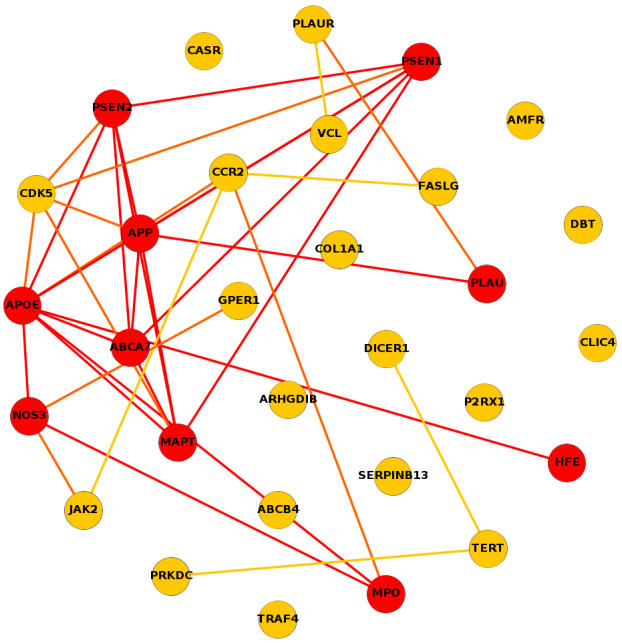


Figure 5.3: Protein-Protein interaction between identified genes(moderate to severe stage) and top 10 AD related genes. AD ●, Identified genes ●.

- SLC45A1: In [149], Ayka et al. found that SLC transporters play an important role in the transportation of molecules across the cell membrane, and the defects in these transporters may contribute to the development of neurodegenerative disorders.
- We also looked for the physical interaction of identified genes with the PD related genes. Figure 5.4 shows the interaction between top 20 identified genes(Control to stage 1) and PD related genes.
- LRRK2: LRRK2 mutations are the major cause of inherited and sporadic Parkinson's disease [150, 151].

- **Stage1 to Stage 2:**
  - BCL2L1: In [152], Chakrabarti et al. use a bioinformatics-based approach to analyze the mechanisms by which the protein α-synuclein contributes to the development of Parkinson's disease. One of the key findings of the study was that the BCL2L1 gene, which encodes the protein Bcl-xL, is a

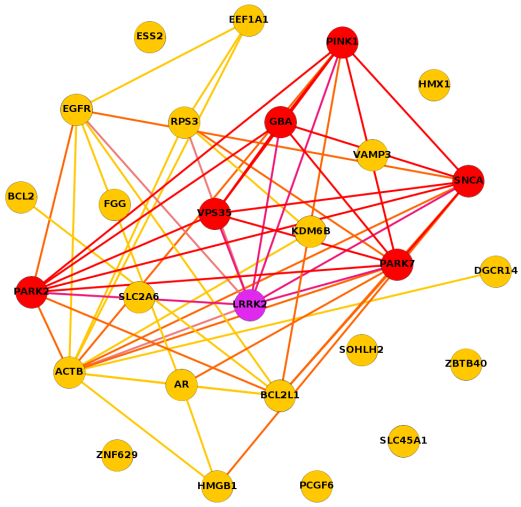


Figure 5.4: Protein-Protein interaction between identified genes(stage 0 to stage 1) and PD related genes. PD related genes ●, Identified genes ●, PD related gene which is also present in identified gene list ●.

potential target for the cytotoxic effects of α-synuclein. The researchers found that α-synuclein can interact with Bcl-xL and disrupts it's normal function, which is to protect cells from apoptosis (programmed cell death). This disruption may lead to the accumulation of damaged cells in the brain, which is a hallmark of Parkinson's disease. Additionally, the study found that the BCL2L1 gene is down-regulated in Parkinson's disease, which may further contribute to the development of the disease by reducing the levels of Bcl-xL and increasing the susceptibility of cells to α-synuclein-mediated toxicity.

– AK8: Adenylate kinase assesses the risk of diseases where oxidative stress plays a crucial role in neurodegenerative diseases [153].

– APOB: The gene found statistically significant in PD[154].

– HSPA8: The finding in [155] shows that the chemical compound rotenone, which is known to be toxic to cells and is believed to contribute to the development of Parkinson's disease, has the ability to decrease the levels of a specific protein called HSPA8/hsc70 in cells in a laboratory setting (in vitro). HSPA8/hsc70 is a type of protein called a chaperone protein,

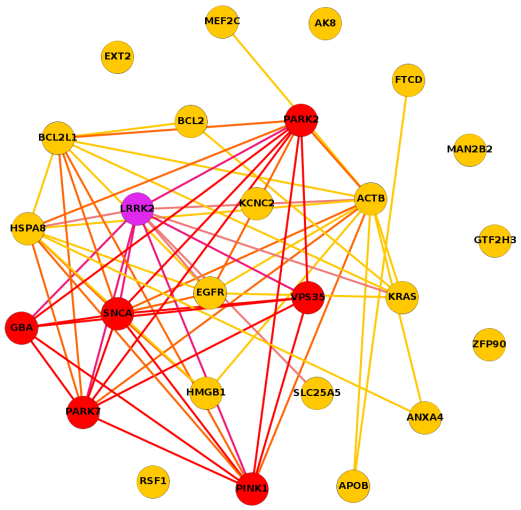


Figure 5.5: Protein-Protein interaction between identified genes(stage 1 to stage 2) and PD related genes. PD related genes ●, Identified genes ●, PD related gene which is also present in identified gene list ●.

which helps other proteins fold and function properly in the cell. The study suggests that this down-regulation of HSPA8/hsc70 by rotenone may be a new mechanism by which the chemical contributes to the development of Parkinson's disease.

- GTF2H3: In [156], the miR-369-3p/GTF2H3 gene was found to be differentially expressed in the midbrains of patients with advanced-stage PD. The miR-369-3p microRNA is known to regulate the expression of genes involved in cell growth and differentiation, and the GTF2H3 gene is a member of the general transcription factor family that is involved in DNA repair and transcriptional regulation.
- Figure 5.5 shows the interaction between top 20 identified genes(stage 1 to stage 2) and PD related genes.

- **Stage 2 to Stage 3**:
  - DLST: In [157], Hansen et al. found that the DLST gene plays an important role in the regulation of the α-ketoglutarate dehydrogenase complex (KGDHC), which is a key enzyme involved in the metabolism of energy in

the brain. The researchers found that mutations in the DLST gene can lead to a decrease in the activity of the KGDHC, which in turn can contribute to the development of neurodegenerative diseases such as Alzheimer's disease and Parkinson's disease.

- MEF2C: Many studies found MEF2C gene as risk factor for multiple neurological disorders, such as Late Onset Alzheimer's disease (LOAD) and Parkinson's disease [158, 159, 160, 161].
- EXT2: The study in [162] suggests that changes in the expression of EXT2 gene may play a role in the development of neurodegenerative diseases and aging-related changes in the brain
- GRIN2C: In [163], Liu et al. found that increasing the activity of a specific subunit of the NMDA receptor, known as GluN2C, in a specific brain region called the external globus pallidus, led to improved motor function in a mouse model of Parkinson's disease. The study also found that this increase in GluN2C activity led to increased firing of a specific type of neuron called fast-spiking neurons in the external globus pallidus. The GRIN2C gene encodes for the GluN2C subunit of the NMDA receptor. The study suggests that increasing the activity of GluN2C-containing NMDA receptors in the external globus pallidus may be a potential therapeutic strategy for improving motor function in Parkinson's disease.
- The interaction between top 20 identified genes(stage 2 to stage 3) and PD related genes are shown in Figure 5.6.

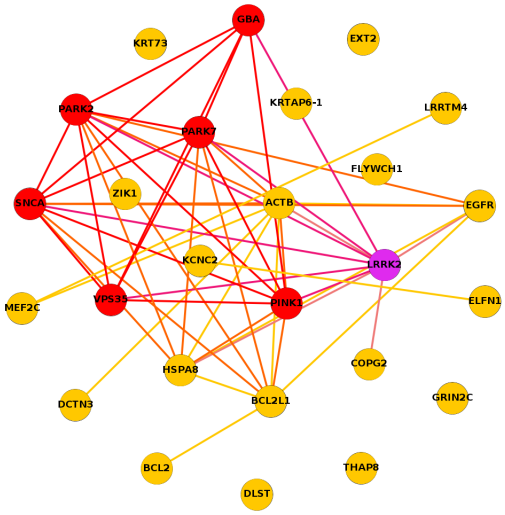


Figure 5.6: Protein-Protein interaction between identified genes(stage 2 to stage 3) and PD related genes. PD related genes ●, Identified genes●, PD related gene which is also present in identified gene list ●.

5.6.3 Human Brest cancer cell cycle

- G1-S Transition: Literature search of first 20 genes from the r1 to r2 transition table reveals that they play an important role in G1-S phase transition. Some of the genes included AKT1 (important for proliferation, observed in cancers)[164], MDM2 (preventing p53 activation and promoting cell cycle progression through G1-S phase)[164], BRCA1 (checkpoint protein involved in all phases of cell cycle)[165].

- G2M Transition: Literature search of first 20 genes from the r2 to r3 transition table reveals that they play an important role in G2-M phase transition. Some of the genes were important for producing proteins like Cyclin B and cdk1 which are crucial for G2-M phase transition and progression into M phase (Fig 1B). Also other genes like ACTB (important cell cycle regulator, control cell shape deformation during mitotic cell division)[166], RHOA (member of RHO-GTPases, important for modulating signalling pathways crucial for cell cycle progression)[166], ITGB1 (used to control YWHAZ which is known to control G2-M checkpoint protein)[167].

- M-G0/G1 transition: Since the data was obtained from a cancer cell line, the genes obtained from r3 to r1 transition play an important role in continuous cell cycle progression and helps in maintaining it's cancerous properties. Some of the genes included NT5E/CD73 were experimentally shown that upon their inhibition, epithelial-to-mesenchymal transition, cell migration and invasion (often linked to many cancer phenotypes) was inhibited[168], RPS27a, which is known to arrest cell cycle at G2-M upon knockout[169], BMP4, upon its inhibition by TGFβ/cyclin D1/Smad proteins promotes breast cancer stem cell self-renewal activity[170].

5.7 Conclusions

Recently, many studies have been carried out comparing different neuro-degenerative diseases. However, the stage-wise analysis of diseases has not been taken up. Believing that different genes can be responsible for different stages of disease progression, we, in this chapter, have introduced a modified version of betweenness centrality named *transitioncentrality* for temporal graphs. We have tested the transition centrality on stage-wise data of Alzheimer's and Parkinson's diseases. Using the transition centrality, we found the genes which may play a crucial role in the disease progression. As a result, we have identified the stage-specific genes. Interestingly, we could validate the identified genes' specificity in a particular stage from the literature.

5.8 Summary

In this chapter, we first discussed temporal graphs and different works that used temporal networks to analyse the dynamicity of different biological networks. Then centrality measures in the temporal networks and their fundamental problems are discussed. Later in the chapter, a formal definition of temporal network, temporal path and temporal shortest path according to our problem specification are given. Introduction to an algorithm to construct the temporal network is provided. We introduced a new centrality measure to identify the central genes in terms of network progression(between any two time stamps). To evaluate the transition centrality measure, this measure is tested on three different disease datasets(AD, PD, HBC). The results are validated using litera-

ture. Interestingly, in every dataset, many genes that are identified found to be related to the disease.

In conclusion, we introduced a new centrality measure called the transition centrality measure for temporal networks. Using the transition centrality measure, we identified the central genes in terms of disease progression between two consecutive stages. We also validated our findings using the literature.

Chapter 6

Conclusion & Future Work

6.1 Conclusion

This PhD thesis explores the analysis of disease networks and the identification of potential biomarker genes using gene expression data. The thesis focuses on Alzheimer's disease, but a generalized framework is also presented for the analysis of other diseases. The thesis starts by introducing the problem of identifying potential biomarker genes for diseases and the challenges associated with it. Then, it provides a review of the existing techniques for the analysis of gene expression data, such as clustering, differential gene expression analysis, pathway analysis, and network analysis. Network analysis is chosen as the primary technique for the analysis of gene expression data because it can provide a global view of the interactions among genes and can reveal the underlying biological processes involved in the disease.

The first contribution of the thesis is the development of a generalized framework(tcGONet) for the construction of disease networks and the identification of potential genes from those networks. The tcGONet framework involves the integration of gene expression data with Gene Ontology. The tcGONet framework includes several steps, such as data preprocessing, network construction, network analysis, and gene prioritization. The framework is evaluated using several datasets, and the results demonstrate its effectiveness in identifying potential genes for diseases.

The second contribution involves the community analysis of the Alzheimer's disease network using a novel neighbour-based community discovery algorithm (NBCD). The goal of this contribution is to identify nodes whose communities are disturbed within the

Alzheimer's disease network as the disease progresses. The first step of this contribution is the construction of the Alzheimer's disease network using the tcGONet framework. Next, the NBCD algorithm is applied to the Alzheimer's disease network to identify genes whose communities got disturbed. The NBCD algorithm differs from many state-of-the-art community discovery algorithms in that it takes into account the tightness relationship among nodes when assigning them to communities. NBCD also handles the conflicts that may arise when multiple communities are a good fit for a particular node. The NBCD algorithm is shown to be superior to other considered algorithms through extensive experiments. This demonstrates its effectiveness in identifying communities within the Alzheimer's disease network. Once the communities have been identified, the next step is to analyze how they change as the disease progresses. This is done by studying genes whose communities are disturbed from one stage to another stage of the disease. The identified genes are found to be related to Alzheimer's disease.

The third contribution focuses on identifying stage-specific genes in disease temporal graphs. A new centrality measure called the temporal transition centrality measure is proposed, which is designed to identify genes that play a crucial role in disease progression. First, the static networks for every stage were constructed using the tcGONet framework. Then for every consecutive network, a temporal network was constructed. Finally, the temporal transition centrality measure is used to rank genes according to their importance in network progression. To demonstrate the efficacy of the proposed centrality measure, the temporal transition centrality measure is applied to three different stage-wise disease datasets, including Alzheimer's, Parkinson's, and breast cancer. For each of these diseases, the temporal transition centrality measure was used to identify genes that were likely to be important in disease progression. The results of this analysis were quite promising. In each of the three diseases studied, the temporal transition centrality measure was able to identify genes that had previously been linked to the disease. Moreover, some of the genes identified by the centrality measure had only recently been associated with the disease, suggesting that the measure was able to detect novel disease-related genes

In summary, three distinct analyses were conducted on networks related to Alzheimer's disease with the aim of identifying potential biomarker genes. However, despite using the same dataset, the genes identified from each analysis are different. This disparity can be attributed to the specific research questions and hypotheses that

guided each analysis, resulting in varying analytical approaches and methods. Consequently, different genes were deemed important or relevant in each analysis, leading to a unique set of significant genes in each case. For instance, the first analysis sought to extract potential biomarker genes from a disease network without considering disease progression stages, whereas the second analysis aimed at identifying genes whose interacting genes had significantly changed during disease progression. Similarly, the third analysis aims at pinpointing genes that played a significant role in disease progression. Although the identified genes differed, efforts were made to identify any functional similarities between the genes identified using different methods. This involved analyzing functional similarities through Gene Ontology (GO) analysis and identifying any shared functions [See 2.2.2]. Additionally, gene expression patterns were compared to identify similarities in gene regulation (up/down-regulation [See 2.2.5]), and pathway analysis was conducted to see if genes belonged to the same pathways. Despite these analyses, no common patterns emerged among the genes identified, although some exhibited direct interactions with one another. Future biological experiments may provide insights into how these genes relate to each other and their involvement in the disease.

It was also observed that different datasets of the same disease give a different set of genes despite using the same methods. Some of the main reasons are:

- Biological variability: Gene expression in Alzheimer's disease is influenced by a variety of biological factors, such as age, gender, ethnicity, medical history and disease severity. These factors can introduce variability into the data, even when using the same experimental protocol.

- Sample size and composition: Differences in sample size and composition can also impact the results. For example, a dataset with a larger sample size may be more powerful at detecting statistically significant differences than a smaller dataset.

- Technical variability: Gene expression measurements are subject to various sources of technical variability, including sample preparation, RNA extraction, microarray or sequencing platform, normalization methods, and statistical analysis. Small differences in these steps can lead to different results, especially when dealing with less expressed genes or subtle changes in gene expression [See 2.2.3].

- Source variability: Gene expressions can be extracted from a variety of tissues and

organs in the body, including blood, skin, muscle, brain, liver, and many others. The specific tissues used for gene expression analysis depend on the research question and the experimental design of the study.

Overall, the PhD thesis contributes significantly to the understanding of Alzheimer's disease and its progression. The three contributions provide a generalized framework for constructing disease networks, a novel community discovery algorithm for analyzing Alzheimer's disease networks, and a new centrality measure for identifying genes that may play an important role in disease progression. The thesis sheds light on the complexity of Alzheimer's disease and highlights the importance of considering multiple factors and using multiple techniques for a comprehensive understanding of the disease.

6.2 Future Work

Our research has demonstrated the immense potential of the tcGONet framework in constructing biologically and statistically significant disease networks, which can serve as a valuable tool for identifying biomarker genes for various diseases. With the integration of additional biological information, we believe that the tcGONet framework can be further enhanced to create even more informative disease networks.

Furthermore, our novel community discovery algorithm, NBCD, has been shown to outperform other similar algorithms in its ability to identify communities in different types of networks. Although NBCD currently works only with undirected networks, we envision its extension to directed and dynamic networks in the future, thereby expanding its applicability to a wider range of real-world networks. Another important feature that could be added is to use it for overlapping community discovery.

Our proposed temporal centrality measure has also proven its worth in identifying genes that play crucial roles in disease progression. This measure, known as transition centrality, has the potential to be applied to other temporal networks, such as social networks, to pinpoint nodes that are of utmost importance in network progression.

In summary, our research has revealed exciting opportunities for further exploration and innovation in the field of network analysis and its applications in various domains, including biological and social networks.

References

[1] U.S. National Institute of Aging. [link]. (4, 12)

[2] Melissa C. Duff, Natalie V. Covington, Caitlin Hilverman, and Neal J. Cohen. **Semantic Memory and the Hippocampus: Revisiting, Reaffirming, and Extending the Reach of Their Critical Relationship**. *Frontiers in Human Neuroscience*, **13**, 2020. (6)

[3] Kuljeet Singh Anand and Vikas Dhikav. **Hippocampus in health and disease: An overview**. *Annals of Indian Academy of Neurology*, **15**(4):239, 2012. (6, 26)

[4] Shailendra Sahu, Pankaj Singh Dholaniya, and T Sobha Rani. **Identifying the candidate genes using co-expression, GO, and machine learning techniques for Alzheimer's disease**. *Network Modeling Analysis in Health Informatics and Bioinformatics*, **11**(1):1–12, 2022. (8, 73, 84, 85)

[5] Hélène-Marie Lanoiselée, Gaël Nicolas, David Wallon, Anne Rovelet-Lecrux, Morgane Lacour, Stéphane Rousseau, Anne-Claire Richard, Florence Pasquier, Adeline Rollin-Sillaire, Olivier Martinaud, et al. **APP, PSEN1, and PSEN2 mutations in early-onset Alzheimer disease: A genetic screening study of familial and sporadic cases**. *PLoS medicine*, **14**(3):e1002270, 2017. (12, 23)

[6] U.S. National Cancer Institute. [link]. (13)

[7] Marianna Milano. **Gene Prioritization Tools**. In Shoba Ranganathan, Michael Gribskov, Kenta Nakai, and Christian Schönbach, editors,

Encyclopedia of Bioinformatics and Computational Biology, pages 907–914. Academic Press, Oxford, 2019. (13)

[8] U.S. National Human Genome Research Institute. **Biological Pathways Fact Sheet**. (14)

[9] Narayan Behera. **Analysis of microarray gene expression data using information theory and stochastic algorithm**. In *Handbook of Statistics*, **43**, pages 349–378. Elsevier, 2020. (15)

[10] Satish Ch. Panigrahi, Md. Shafiul Alam, and Asish Mukhopadhyay. **Chapter 12 - Feature Selection and Analysis of Gene Expression Data Using Low-Dimensional Linear Programming**. In Quoc Nam Tran and Hamid Arabnia, editors, *Emerging Trends in Computational Biology, Bioinformatics, and Systems Biology*, Emerging Trends in Computer Science and Applied Computing, pages 235–264. Morgan Kaufmann, Boston, 2015. (15)

[11] Raul Rodriguez-Esteban and Xiaoyu Jiang. **Differential gene expression in disease: a comparison between high-throughput studies and the literature**. *BMC medical genomics*, **10**(1):1–10, 2017. (15)

[12] Arfa Anjum, Seema Jaggi, Eldho Varghese, Shwetank Lall, Arpan Bhowmik, and Anil Rai. **Identification of differentially expressed genes in rna-seq data of arabidopsis thaliana: A compound distribution approach**. *Journal of Computational Biology*, **23**(4):239–247, 2016. (15)

[13] Emily K Mallory, Ce Zhang, Christopher Ré, and Russ B Altman. **Large-scale extraction of gene interactions from full-text literature using DeepDive**. *Bioinformatics*, **32**(1):106–113, 2016. (17)

[14] Mathieu Garand, Manoj Kumar, Susie Shih Yin Huang, and Souhaila Al Khodor. **A literature-based approach for curating gene signatures in multifaceted diseases**. *Journal of translational medicine*, **18**(1):1–8, 2020. (17)

[15] Gwenaëlle G Lemoine, Marie-Pier Scott-Boyer, Bathilde Ambroise, Olivier Périn, and Arnaud Droit. **GWENA: gene co-expression net-**

works analysis and extended modules characterization in a single Bioconductor package. *BMC bioinformatics*, **22**(1):1–20, 2021. (18)

[16] Li Yieng Lau, Antonio Reverter, Nicholas J Hudson, Marina Naval-Sanchez, Marina RS Fortes, and Pâmela A Alexandre. **Dynamics of gene co-expression networks in time-series data: A case study in drosophila melanogaster embryogenesis**. *Frontiers in genetics*, **11**:517, 2020. (18)

[17] Joshua JR Burns, Benjamin T Shealy, Mitchell S Greer, John A Hadish, Matthew T McGowan, Tyler Biggs, Melissa C Smith, F Alex Feltus, and Stephen P Ficklin. **Addressing noise in co-expression network construction**. *Briefings in Bioinformatics*, **23**(1):bbab495, 2022. (18)

[18] Linlin Tian, Tong Chen, Jiaju Lu, Jianguo Yan, Yuting Zhang, Peifang Qin, Sentai Ding, and Yali Zhou. **Integrated Protein–Protein Interaction and Weighted Gene Co-expression Network Analysis Uncover Three Key Genes in Hepatoblastoma**. *Frontiers in cell and developmental biology*, **9**:631982, 2021. (19)

[19] Jiancheng Zhong, Chao Tang, Wei Peng, Minzhu Xie, Yusui Sun, Qiang Tang, Qiu Xiao, and Jiahong Yang. **A novel essential protein identification method based on PPI networks and gene expression data**. *BMC bioinformatics*, **22**(1):1–21, 2021. (19)

[20] S Mahapatra, R Bhuyan, J Das, and T Swarnkar. **Integrated multiplex network based approach for hub gene identification in oral cancer**. *Heliyon*, **7**(7):e07418, 2021. (19)

[21] Pascal Pons and Matthieu Latapy. **Computing Communities in Large Networks Using Random Walks**. In *Computer and Information Sciences - ISCIS 2005*, pages 284–293, 2005. (20)

[22] Zhao Yang, René Algesheimer, and Claudio J Tessone. **A comparative analysis of community detection algorithms on artificial networks**. *Scientific reports*, **6**(1):1–18, 2016. (20)

[23] Benjamin H Good, Yves-Alexandre De Montjoye, and Aaron Clauset. **Performance of modularity maximization in practical contexts**. *Physical Review E*, **81**(4):046106, 2010. (20)

[24] Christian Toth, Denis Helic, and Bernhard C Geiger. **Synwalk: community detection via random walk modelling**. *Data Mining and Knowledge Discovery*, pages 1–42, 2022. (20)

[25] Xingwang Zhao, Jiye Liang, and Jie Wang. **A community detection algorithm based on graph compression for large-scale social networks**. *Information Sciences*, **551**:358–372, 2021. (20, 57)

[26] Volkan Tunali. **Large-Scale Network Community Detection Using Similarity-Guided Merge and Refinement**. *IEEE Access*, **9**:78538–78552, 2021. (20, 57)

[27] Mark EJ Newman and Michelle Girvan. **Finding and evaluating community structure in networks**. *Physical review E*, **69**(2):026113, 2004. (21, 46, 61)

[28] Usha Nandini Raghavan, Réka Albert, and Soundar Kumara. **Near linear time algorithm to detect community structures in large-scale networks**. *Phys. Rev. E*, **76**(3):036106, 2007. (21)

[29] Mark EJ Newman. **Analysis of weighted networks**. *Physical review E*, **70**(5):056131, 2004. (21)

[30] Linton C Freeman. **A set of measures of centrality based on betweenness**. *Sociometry*, pages 35–41, 1977. (21)

[31] Mark EJ Newman. **Scientific collaboration networks. I. Network construction and fundamental results**. *Physical review E*, **64**(1):016131, 2001. (21)

[32] Phillip Bonacich. **Factoring and weighting approaches to status scores and clique identification**. *Journal of mathematical sociology*, **2**(1):113–120, 1972. (22)

[33] Sumanta Ray, Sk Md Mosaddek Hossain, Lutfunnesa Khatun, and Anirban Mukhopadhyay. **A comprehensive analysis on preservation patterns of gene co-expression networks during Alzheimer's disease progression**. *BMC bioinformatics*, **18**(1):1–21, 2017. (23, 24)

[34] Jurii V Prokhorov and Kiyosi Itô. *Probability theory and mathematical statistics*. Springer, 1983. (23)

[35] Gui-Qiong Zhu and Pei-Hui Yang. **Identifying the candidate genes for Alzheimer's disease based on the rejection region of T test**. In *2016 International Conference on Machine Learning and Cybernetics (ICMLC)*, **2**, pages 732–736. IEEE, 2016. (23, 25)

[36] Rui-ting Hu, Qian Yu, Shao-dan Zhou, Yi-xin Yin, Rui-guang Hu, Hai-peng Lu, and Bang-li Hu. **Co-expression network analysis reveals novel genes underlying Alzheimer's disease pathogenesis**. *Frontiers in Aging Neuroscience*, **12**:605961, 2020. (24)

[37] Jing Xia, David M Rocke, George Perry, and Monika Ray. **Differential network analyses of Alzheimer's disease identify early events in Alzheimer's disease pathology**. *International Journal of Alzheimer's Disease*, **2014**, 2014. (24)

[38] Jianhua Ruan and Weixiong Zhang. **Identification and evaluation of functional modules in gene co-expression networks**. In *Systems Biology and Computational Proteomics*, pages 57–76. Springer, 2006. (24)

[39] Monika Ray and Weixiong Zhang. **Analysis of Alzheimer's disease severity across brain regions by topological analysis of gene co-expression networks**. *BMC systems biology*, **4**(1):1–11, 2010. (24)

[40] Monika Ray, Reem Yunis, Xiucui Chen, and David M Rocke. **Comparison of low and high dose ionising radiation using topological analysis of gene coexpression networks**. *BMC genomics*, **13**(1):1–17, 2012. (24)

[41] Takahiro Koiwa, Kazutaka Nishiwaki, and Hayato Ohwada. **Finding unknown disease-related genes by comparing random forest results to**

secondary data in medical science study. In *Proceedings of the 7th International Conference on Computational Systems-Biology and Bioinformatics*, pages 24–27, 2016. (25)

[42] Kazutaka Nishiwaki, Katsutoshi Kanamori, and Hayato Ohwada. **Finding a disease-related gene from microarray data using random forest**. In *2016 IEEE 15th International Conference on Cognitive Informatics & Cognitive Computing (ICCI* CC)*, pages 542–546. IEEE, 2016. (25)

[43] AL-Dlaeen Dana and Abdallah Alashqur. **Using decision tree classification to assist in the prediction of Alzheimer's disease**. In *2014 6th international conference on computer science and information technology (CSIT)*, pages 122–126. IEEE, 2014. (25)

[44] Abhibhav Sharma and Pinki Dey. **A machine learning approach to unmask novel gene signatures and prediction of Alzheimer's disease within different brain regions**. *Genomics*, **113**(4):1778–1789, 2021. (25, 35)

[45] Ramya Ramaswamy, Premalatha Kandhasamy, and Swathypriyadharsini Palaniswamy. **Feature selection for Alzheimer's gene expression data using modified binary particle swarm optimization**. *IETE Journal of Research*, pages 1–12, 2021. (25, 36)

[46] Jack Cheng, Hsin-Ping Liu, Wei-Yong Lin, and Fuu-Jen Tsai. **Machine learning compensates fold-change method and highlights oxidative phosphorylation in the brain transcriptome of Alzheimer's disease**. *Scientific reports*, **11**(1):1–13, 2021. (25)

[47] RA Saputra, C Agustina, D Puspitasari, R Ramanda, D Pribadi, K Indriani, et al. **Detecting Alzheimer's disease by the decision tree methods based on particle swarm optimization**. In *Journal of Physics: Conference Series*, **1641**, page 012025. IOP Publishing, 2020. (25)

[48] Jie Kuang, Pin Zhang, TianPan Cai, ZiXuan Zou, Li Li, Nan Wang, and Lei Wu. **Prediction of transition from mild cognitive impairment to Alzheimer's disease based on a logistic regression–artificial neu-**

ral network–decision tree model. *Geriatrics & Gerontology International*, **21**(1):43–47, 2021. (25)

[49] Megan Crow, Nathaniel Lim, Sara Ballouz, Paul Pavlidis, and Jesse Gillis. **Predictability of human differential gene expression**. *Proceedings of the National Academy of Sciences*, **116**(13):6491–6500, 2019. (25)

[50] Tukur Dahiru. **P-value, a true test of statistical significance? A cautionary note**. *Annals of Ibadan postgraduate medicine*, **6**(1):21–26, 2008. (29)

[51] Haldun Akoglu. **User's guide to correlation coefficients**. *Turkish journal of emergency medicine*, **18**(3):91–93, 2018. (30)

[52] MJMMJ Mukaka. **Statistics corner: a guide to appropriate use of correlation in medical research**. *Malawi Med J*, **24**(3):69–71, 2012. (30)

[53] Michael Ashburner, Catherine A Ball, Judith A Blake, David Botstein, Heather Butler, J Michael Cherry, Allan P Davis, Kara Dolinski, Selina S Dwight, Janan T Eppig, et al. **Gene ontology: tool for the unification of biology**. *Nature genetics*, **25**(1):25–29, 2000. (30)

[54] Chenguang Zhao and Zheng Wang. **GOGO: an improved algorithm to measure the semantic similarity between gene ontology terms**. *Scientific reports*, **8**(1):1–10, 2018. (30, 31)

[55] Da Wei Huang, Brad T Sherman, Qina Tan, Jack R Collins, W Gregory Alvord, Jean Roayaei, Robert Stephens, Michael W Baseler, H Clifford Lane, and Richard A Lempicki. **The DAVID Gene Functional Classification Tool: a novel biological module-centric algorithm to functionally analyze large gene lists**. *Genome biology*, **8**(9):1–16, 2007. (30)

[56] David Martin, Christine Brun, Elisabeth Remy, Pierre Mouren, Denis Thieffry, and Bernard Jacq. **GOToolBox: functional analysis of gene datasets based on Gene Ontology**. *Genome biology*, **5**(12):1–8, 2004. (31)

[57] Aravind Subramanian, Pablo Tamayo, Vamsi K Mootha, Sayan Mukherjee, Benjamin L Ebert, Michael A Gillette, Amanda Paulovich, Scott L Pomeroy, Todd R Golub, Eric S Lander, et al. **Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles**. *Proceedings of the National Academy of Sciences*, **102**(43):15545–15550, 2005. (32)

[58] Mark A Hall. *Correlation-based feature selection for machine learning*. PhD thesis, The University of Waikato, 1999. (35)

[59] Krishna BS Swamy, Scott C Schuyler, and Jun-Yi Leu. **Protein complexes form a basis for complex hybrid incompatibility**. *Frontiers in Genetics*, **12**:609766, 2021. (39)

[60] James Keaney, Julien Gasser, Gaëlle Gillet, Diana Scholz, and Irena Kadiu. **Inhibition of Bruton's tyrosine kinase modulates microglial phagocytosis: therapeutic implications for Alzheimer's disease**. *Journal of Neuroimmune Pharmacology*, **14**(3):448–461, 2019. (40)

[61] Elhanan Pinner, Yaron Gruper, Micha Ben Zimra, Don Kristt, Moshe Laudon, David Naor, and Nava Zisapel. **CD44 splice variants as potential players in Alzheimer's disease pathology**. *Journal of Alzheimer's Disease*, **58**(4):1137–1149, 2017. (41)

[62] Ran-Sook Woo, Ji-Hye Lee, Ha-Nul Yu, Dae-Yong Song, and Tai-Kyoung Baik. **Expression of ErbB4 in the neurons of Alzheimer's disease brain and APP/PS1 mice, a model of Alzheimer's disease**. *Anatomy & cell biology*, **44**(2):116–127, 2011. (41)

[63] Abhik Ray Chaudhury, Kimberly M Gerecke, J Michael Wyss, David G Morgan, Marcia N Gordon, and Steven L Carroll. **Neuregulin-1 and erbB4 immunoreactivity is associated with neuritic plaques in Alzheimer disease brain and in a transgenic model of Alzheimer disease**. *Journal of Neuropathology & Experimental Neurology*, **62**(1):42–54, 2003. (41)

[64] Eric M Norstrom, Can Zhang, Rudolph Tanzi, and Sangram S Sisodia. **Identification of NEEP21 as a β-amyloid precursor protein-interacting protein in vivo that modulates amyloidogenic processing in vitro**. *Journal of Neuroscience*, **30**(46):15677–15685, 2010. (42)

[65] Marco Magistri, Dmitry Velmeshev, Madina Makhmutova, and Mohammad Ali Faghihi. **Transcriptomics profiling of Alzheimer's disease reveal neurovascular defects, altered amyloid-β homeostasis, and deregulated expression of long noncoding RNAs**. *Journal of Alzheimer's disease*, **48**(3):647–665, 2015. (42)

[66] K Blennow, N Bogdanovic, I Alafuzoff, R Ekman, and P Davidsson. **Synaptic pathology in Alzheimer's disease: relation to severity of dementia, but not to senile plaques, neurofibrillary tangles, or the ApoE4 allele**. *Journal of neural transmission*, **103**(5):603–618, 1996. (42)

[67] Bianca Seifert, Robert Eckenstaler, Raik Rönicke, Julia Leschik, Beat Lutz, Klaus Reymann, Volkmar Lessmann, and Tanja Brigadski. **Amyloid-beta induced changes in vesicular transport of BDNF in hippocampal neurons**. *Neural Plasticity*, **2016**, 2016. (42)

[68] Brent L Kelly and Adriana Ferreira. **Beta-amyloid disrupted synaptic vesicle endocytosis in cultured hippocampal neurons**. *Neuroscience*, **147**(1):60–70, 2007. (42)

[69] Yili Wu, Si Zhang, Qin Xu, Haiyan Zou, Weihui Zhou, Fang Cai, Tingyu Li, and Weihong Song. **Regulation of global gene expression and cell proliferation by APP**. *Scientific reports*, **6**(1):1–9, 2016. (42)

[70] Nicole T Watt, Isobel J Whitehouse, and Nigel M Hooper. **The role of zinc in Alzheimer's disease**. *International Journal of Alzheimer's disease*, **2011**, 2011. (42)

[71] Janice K Kiecolt-Glaser, Phillip T Marucha, Ana M Mercado, William B Malarkey, and Ronald Glaser. **Slowing of wound healing by psychological stress**. *The Lancet*, **346**(8984):1194–1196, 1995. (42)

[72] Yu-Hung Chen and Raymond Y Lo. **Alzheimer's disease and osteoporosis**. *Tzu-Chi Medical Journal*, **29**(3):138, 2017. (42)

[73] Filippo Radicchi, Claudio Castellano, Federico Cecconi, Vittorio Loreto, and Domenico Parisi. **Defining and identifying communities in networks**. *Proceedings of the National Academy of Sciences*, **101**(9):2658–2663, 2004. (45)

[74] Santo Fortunato and Marc Barthelemy. **Resolution limit in community detection**. *Proceedings of the national academy of sciences*, **104**(1):36–41, 2007. (46)

[75] Laura Cantini, Enzo Medico, Santo Fortunato, and Michele Caselle. **Detection of gene communities in multi-networks reveals cancer drivers**. *Scientific reports*, **5**(1):1–10, 2015. (47)

[76] Genís Calderer and Marieke L Kuijjer. **Community detection in large-scale bipartite biological networks**. *Frontiers in Genetics*, page 520, 2021. (47)

[77] Stephen J Wilson, Angela D Wilkins, Chih-Hsu Lin, Rhonald C Lua, and Olivier Lichtarge. **Discovery of functional and disease pathways by community detection in protein-protein interaction networks**. In *PACIFIC SYMPOSIUM ON BIOCOMPUTING 2017*, pages 336–347. World Scientific, 2017. (47)

[78] Marwa Ben M'barek, Amel Borgi, Walid Bedhiafi, and Sana Ben Hmida. **Genetic algorithm for community detection in biological networks**. *Procedia Computer Science*, **126**:195–204, 2018. (47)

[79] Ke Hu, Ju Xiang, Yun-Xia Yu, Liang Tang, Qin Xiang, Jian-Ming Li, Yong-Hong Tang, Yong-Jun Chen, and Yan Zhang. **Significance-based multi-scale method for network community detection and its application in disease-gene prediction**. *Plos one*, **15**(3):e0227244, 2020. (48)

[80] Akshat Singhal, Song Cao, Christopher Churas, Dexter Pratt, Santo Fortunato, Fan Zheng, and Trey Ideker. **Multiscale commu-**

nity detection in Cytoscape. *PLoS computational biology*, **16**(10):e1008239, 2020. (48)

[81] Carter Allen, Kyeong Joo Jung, Yuzhou Chang, Qin Ma, and Dongjun Chung. **Analysis of community connectivity in spatial transcriptomics data**. *bioRxiv*, 2022. (48)

[82] Shuyue Xue, Lavida RK Rogers, Minzhang Zheng, Jin He, Carlo Piermarocchi, and George I Mias. **Applying differential network analysis to longitudinal gene expression in response to perturbations**. *Frontiers in Genetics*, **13**, 2022. (48)

[83] Saharnaz Dilmaghani, Matthias R Brust, Carlos HC Ribeiro, Emmanuel Kieffer, Grégoire Danoy, and Pascal Bouvry. **From communities to protein complexes: A local community detection algorithm on PPI networks**. *Plos one*, **17**(1):e0260484, 2022. (48)

[84] Wayne Zachary. **An Information Flow Model for Conflict and Fission in Small Groups1**. *Journal of anthropological research*, **33**, 11 1976. (54, 57, 58)

[85] Gennaro Cordasco and Luisa Gargano. **Community detection via semi-synchronous label propagation algorithms**. In *2010 IEEE International Workshop on: Business Applications of Social Network Analysis (BASNA)*, pages 1–8, 2010. (56)

[86] Haijuan Yang, Jianjun Cheng, Zeyi Yang, Handong Zhang, Wenbo Zhang, Ke Yang, and Xiaoyun Chen. **A Node Similarity and Community Link Strength-Based Community Discovery Algorithm**. *Complexity*, **2021**:1–17, 2021. (56, 57)

[87] Giulio Rossetti, Letizia Milli, and Remy Cazabet. **CDLIB: a python library to extract, compare and evaluate communities from complex networks**. *Applied Network Science*, **4**:52, 2019. (56, 57, 58)

[88] Chuanwei Li, Hongmei Chen, Tianrui Li, and Xiaoling Yang. **A stable community detection approach for complex network based on density**

peak clustering and label propagation. *Applied Intelligence*, **52**(2):1188–1208, 2022. (56, 61)

[89] Yunfei Feng, Hongmei Chen, Tianrui Li, and Chuan Luo. **A novel community detection method based on whale optimization algorithm with evolutionary population**. *Applied Intelligence*, **50**:2503–2522, 2020. (57)

[90] Hamid Roghani, Asgarali Bouyer, and Esmaeil Nourani. **PLDLS: A novel parallel label diffusion and label Selection-based community detection algorithm based on Spark in social networks**. *Expert Systems with Applications*, **183**:115377, 2021. (57)

[91] Aric Hagberg, Pieter Swart, and Daniel S Chult. **Exploring network structure, dynamics, and function using NetworkX**. Technical report, Los Alamos National Lab.(LANL), Los Alamos, NM (United States), 2008. (57)

[92] Karsten Steinhaeuser and Nitesh V. Chawla. **Identifying and evaluating community structure in complex networks**. *Pattern Recognition Letters*, **31**(5):413–421, 2010. (57, 58)

[93] David Lusseau. **The emergent properties of a dolphin social network**. *Proceedings. Biological sciences / The Royal Society*, **270 Suppl 2**:S186–8, 12 2003. (57, 58)

[94] M. Girvan and M. E. J. Newman. **Community structure in social and biological networks**. *Proceedings of the National Academy of Sciences*, **99**(12):7821–7826, 2002. (57, 58)

[95] Dongbo Bu, Yi Zhao, Lun Cai, Hong Xue, Xiaopeng Zhu, Hongchao Lu, Jingfen Zhang, Shiwei Sun, Lunjiang Ling, Nan Zhang, Guo-Jie Li, and Runsheng Chen. **Topological structure analysis of the protein-protein interaction network in budding yeast**. *Nucleic acids research*, **31**:2443–50, 06 2003. (58)

[96] Jaewon Yang and Jure Leskovec. **Defining and Evaluating Network Communities based on Ground-truth**, 2015. (58)

[97] Alan Mislove, Massimiliano Marcon, Krishna P. Gummadi, Peter Druschel, and Bobby Bhattacharjee. **Measurement and Analysis of Online Social Networks**. In *Proceedings of the 5th ACM/Usenix Internet Measurement Conference (IMC'07)*, San Diego, CA, October 2007. (58)

[98] Swarup Chattopadhyay, Tanmay Basu, Asit K. Das, Kuntal Ghosh, and Late C. A. Murthy. **Towards effective discovery of natural communities in complex networks and implications in e-commerce**. *Electronic Commerce Research*, **21**(4):917–954, 2021. (59, 60)

[99] Steve Harenberg, Gonzalo Bello, L. Gjeltema, Stephen Ranshous, Jitendra Harlalka, Ramona Seay, Kanchana Padmanabhan, and Nagiza Samatova. **Community detection in large-scale networks: a survey and empirical evaluation**. *WIREs Computational Statistics*, **6**(6):426–439, 2014. (59)

[100] Andrea Lancichinetti, Santo Fortunato, and Filippo Radicchi. **Benchmark graphs for testing community detection algorithms**. *Physical review. E, Statistical, nonlinear, and soft matter physics*, **78**:046110, 11 2008. (59)

[101] Jianjun Cheng, Xing Su, Haijuan Yang, Longjie Li, Jingming Zhang, Shiyan Zhao, and Xiaoyun Chen. **Neighbor Similarity Based Agglomerative Method for Community Detection in Networks**. *Complexity*, **2019**:8292485, 2019. (59)

[102] Alexander Strehl and Joydeep Ghosh. **Cluster ensembles—a knowledge reuse framework for combining multiple partitions**. *Journal of machine learning research*, **3**(Dec):583–617, 2002. (60)

[103] Tanmay Basu and CA Murthy. **A similarity assessment technique for effective grouping of documents**. *Information Sciences*, **311**:149–162, 2015. (60)

[104] Ling Liu and M Tamer Özsu. *Encyclopedia of database systems*, **6**, chapter F-Measure. Springer, 2009. (61)

[105] Lawrence Hubert and Phipps Arabie. **Comparing partitions**. *Journal of classification*, **2**(1):193–218, 1985. (61)

[106] Nguyen Xuan Vinh, Julien Epps, and James Bailey. **Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance**. *The Journal of Machine Learning Research*, **11**:2837–2854, 2010. (61)

[107] Reihaneh Rabbany, Mansoureh Takaffoli, Justin Fagnan, Osmar R Zaïane, and Ricardo JGB Campello. **Communities validity: methodical evaluation of community mining algorithms**. *Social Network Analysis and Mining*, **3**(4):1039–1062, 2013. (61)

[108] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. **Scikit-learn: Machine learning in Python**. *the Journal of machine Learning research*, **12**:2825–2830, 2011. (62)

[109] Gita A Pathak, Talisa K Silzer, Jie Sun, Zhengyang Zhou, Ann A Daniel, Leigh Johnson, Sid O'Bryant, Nicole R Phillips, and Robert C Barber. **Genome-wide methylation of mild cognitive impairment in mexican americans highlights genes involved in synaptic transport, alzheimer's disease-precursor phenotypes, and metabolic morbidities**. *Journal of Alzheimer's Disease*, **72**(3):733–749, 2019. (76)

[110] Christopher J Folts, Stefanie Giera, Tao Li, and Xianhua Piao. **Adhesion G protein-coupled receptors as drug targets for neurological diseases**. *Trends in pharmacological sciences*, **40**(4):278–293, 2019. (76)

[111] Carlotta Camilli, Alexandra E Hoeh, Giulia De Rossi, Stephen E Moss, and John Greenwood. **LRG1: an emerging player in disease pathogenesis**. *Journal of Biomedical Science*, **29**(1):1–29, 2022. (76)

[112] Li-ping Su, Min Ji, Li Liu, Wei Sang, Jing Xue, Bo Wang, Hong-Wei Pu, and Wei Zhang. **The expression of ASAP3 and NOTCH3 and**

the clinicopathological characteristics of adult glioma patients. *Open Medicine*, **17**(1):1724–1741, 2022. (76)

[113] Dervis A Salih, Sevinc Bayram, Sebastian Guelfi, Regina H Reynolds, Maryam Shoai, Mina Ryten, Jonathan Brenton, David Zhang, Mar Matarin, Juan A Botia, et al. **Genetic variability in response to amyloid beta deposition influences Alzheimer's disease risk**. *Brain communications*, 2019. (76)

[114] Laura E Donovan, Eric B Dammer, Duc M Duong, John J Hanfelt, Allan I Levey, Nicholas T Seyfried, and James J Lah. **Exploring the potential of the platelet membrane proteome as a source of peripheral biomarkers for Alzheimer's disease**. *Alzheimer's research & therapy*, **5**(3):1–16, 2013. (77)

[115] Yuanyuan Zhao, Di Hu, Rihua Wang, Xiaoyan Sun, Philip Ropelewski, Zita Hubler, Kathleen Lundberg, Quanqiu Wang, Drew J Adams, Rong Xu, et al. **ATAD3A oligomerization promotes neuropathology and cognitive deficits in Alzheimer's disease models**. *Nature communications*, **13**(1):1–20, 2022. (77)

[116] Mohammad Mehdi Hosseinzadeh, Mario Cannataro, Pietro Hiram Guzzi, and Riccardo Dondi. **Temporal networks in biology and medicine: a survey on models, algorithms, and tools**. *Network Modeling Analysis in Health Informatics and Bioinformatics*, **12**(1):1–22, 2023. (79, 82)

[117] William Hedley Thompson, Per Brantefors, and Peter Fransson. **From static to temporal network theory: Applications to functional brain connectivity**. *Network Neuroscience*, **1**(2):69–99, 2017. (79, 81)

[118] Min Li, Xiangmao Meng, Ruiqing Zheng, Fang-Xiang Wu, Yaohang Li, Yi Pan, and Jianxin Wang. **Identification of protein complexes by using a spatial and temporal active protein interaction network**. *IEEE/ACM transactions on computational biology and bioinformatics*, **17**(3):817–827, 2017. (80)

[119] Xiangmao Meng, Wenkai Li, Ju Xiang, Hayat Dino Bedru, Wenkang Wang, Fang-Xiang Wu, and Min Li. **Temporal-spatial analysis of the essentiality of hub proteins in protein-protein interaction networks**. *IEEE Transactions on Network Science and Engineering*, **9**(5):3504–3514, 2022. (80)

[120] Pietro Hiram Guzzi, Francesco Petrizzelli, and Tommaso Mazza. **Disease spreading modeling and analysis: A survey**. *Briefings in Bioinformatics*, **23**(4), 2022. (80)

[121] Rory Humphries, Kieran Mulchrone, Jamie Tratalos, Simon J More, and Philipp Hövel. **A systematic framework of modelling epidemics on temporal networks**. *Applied Network Science*, **6**(1):1–19, 2021. (80)

[122] Rong Gao, Jinling Yan, Peiluan Li, and Luonan Chen. **Detecting the critical states during disease development based on temporal network flow entropy**. *Briefings in Bioinformatics*, 2022. (80)

[123] Zichen Wang, Parth Natekar, Challana Tea, Sharon Tamir, Hiroyuki Hakozaki, and Johannes Schoeneberg. **Mitotnt: mitochondrial temporal network tracking for 4d live-cell fluorescence microscopy data**. *bioRxiv*, 2022. (80)

[124] B Bui Xuan, Afonso Ferreira, and Aubin Jarry. **Computing shortest, fastest, and foremost journeys in dynamic networks**. *International Journal of Foundations of Computer Science*, **14**(02):267–285, 2003. (82)

[125] Hyoungshick Kim and Ross Anderson. **Temporal node centrality in complex networks**. *Physical Review E*, **85**(2):026107, 2012. (82)

[126] Amir Afrasiabi Rad, Paola Flocchini, and Joanne Gaudet. **Computation and analysis of temporal betweenness in a knowledge mobilization network**. *Computational social networks*, **4**(1):1–22, 2017. (82)

[127] Sebastian Buss, Hendrik Molter, Rolf Niedermeier, and Maciej Rymar. **Algorithmic aspects of temporal betweenness**. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2084–2092, 2020. (82)

[128] ED Israels and LG Israels. **The cell cycle**. *The oncologist*, **5**(6):510–513, 2000. (84)

[129] Suping Cai, Fan Yang, Xuwen Wang, Sijia Wu, Liyu Huang, Alzheimer's Disease Neuroimaging Initiative, et al. **Structural brain characteristics and gene co-expression analysis: A study with outcome label from normal cognition to mild cognitive impairment**. *Neurobiology of Learning and Memory*, **191**:107620, 2022. (86)

[130] Deepanshi Vijh, Md Ali Imam, Mohd Maksuf Ul Haque, Subhajit Das, Asimul Islam, and Md Zubbair Malik. **Network pharmacology and bioinformatics approach reveals the therapeutic mechanism of action of curcumin in Alzheimer disease**. *Metabolic Brain Disease*, pages 1–16, 2023. (90)

[131] Yoshihisa Kitamura, Shun Shimohama, Wataru Kamoshima, Takashi Ota, Yasuji Matsuoka, Yasuyuki Nomura, Mark A Smith, George Perry, Peter J Whitehouse, and Takashi Taniguchi. **Alteration of proteins regulating apoptosis, Bcl-2, Bcl-x, Bax, Bak, Bad, ICH-1 and CPP32, in Alzheimer's disease**. *Brain research*, **780**(2):260–269, 1998. (90)

[132] Joseph El Khoury, Michelle Toft, Suzanne E Hickman, Terry K Means, Kinya Terada, Changiz Geula, and Andrew D Luster. **Ccr2 deficiency impairs microglial accumulation and accelerates progression of Alzheimer-like disease**. *Nature medicine*, **13**(4):432–438, 2007. (90)

[133] Mitsuru Shinohara, Masaya Tachibana, Takahisa Kanekiyo, and Guojun Bu. **Role of LRP1 in the pathogenesis of Alzheimer's disease: evidence from clinical and preclinical studies: Thematic Review Series: ApoE and Lipid Homeostasis in Alzheimer's Disease**. *Journal of lipid research*, **58**(7):1267–1281, 2017. (90)

[134] Woojin Scott Kim, Cynthia Shannon Weickert, and Brett Garner. **Role of ATP-binding cassette transporters in brain lipid transport and neurological disease**. *Journal of neurochemistry*, **104**(5):1145–1166, 2008. (91)

[135] Laura Madrid, Sonia Moreno-Grau, Shahzad Ahmad, Antonio González-Pérez, Itziar de Rojas, Rui Xia, Pamela V Martino Adami, Pablo García-González, Luca Kleineidam, Qiong Yang, et al. **Multiomics integrative analysis identifies APOE allele-specific blood biomarkers associated to Alzheimer's disease etiopathogenesis**. *Aging (Albany NY)*, **13**(7):9277, 2021. (91)

[136] Alejo J Nevado-Holgado, Elena Ribe, Laura Thei, Laura Furlong, Miguel-Angel Mayer, Jie Quan, Jill C Richardson, Jonathan Cavanagh, NIMA Consortium, and Simon Lovestone. **Genetic and real-world clinical data, combined with empirical validation, nominate Jak-Stat signaling as a target for Alzheimer's disease therapeutic development**. *Cells*, **8**(5):425, 2019. (91)

[137] Lars Klimaschewski and Peter Claus. **Fibroblast growth factor signalling in the diseased nervous system**. *Molecular Neurobiology*, **58**(8):3884–3902, 2021. (92)

[138] Young Ho Park, Angela Hodges, Shannon L Risacher, Kuang Lin, Jae-Won Jang, Soyeon Ahn, SangYun Kim, Simon Lovestone, Andrew Simmons, Michael W Weiner, et al. **Dysregulated Fc gamma receptor–mediated phagocytosis pathway in Alzheimer's disease: network-based gene expression analysis**. *Neurobiology of aging*, **88**:24–32, 2020. (92)

[139] Rommy Von Bernhardi, Francisca Cornejo, Guillermo E Parada, and Jaime Eugenín. **Role of TGFβ signaling in the pathogenesis of Alzheimer's disease**. *Frontiers in cellular neuroscience*, **9**:426, 2015. (92)

[140] Nuripa Jenishbekovna Aidaralieva, Kouzin Kamino, Ryo Kimura, Mitsuko Yamamoto, Takeshi Morihara, Hiroaki Kazui, Ryota Hashimoto, Toshihisa Tanaka, Takashi Kudo, Tomoyuki Kida, et al. **Dynamin 2 gene is a novel susceptibility gene for late-onset Alzheimer disease in non-APOE-ε4 carriers**. *Journal of human genetics*, **53**(4):296–302, 2008. (92)

[141] Maria M Hadjimarkou and Nandini Vasudevan. **GPER1/GPR30 in the brain: Crosstalk with classical estrogen receptors and implications for**

behavior. *The Journal of Steroid Biochemistry and Molecular Biology*, **176**:57–64, 2018. (92)

[142] Emanuela Gardenal, Anna Chiarini, Ubaldo Armato, Ilaria Dal Prà, Alexei Verkhratsky, and José J Rodríguez. **Increased calcium-sensing receptor immunoreactivity in the hippocampus of a triple transgenic mouse model of Alzheimer's disease**. *Frontiers in neuroscience*, **11**:81, 2017. (93)

[143] Yong Yang, Xiao-Rui Cheng, Gui-Rong Zhang, Wen-Xia Zhou, and Yong-Xiang Zhang. **Autocrine motility factor receptor is involved in the process of learning and memory in the central nervous system**. *Behavioural brain research*, **229**(2):412–418, 2012. (93)

[144] Raquel Domingo-Fernández, Rebecca C Coll, Jay Kearney, Samuel Breit, and Luke AJ O'Neill. **The intracellular chloride channel proteins CLIC1 and CLIC4 induce IL-1β transcription and activate the NLRP3 inflammasome**. *Journal of Biological Chemistry*, **292**(29):12077–12087, 2017. (94)

[145] Josianne T Fukasawa, Roger W de Labio, Lucas T Rasmussen, Lucieni C de Oliveira, Elizabeth Chen, Joao Villares, Gustavo Tureck, Marilia de Arruda C Smith, and Spencer LM Payao. **CDK5 and MAPT gene expression in Alzheimer's disease brain samples**. *Current Alzheimer Research*, **15**(2):182–186, 2018. (94)

[146] Stijn De Graeve, Sarah Marinelli, Frank Stolz, Jelle Hendrix, Jurgen Vandamme, Yves Engelborghs, Patrick Van Dijck, and Johan M Thevelein. **Mammalian ribosomal and chaperone protein RPS3A counteracts α-synuclein aggregation and toxicity in a yeast model system**. *Biochemical Journal*, **455**(3):295–306, 2013. (94)

[147] Giulia Straccia, Chiara Reale, Massimo Castellani, Isabel Colangelo, Eva Orunesu, Sara Meoni, Elena Moro, Paul Krack, Holger Prokisch, Michael Zech, et al. **ACTB gene mutation in combined**

Dystonia-Deafness syndrome with parkinsonism: Expanding the phenotype and highlighting the long-term GPi DBS outcome. *Parkinsonism & related disorders*, 2022. (94)

[148] In-Su Kim, Sushruta Koppula, Shin-Young Park, and Dong-Kug Choi. **Analysis of epidermal growth factor receptor related gene expression changes in a cellular and animal model of parkinson's disease**. *International journal of molecular sciences*, **18**(2):430, 2017. (94)

[149] Asli Aykaç and Ahmet Ozer Sehirli. **The role of the SLC transporters protein in the neurodegenerative disorders**. *Clinical Psychopharmacology and Neuroscience*, **18**(2):174, 2020. (95)

[150] Jie-Qiong Li, Lan Tan, and Jin-Tai Yu. **The role of the LRRK2 gene in Parkinsonism**. *Molecular neurodegeneration*, **9**(1):1–17, 2014. (95)

[151] Pilar Rivero-Ríos, María Romo-Lozano, Rachel Fasiczka, Yahaira Naaldijk, and Sabine Hilfiker. **LRRK2-Related Parkinson's disease due to altered endolysosomal biology with variable lewy body pathology: A hypothesis**. *Frontiers in Neuroscience*, **14**:556, 2020. (95)

[152] Sankha S Chakrabarti, Venkatadri S Sunder, Upinder Kaur, Sapna Bala, Priyanka Sharma, Manjari Kiran, Ravindra K Rawal, and Sasanka Chakrabarti. **Identifying the mechanisms of α-synuclein-mediated cytotoxicity in Parkinson's disease: new insights from a bioinformatics-based approach**. *Future Neurology*, **15**(3):FNL49, 2020. (95)

[153] Mihaela Ileana Ionescu. **Adenylate kinase: a ubiquitous enzyme correlated with medical conditions**. *The Protein Journal*, **38**(2):120–133, 2019. (96)

[154] Fang Fang, Yiqiang Zhan, Niklas Hammar, Xia Shen, Karin Wirdefeldt, Göran Walldius, and Daniela Mariosa. **Lipids, Apolipoproteins, and the risk of Parkinson disease: a prospective cohort study and a Mendelian randomization analysis**. *Circulation research*, **125**(6):643–652, 2019. (96)

[155] Gessica Sala, Daniele Marinig, Chiara Riva, Alessandro Arosio, Giovanni Stefanoni, Laura Brighina, Matteo Formenti, Lilia Alberghina, Anna Maria Colangelo, and Carlo Ferrarese. **Rotenone down-regulates HSPA8/hsc70 chaperone protein in vitro: a new possible toxic mechanism contributing to Parkinson's disease**. *Neurotoxicology*, **54**:161–169, 2016. (96)

[156] Lucas Caldi Gomes, Ana Galhoz, Gaurav Jain, Anna-Elisa Roser, Fabian Maass, Eleonora Carboni, Elisabeth Barski, Christof Lenz, Katja Lohmann, Christine Klein, et al. **Multi-omic landscaping of human midbrains identifies disease-relevant molecular targets and pathways in advanced-stage Parkinson's disease**. *Clinical and translational medicine*, **12**(1):e692, 2022. (97)

[157] Grace E Hansen and Gary E Gibson. **The α-Ketoglutarate Dehydrogenase Complex as a Hub of Plasticity in Neurodegeneration and Regeneration**. *International Journal of Molecular Sciences*, **23**(20):12403, 2022. (97)

[158] Hua She, Qian Yang, and Zixu Mao. **Neurotoxin-induced selective ubiquitination and regulation of MEF2A isoform in neuronal stress response**. *Journal of neurochemistry*, **122**(6):1203–1210, 2012. (98)

[159] Jean-Bernard Dietrich. **The MEF2 family and the brain: from molecules to memory**. *Cell and tissue research*, **352**(2):179–190, 2013. (98)

[160] Gail Davies, Nicola Armstrong, Joshua C Bis, Jan Bressler, Vincent Chouraki, Sudheer Giddaluru, Edith Hofer, Carla A Ibrahim-Verbaas, Mirna Kirin, J Lahti, et al. **Genetic contributions to variation in general cognitive function: a meta-analysis of genome-wide association studies in the CHARGE consortium (N= 53 949)**. *Molecular psychiatry*, **20**(2):183–192, 2015. (98)

[161] Preeti Sunderaraman, Stephanie Cosentino, Nicole Schupf, Jennifer Manly, Yian Gu, and Sandra Barral. **MEF2C common genetic variation is associated with different aspects of cognition in non-hispanic**

white and caribbean hispanic non-demented older adults. *Frontiers in genetics*, **12**, 2021. (98)

[162] OB Shevelev, VI Rykova, LA Fedoseeva, E Yu Leberfarb, GM Dymshits, and NG Kolosova. **Expression of Ext1, Ext2, and heparanase genes in brain of senescence-accelerated OXYS rats in early ontogenesis and during development of neurodegenerative changes**. *Biochemistry (Moscow)*, **77**(1):56–61, 2012. (98)

[163] Jinxu Liu, Gajanan P Shelkar, Lopmudra P Sarode, Dinesh Y Gawande, Fabao Zhao, Rasmus Praetorius Clausen, Rajesh R Ugale, and Shashank Manohar Dravid. **Facilitation of GluN2C-containing NMDA receptors in the external globus pallidus increases firing of fast spiking neurons and improves motor function in a hemiparkinsonian mouse model**. *Neurobiology of disease*, **150**:105254, 2021. (98)

[164] Naihan Xu, Yuanzhi Lao, Yaou Zhang, and David A Gillespie. **Akt: a double-edged sword in cell proliferation and genome stability**. *Journal of oncology*, **2012**, 2012. (99)

[165] Chu-Xia Deng. **BRCA1: cell cycle checkpoint, genetic instability, DNA damage response and cancer evolution**. *Nucleic acids research*, **34**(5):1416–1426, 2006. (99)

[166] Paulius Gibieža and Vilma Petrikaitė. **The regulation of actin dynamics during cell division and malignancy**. *American Journal of Cancer Research*, **11**(9):4050, 2021. (99)

[167] Jinghe Xie, Tingting Guo, Zhiyong Zhong, Ning Wang, Yan Liang, Weiping Zeng, Shoupei Liu, Qicong Chen, Xianglian Tang, Haibin Wu, et al. **ITGB1 Drives Hepatocellular Carcinoma Progression by Modulating Cell Cycle Process Through PXN/YWHAZ/AKT Pathways**. *Frontiers in cell and developmental biology*, **9**, 2021. (99)

[168] Jianjie Zhu, Yuanyuan Zeng, Wei Li, Hualong Qin, Zhe Lei, Dan Shen, Dongmei Gu, Jian-an Huang, and Zeyi Liu. **CD73/NT5E is a target of**

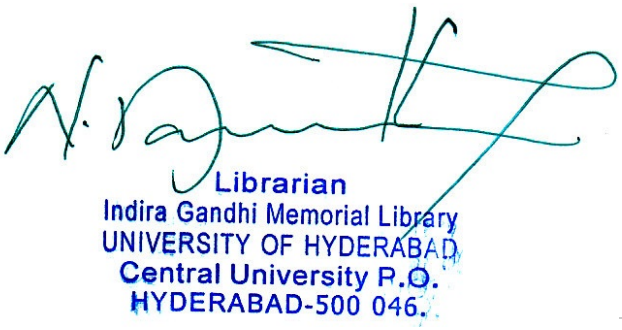
miR-30a-5p and plays an important role in the pathogenesis of non-small cell lung cancer. *Molecular cancer*, **16**(1):1–15, 2017. (100)

[169] Hongyan Li, Hong Zhang, Guomin Huang, Zhitong Bing, Duling Xu, Jiadi Liu, Hongtao Luo, and Xiaoli An. **Loss of RPS27a expression regulates the cell cycle, apoptosis, and proliferation via the RPL11-MDM2-p53 pathway in lung adenocarcinoma cells**. *Journal of Experimental & Clinical Cancer Research*, **41**(1):1–20, 2022. (100)

[170] Gang Yan, Meiou Dai, Chenjing Zhang, Sophie Poulet, Alaa Moamer, Ni Wang, Julien Boudreault, Suhad Ali, and Jean-Jacques Lebrun. **TGFβ/cyclin D1/Smad-mediated inhibition of BMP4 promotes breast cancer stem cell self-renewal activity**. *Oncogenesis*, **10**(3):1–14, 2021. (100)

Disease Network Construction and Analysis: A case study of Alzheimer's Disease

by Shailendra Sahu



Submission date: 15-Mar-2023 03:11PM (UTC+0530)
Submission ID: 2037683229
File name: Shailendra_Sahu.pdf (5.11M)
Word count: 30157
Character count: 151212

Disease Network Construction and Analysis: A case study of Alzheimer's Disease

ORIGINALITY REPORT Overall Similarity: 40-(8+13)=9% 15/3/2023

40%	9%	38%	3%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Shailendra Sahu, T. Sobha Rani. "A neighbour-similarity based community discovery algorithm", Expert Systems with Applications, 2022 Publication This publication belongs to my student.	18%
2	Shailendra Sahu, Pankaj Singh Dholaniya, T. Sobha Rani. "Identifying the candidate genes using co-expression, GO, and machine learning techniques for Alzheimer's disease", Network Modeling Analysis in Health Informatics and Bioinformatics, 2022 Publication This publication belongs to my student.	13%
3	www.researchgate.net Internet Source	2%
4	www.ncbi.nlm.nih.gov Internet Source	1%
5	J. Beringer, J. -F. Arguin, R. M. Barnett, K. Copic et al. "Review of Particle Physics", Physical Review D, 2012 Publication	<1%